*ИЗВЕСТИЯ РАН. ТЕОРИЯ И СИСТЕМЫ УПРАВЛЕНИЯ, 2023, № 3, с. 76–89*

## ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ

*УДК 004.81;004.852;004.855.5*

# ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ ДЛЯ МОДЕЛЬНЫХ ЗАДАЧ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ

**© 2023 г. С. С. Семенов[a],*, В. И. Цурков[b],****

[a] *МФТИ, Долгопрудный, МО, Россия*

[b] *ФИЦ ИУ РАН, Москва, Россия*

**e-mail: semenov.ss@phystech.edu*

***e-mail: tsur@ccas.ru*

Поступила в редакцию 10.11.2022 г.
После доработки 08.01.2023 г.
Принята к публикации 06.02.2023 г.

Оптимизируются функционалы динамических систем различного вида с помощью современных методов обучения с подкреплением. Рассматриваются линейная задача распределения ресурсов, задача оптимального потребления и ее стохастические модификации. В обучении с подкреплением использовались методы градиента стратегии.

**DOI:** 10.31857/S0002338823030125, **EDN:** EVAFAM

**Введение.** К настоящему времени известны широко применяемые классические подходы в оптимальном управлении, такие, как принцип максимума Понтрягина, принцип оптимальности Беллмана, численные методы и т.д. Однако круг их использования ограничен. В последнее время в данной области получило развитие машинное обучение на нейронных сетях.

Обучение с подкреплением формулирует задачу оптимального управления на языке марковского процесса принятия решений и осуществляет переход к эквивалентной задаче оптимизации. Задача оптимального управления может быть переведена на язык принятия решения. Для осуществления перехода к эквивалентной задаче оптимизации вводятся такие объекты, как агент и среда. Среда – это некоторый марковский процесс принятия решений, который характеризуется пространством действий $s \in S$, пространством состояний $a \in A$, динамикой среды $\mathcal{P}^a_{ss'}$, функцией вознаграждения $\mathcal{R}^a_s$ и дисконтирующим фактором $\gamma$. Агент представляет собой некоторый стохастический алгоритм, который принимает на вход состояние среды $s$ и возвращает действие $a$, которое необходимо принять, чтобы максимизировать итоговое вознаграждение. Взаимодействуя со средой, агент накапливает опыт игры и с каждой новой попыткой улучшает стратегию своей игры. Осуществив сотни тысяч попыток, стратегия агента сходится к оптимальной. Полученная стратегия максимизирует дисконтированную награду агента по траекториям при взаимодействии со средой и, как следствие, решает исходную задачу оптимального управления.

В настоящее время существуют два основных семейства алгоритмов обучения с подкреплением, в основе которых лежат различные принципы, а именно семейство Actor-Critic и семейство Policy Gradient. В решении практических задач наиболее эффективными являются алгоритмы DDPG [1], TRPO [2] из Policy Gradient и алгоритмы SAC [3], A2C [4] из Actor-Critic. В данной работе используется алгоритм Proximal Policy Optimization [5–7], который базируется на идее обновления весов не только с помощью подсчета градиента стратегии как в Policy Gradient, но и на выборе наиболее релевантного действия актором и оценке правдоподобия выбранного действия критиком как в Actor-Critic. Обучение с подкреплением успешно применяется в работах [8–12].

Рассматриваются три модельных задачи. В первой алгоритм эффективно строит точки переключения в разрывных управлениях. Во второй происходит совпадение с решением, полученным с помощью рекуррентных соотношений Беллмана. Наконец, в третьей строятся оптимальные решения для различных видов случайных параметров.

**1. Постановка задачи.** 1.1. Линейная задача оптимального управления. Имеем простейшую динамическую модель распределения ресурсов с двумя подсистемами. Близкие задачи оптимального управления широко представлены в книге [13]:

$$\begin{gathered}\mathcal{F}(u) = x_1(T) + x_2(T) \to \max,\\ \frac{dx_1}{dt} = a_1(t)u_1(t),\\ \frac{dx_2}{dt} = a_2(t)u_2(t),\\ u_1 \geq 0, \quad u_2 \geq 0, \quad u_1 + u_2 \leq W = \text{const},\\ t \in [0,T], \quad u = [u_1,u_2]^{\mathrm{T}},\\ x_1(0) = x^1, \quad x_2(0) = x^2,\end{gathered} \tag{1.1}$$

где $u_1(t)$, $u_2(t)$ – управления, $x_1(t)$, $x_2(t)$ – фазовые переменные, $t \in \mathbb{R}_+$ – время, $T \in \mathbb{R}_+$ – конечное время, $a_1(t)$, $a_2(t)$ – заданные функции, $W$, $x^1$, $x^2$ – константы.

1.2. Задача оптимального потребления [14]. Имеем

$$\begin{gathered}\mathcal{F}(u) = \int_0^T \exp(-\alpha t)g(u(t))dt \to \max,\\ \dot{x} = \rho x - u, \quad x(0) = x_0, \quad x(T) = 0,\\ u(t) \geq 0, \quad t \in [0,T],\end{gathered} \tag{1.2}$$

где $u(t)$ – интенсивность потребления или мгновенное значение потребления в момент времени $t$; $x(t)$ – денежный капитал в момент времени $t$; $g(u)$ – функция полезности потребления; $\alpha$ – ставка дисконтирования; $\rho$ – банковская депозитная ставка.

Функция полезности потребления $g(u)$ удовлетворяет условиям:

1) $g(u) \geq 0$,

2) $g(u)$ строго возрастает для $u \geq 0$,

3) $g(u)$ строго вогнутая дифференцируемая функция в области $u \geq 0$.

1.3. Дискретная стохастическая задача. Пусть депозитная ставка является случайной величиной:

$$\begin{gathered}\mathcal{F}(u) = \mathbb{E}_\rho \sum_{k=0}^{K} \exp(-\alpha k)g(u_k) \to \max,\\ x_{k+1} - x_k = \rho x_k - u_k,\\ x_0 = a, \quad x_K = 0,\\ u_k \geq 0, \quad t \in \overline{0,K},\end{gathered} \tag{1.3}$$

где $\rho$ – депозитная ставка; $x_k$ – значение капитала в момент времени $k$; $u_k$ – значение потребления в момент времени $k$; $\mathbb{E}_\rho[f]$ – математическое ожидание $f$ по $\rho$, $k$ и $K \in \mathcal{J} = \{0, 1, 2, \ldots\}$.

Будут представлены дискретное распределение вероятностей, винеровский процесс и пуассоновский процесс.

Рассмотрим дискретное распределение вероятностей. Пусть $\{t\}_{t\in\mathbb{N}}$ – процесс, заданный в виде последовательности независимых и одинаково распределенных случайных величин. Пример случайного распределения представлен на рис. 1 с параметрами:

$$\mathbb{P}(\rho = r) = p, \quad r \in \{1.05, 1.06, 1.07, 1.08, 1.09\}, \tag{1.4}$$

где $\mathbb{P}(\rho = r)$ – вероятность того, что $\rho$ принимает значение $r$.

Рассмотрим винеровский процесс. Пусть $\{\rho_t\}_{t\in[0,\,T]}$ – геометрическое броуновское движение, т.е. процесс, подчиняющийся дифференциальному уравнению:

$$\frac{d\rho_t}{\rho_t} = \alpha dt + \sigma dW_t, \tag{1.5}$$

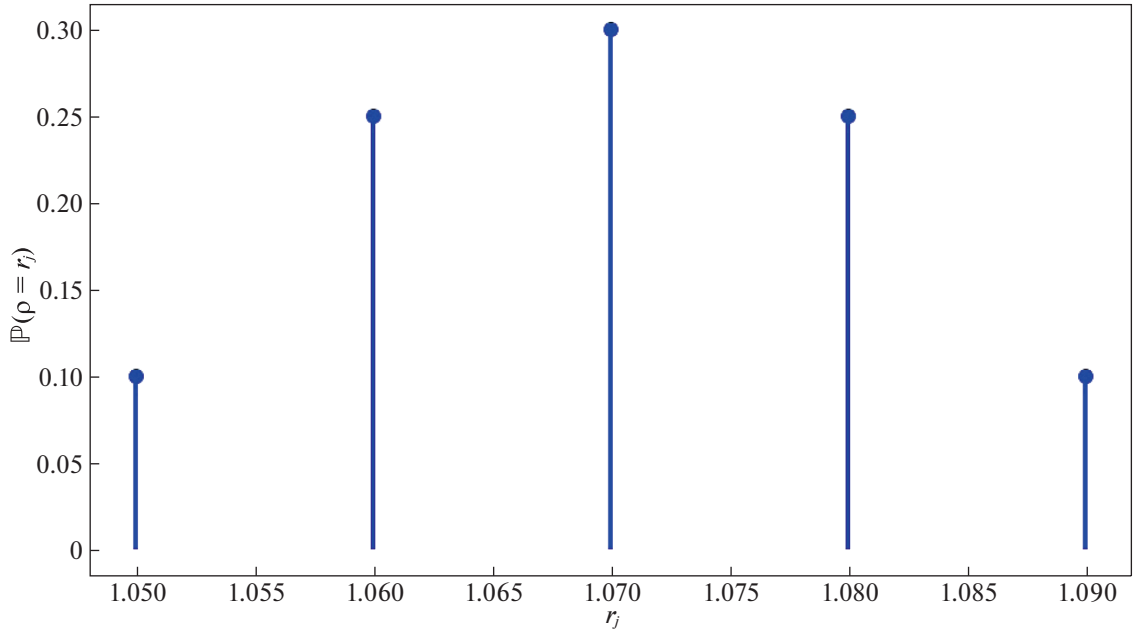


**Рис. 1.** Коэффициент прироста капитала как дискретное распределение вероятностей

где $W_t$ – винеровский процесс, $dW_t$ – дифференциал винеровского процесса $t \in \mathbb{R}_+$.

Для решения этого уравнения нам понадобится лемма Ито, сформулированная и доказанная, например, в [15].

О п р е д е л е н и е 1. Процесс $X_t$ будем называть процессом Ито, если

$$X_t = X_0 + \int_0^t u(s,\omega)ds + \int_0^t v(t,\omega)dW_t,$$

$$dX_t = udt + wdW_t.$$

Л е м м а И т о. Пусть $X_t$ – процесс Ито, задаваемый дифференциалом:

$$dX_t = udt + wdW_t,$$

и пусть $g(t, x)$ дважды непрерывно дифференцируемая функция на $[0, +\infty) \times \mathbb{R}$. Тогда $Y_t = g(t, X_t)$ – снова есть процесс Ито и

$$dY_t = \frac{\partial g}{\partial t}dt + \frac{\partial g}{\partial x}dX_t + \frac{1}{2}\frac{\partial^2 g}{\partial^2 x}(dX_t)^2,$$

где $(cX_t)^2$ вычисляется с использованием следующих правил:

$$dtd \sim dtdW_t \sim 0, \quad dW_t dW_t \sim dt.$$

Воспользовавшись леммой Ито, получаем

$$\rho_t = \rho_0 \exp\left(\left(\alpha - \frac{\sigma^2}{2}\right)t + \sigma W_t)\right).$$

Поведение траекторий представлено на рис. 2.

Рассмотрим пуассоновский процесс. Пусть $\{\rho_t\}_{t \in [0,\ T]}$ – случайная функция, подчиняющаяся дифференциальному уравнению:

$$\frac{d\rho_t}{\rho_t} = (\alpha - \lambda)dt + \sigma dN_t, \tag{1.6}$$

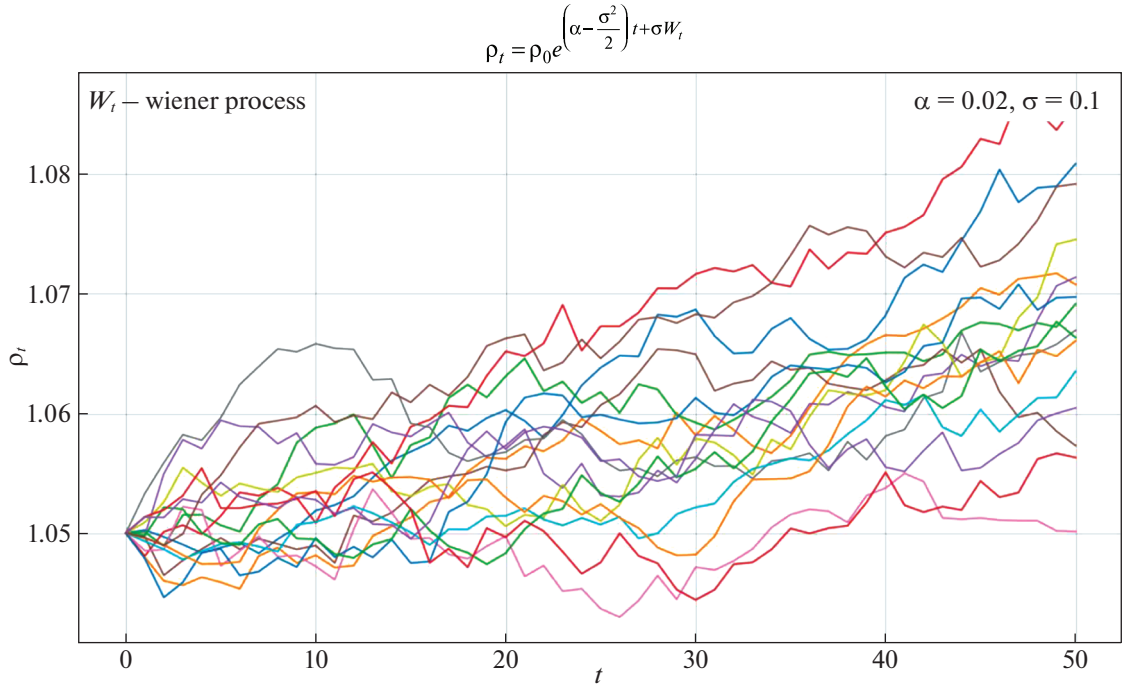


**Рис. 2.** Коэффициент прироста капитала как марковский процесс (16 реализаций)

где $N_t$ – счетчик пуассоновского процесса с интенсивностью $\lambda$. Решение стохастического дифференциального уравнения получается с использованием леммы Ито аналогично случаю винеровского процесса. Имеем

$$\rho_t = \rho_0 \exp((\alpha - \lambda)t + N_t \ln(1 + \sigma)).$$

Поведение траекторий представлено на рис. 3.

**2. Решения детерминированных задач.** 2.1. Линейная задача распределения ресурсов. Для задачи (1.1) применим принцип максимума Понтрягина (см., например, в [16]).

Определим класс (множество) допустимых управлений как множество $m$-мерных вектор-функций $u(t) = (u_1(t), u_2(t), \ldots, u_m(t))$, кусочно-непрерывных на заданном отрезке $\mathbf{T} = [0, T]$ с условием (ограничением) $u(t) \in \mathcal{U}$, $t \in \mathbf{T}$, где $\mathcal{U} \subseteq \mathbb{R}^m$ – заданное множество. В формальной записи множество допустимых управлений $\mathcal{W}$ имеет следующий вид:

$$\mathcal{W} = \{u \in \mathcal{C}_m(T), \mathbf{u}(t) \in \mathcal{U}, t \in \mathbf{T}\}.$$

Рассмотрим обобщенный вариант простейшей задачи для линейной по состоянию системы с неразделенными переменными $x$, $u$:

$$\begin{gathered} \mathcal{F}(\mathbf{u}) = \langle \mathbf{c}, \mathbf{x}(t_1) \rangle \to \min, \quad \mathbf{u} \in \mathcal{W}, \\ \dot{\mathbf{x}} = \mathbf{A}(u,t)x + \mathbf{b}(\mathbf{u},t), \quad \mathbf{x}(t_0) = \mathbf{x}_0, \\ \mathcal{W} = \{\mathbf{u} \in \mathcal{C}_m(T), \mathbf{u}(t) \in \mathcal{U}, t \in T = [0,T]\}, \\ \mathbf{A} \in \mathbb{R}^{2\times 2}, \quad \mathbf{x} \in \mathbb{R}^2, \quad \mathbf{u} \in \mathbb{R}^2, \quad \mathbf{b} \in \mathbb{R}^2, \quad \mathbf{c} \in \mathbb{R}^2, \end{gathered} \tag{2.1}$$

где $\langle \cdot, \cdot \rangle$ – скалярное произведение.

Пусть множество $\mathcal{U}$ допустимых значений управления является компактом в $\mathbb{R}^n$. Пусть матричная функция $\mathbf{A}(u, t)$ и вектор-функция $\mathbf{b}(\mathbf{u}, t)$ непрерывны на произведении $\mathcal{U} \times \mathcal{T}$.

Рассмотрим функцию Гамильтона–Понтрягина для (2.1):

$$H(\psi, \mathbf{x}, \mathbf{u}, t) = \langle \psi, \mathbf{A}^{\mathrm{T}}(\mathbf{u}, t)\mathbf{x} + b(\mathbf{u}, t) \rangle,$$

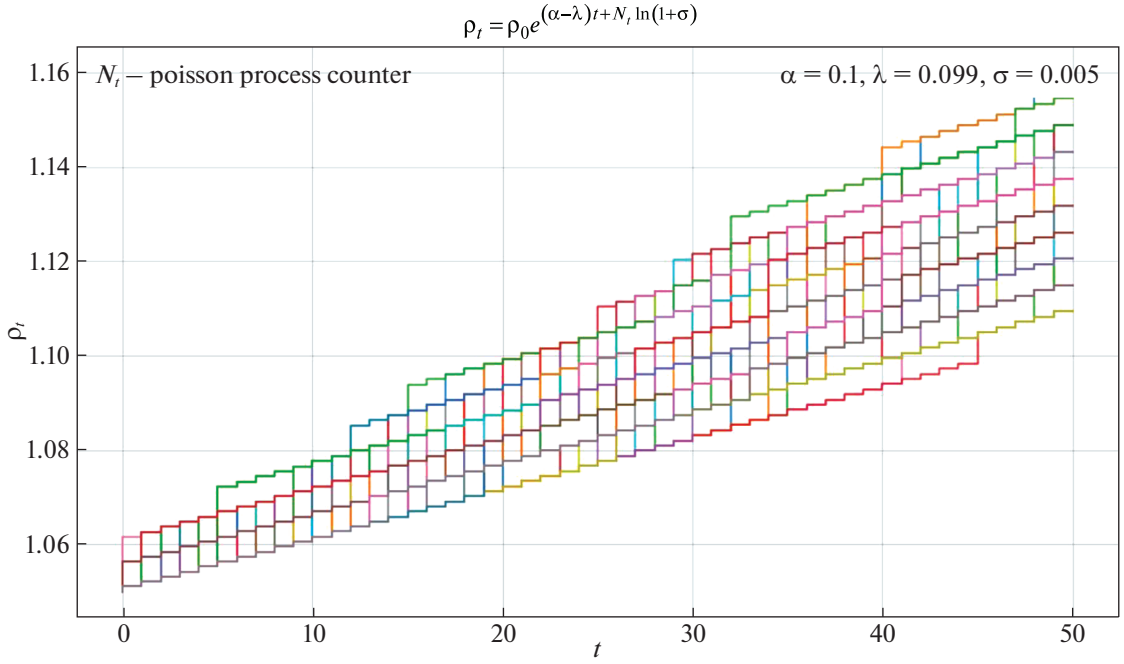


**Рис. 3.** Коэффициент прироста капитала как пуассоновский процесс (16 реализаций)

а также сопряженную систему:

$$\dot{\psi} = -H_x(\psi, \mathbf{x}, \mathbf{u}, t) = -\mathbf{A}^{\mathrm{T}}(\mathbf{u}, t)\psi, \qquad \psi(t_1) = -\mathbf{c},$$

где $\psi$ и $\mathbf{c}$ – двумерные векторы, причем $\mathbf{c} = (1, 1)^{\mathrm{T}}$.

Чтобы допустимое управление $u(t)$, $t \in \mathbf{T}$ было оптимально в задаче (2.1), необходимо, чтобы всюду на $\mathbf{T}$ выполнялось условие

$$\mathbf{u}(t) = \arg\max_{v \in \mathcal{U}} H(\psi(t, \mathbf{u}), \mathbf{x}(t, \mathbf{u}), v, t).$$

Для задачи (1.1) получаем окончательно

$$H = a_1 y_1 \psi_1(t) + a_2 y_2 \psi_2(t),$$

$$\dot{\psi}_1 = \frac{\partial H}{\partial x_1} = 0,$$

$$\dot{\psi}_2 = \frac{\partial H}{\partial x_2} = 0.$$

Поэтому

$$\frac{\partial H}{\partial x_1} = \frac{\partial H}{\partial x_2} = 0,$$

$$\psi = \mathrm{const},$$

$$\psi(t_1) = -\mathbf{c} = \mathbf{1},$$

где $\mathbf{1}$ – единичный вектор.

Решение выглядит так:

$$u_1^* = \begin{cases} W, & a_1(t) \geq a_2(t), \\ 0, & a_1(t) < a_2(t), \end{cases}$$
$$u_2^* = \begin{cases} 0, & a_1(t) \geq a_2(t), \\ W, & a_1(t) < a_2(t). \end{cases} \qquad (2.2)$$

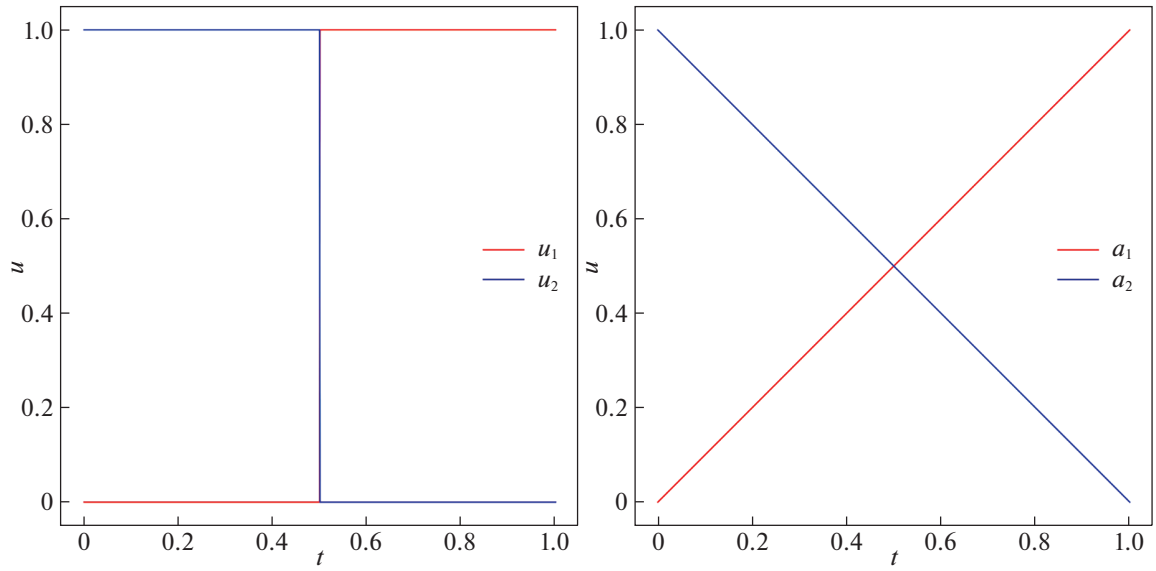


**Рис. 4.** Аналитическое решение линейной задачи оптимального управления с одной точкой переключения

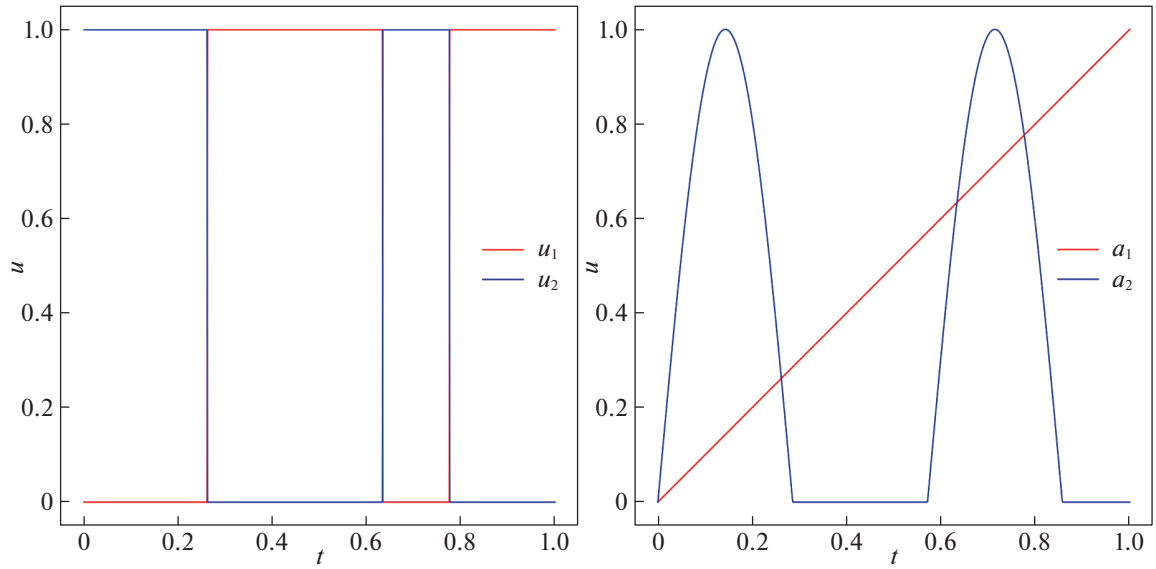


**Рис. 5.** Аналитическое решение линейной задачи оптимального управления с тремя точками переключения

Если $a_1(t) = t$, $a_2(t) = 1 - t$, $t \in [0, 1]$, то решение иллюстрирует рис. 4. Имеем одну точку переключения. Если $a_1(t) = t$, $a_2(t) = \sin(3.5\pi t)$, $t \in [0, 1]$, то решение иллюстрируется с помощью рис. 5. Имеем три точки переключения.

2.2. Задача оптимального потребления. Безусловно, эту задачу можно было решать принципом максимума, но можно и принципом Беллмана.

Для нас самое важное то, что нейросети получат то же самое решение, если мы используем обучение с подкреплением.

Дискретизация (1.2) по времени дает:

$$\begin{gathered}\sum_{t=0}^{T}\beta^t \ln(u_t) \to \max_{u_t},\\ x_{t+1} = \alpha x_t - u_t,\\ x_0 = a,\\ x_T = 0.\end{gathered} \tag{2.3}$$

Найдем решение с помощью принципа оптимальности Беллмана (см., например, в [14]).

Функция Беллмана имеет вид, где $k$ – текущий номер:

$$S_k(x) = \max_{u_t, t=k,T} \sum_{t=k}^{T} \beta^t \ln u_t$$

при условиях

$$x_{t+1} = \alpha x_t - u_t,$$
$$x_k = x.$$

Запишем рекуррентное соотношение Беллмана

$$S_t(x) = \max_{u_t} \{\beta^t \ln(u_t) + S_{t+1}(\alpha x - u_t)\}$$

с терминальным значением

$$S_T(x) = \begin{cases} 0, & x \geq 0, \\ -\infty, & x < 0. \end{cases}$$

Имеем

$$u_k(x) = \frac{\alpha x}{\sum\limits_{i=0}^{T-k} \beta^i}. \tag{2.4}$$

В самом деле, подставим решение в правую часть уравнения Беллмана:

$$S_j(x) = \max_{u_j \geq 0} \left[ \beta^j \ln(u_j) + \beta^{j+1} \sum_{t=0}^{T-j-1} \left\{ \beta^t \ln \left[ \frac{\alpha(\alpha x - u_j)(\alpha\beta)^t}{\sum\limits_{i=0}^{T-j-1} \beta^i} \right] \right\} \right]. \tag{2.5}$$

Максимум получим дифференцированием правой части (2.5):

$$\frac{\beta^j}{u_j} - \beta^{j+1} \sum_{t=0}^{T-j-1} \frac{\beta^t}{\alpha x - u_j} = 0 \Rightarrow u_j = \frac{\beta^j \alpha x}{\beta^j + \beta^{j+1} \sum\limits_{t=0}^{T-j-1} \beta^t} = \frac{\alpha x}{\sum\limits_{t=0}^{T-j} \beta^t}. \tag{2.6}$$

Знак второй производной проверяется аналогично.

При подстановке (2.6) в правую часть соотношения (2.5) получаем

$$S_j(x) = \beta^j \ln \left[ \frac{\alpha x}{\sum\limits_{i=0}^{T-j} \beta^i} \right] + \beta^{j+1} \sum_{t=0}^{T-j-1} \left\{ \beta^t \ln \left[ \alpha \left( \alpha x - \frac{\alpha x}{\sum\limits_{i=0}^{T-k} \beta^i} \right) (\alpha\beta)^t \left( \sum_{i=0}^{T-j-1} \beta^i \right)^{-1} \right] \right\} =$$

$$= \beta^j \ln \left[ \frac{\alpha x}{\sum\limits_{i=0}^{T-j} \beta^i} \right] + \beta^{j+1} \sum_{t=0}^{T-j-1} \left\{ \beta^t \ln \left[ \frac{x\alpha^2 (\alpha\beta)^t \beta}{\sum\limits_{i=0}^{T-j} \beta^i} \right] \right\} = \beta^j \ln \left[ \frac{\alpha x}{\sum\limits_{i=0}^{T-j} \beta^i} \right] + \beta^{j+1} \sum_{u=0}^{T-j} \left\{ \beta^{u-1} \ln \left[ \frac{x\alpha^2 (\alpha\beta)^{u-1} \beta}{\sum\limits_{i=0}^{T-j} \beta^i} \right] \right\} =$$

$$= \beta^j \ln \left[ \frac{\alpha x}{\sum\limits_{i=0}^{T-j} \beta^i} \right] + \beta^j \sum_{u=0}^{T-j} \left\{ \beta^u \ln \left[ \frac{x\alpha (\alpha\beta)^u}{\sum\limits_{i=0}^{T-j} \beta^i} \right] \right\} = \beta^j \sum_{t=0}^{T-j} \left\{ \beta^t \ln \left[ \frac{\alpha x (\alpha\beta)^t}{\sum\limits_{i=0}^{T-j} \beta^i} \right] \right\}.$$

Решение уравнения Беллмана запишем как

$$S_k(x) = \beta^k \sum_{t=0}^{T-k} \left\{ \beta^t \ln \left[ \frac{\alpha x (\alpha\beta)^t}{\sum_{i=0}^{T-k} \beta^i} \right] \right\}. \tag{2.7}$$

**3. Построение решений с помощью обучения с подкреплением.** В работе основой для построения численного решения задач оптимального управления служит алгоритм Proximal Policy Optimization [5]. Вводится ряд понятий.

О п р е д е л е н и е 2. Марковский процесс принятия решений: $\langle S, A, P, R, \gamma \rangle$, где $S$ – пространство состояний; $A$ – пространство действий, $P^a_{ss'} = \mathbb{P}(s_{t+1} = s' | s_t = s, a_t = a)$ – динамика среды или вероятность перехода агента из состояния $s$ в состояние $s'$ при условии действия $a$ в момент времени $t$; $R^a_s = \mathbb{E}[r_t | s_t = s, a_t = a]$ – функция вознаграждения. Она характеризует среднее вознаграждение агента в состоянии $s$ при условии действия $a$ в момент времени $t$; $\gamma \in (0, 1]$ – дисконтирующий фактор.

О п р е д е л е н и е 3. Стратегия агента:

$$\pi_\theta(a \mid s) = \mathbb{P}(a_t = a \mid s_t = s, \theta).$$

О п р е д е л е н и е 4. Награда за эпизод:

$$r_t = \mathbb{E}[r \mid a_t = a, s_t = s],$$

$$G_t = r_{t+1} + \gamma r_{t+2} + \ldots + \gamma^{T-t-1} r_T = \sum_{t'=t}^{T-1} \gamma^{t'=t} r_{t'+1}.$$

О п р е д е л е н и е 5. Полезность состояния:

$$V(s) = \mathbb{E}_\pi[G_t \mid s_t = s].$$

О п р е д е л е н и е 6. Полезность состояния-действия:

$$Q(a, s) = \mathbb{E}_\pi[G_t \mid s_t = s, a_t = a].$$

О п р е д е л е н и е 7. Функция преимущества:

$$A(a, s) = Q(a, s) - V(s).$$

3.1. Д в о й с т в е н н а я з а д а ч а о п т и м и з а ц и и. Определим отношение вероятностей $r_t(\theta)$ как

$$r_t(\theta) = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)}, \quad r(\theta_{old}) = 1,$$

где $\theta_{\text{old}}$ – параметры текущей стратегии.

TRPO [2] максимизирует суррогатный функционал:

$$L^{CLI}(\theta) = \hat{\mathbb{E}}_t \left[ \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{old}}(a_t \mid s_t)} \hat{A}_t \right],$$

где $\hat{\mathbb{E}}_t$ – математическое ожидание по времени, а $\hat{A}_t$ – функция преимущества в момент времени $t$.

Без дополнительных ограничений максимизация $L^{CLI}$ приведет к чрезмерно большому обновлению стратегии. Введем дополнительный штраф за изменения стратегии, которые удаляют $r_t(\theta)$ от 1.

Двойственная задача оптимизации

$$L^{CLIP}(\theta) = \hat{\mathbb{E}}_t \left[ \min\{r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1 - \epsilon, 1 + \epsilon)\hat{A}_t\} \right],$$

$$L^{CLIP}(\theta) \to \max_{\theta \in \mathbb{R}^n},$$

где $\epsilon$ – гиперпараметр, а

$$\mathrm{clip}(x,a,b) = \begin{cases} a, & \text{если} \quad x < a, \\ x, & \text{если} \quad a \le x \le b, \\ b, & \text{если} \quad x > b. \end{cases}$$

3.2. Proximal Policy Optimization (PPO). В [5] PPO содержит две независимые параметризации для актора и критика. Процесс принятия решений обеспечивается актором, который имеет параметризацию $\theta$. Процедура оценки полезности действий актора осуществляется критиком, который имеет параметризацию $\phi$.

Веса актора обновляются максимизацией функции преимущества по парам состояние-действие для полученных траекторий. Веса критика обновляются минимизацией среднеквадратичной ошибки между отложенными дисконтированными значениями вознаграждения и параметризованной функцией полезности.

**4. Результаты вычислений.** Для решения поставленных задач наряду с PPO применялись различные алгоритмы обучения с подкреплением семейства Policy Gradient, а именно TRPO [2], DDPG [1], SAC [3], A2C [4].

Для проведения экспериментов использовались фреймворки обучения с подкреплением RLlib и stablebaseline3.

4.1. Линейная задача оптимального управления (1.1). Аналитическое решение задачи (1.1) задается формулой (2.2). Для задачи (1.1) применяем дискретизацию по времени:

$$\mathcal{F}(\mathbf{u}) = x_K^1 + x_K^2 \to \max,$$
$$x_{k+1}^1 - x_k^1 = a_k^1 u_k^1 \Delta t, \quad x_{k+1}^2 - x_k^2 = a_k^2 u_k^2 \Delta t,$$
$$u_k^1 \ge 0, \quad u_k^2 \ge 0, \quad x_0 = 0,$$
$$u_k^1 + u_k^2 \le W, \quad k \in \overline{0,K},$$
$$x_k = x(t_k), \quad k \in \overline{0,K}, \quad t_{k+1} - t_k = \Delta t.$$

Использование алгоритмов обучения с подкреплением требует формулировки задачи оптимального управления на языке принятия решений. Марковский процесс принятия решений подразумевает задание пространства действий, пространства состояний, функции вознаграждения, динамики среды и дисконтирующего фактора.

Под состоянием будем понимать вектор из четырех компонент:

$$s = (a_t^1, a_t^2, t, T - t),$$

где $a_t^1$, и $a_t^2$ – значения динамических коэффициентов в текущий момент времени, $t$ – текущее время с момента начала эпизода и $T - t$ – оставшееся время до конца эпизода.

Остановимся на функции вознаграждения. Ввиду специфики задачи оптимизации и граничных условий наиболее естественным будет определить функцию вознаграждения для марковского процесса принятия решений следующим образом. Выразим $x_1(t)$ и $x_1(t)$ из уравнения (1.1):

$$x_1(t) = \int_0^t u_1(t) a_1(t) dt,$$

$$x_2(t) = \int_0^t u_2(t) a_2(t) dt.$$

Проведем дискретизацию по времени ($x_k = x(t_k)$, $k \in \overline{0,K}$, $t_{k+1} - t_k = \Delta t$):

$$x_t^1 = \sum_{i=0}^t u_i^1 a_i^1 \Delta t,$$

$$x_t^2 = \sum_{i=0}^t u_i^2 a_i^2 \Delta t.$$

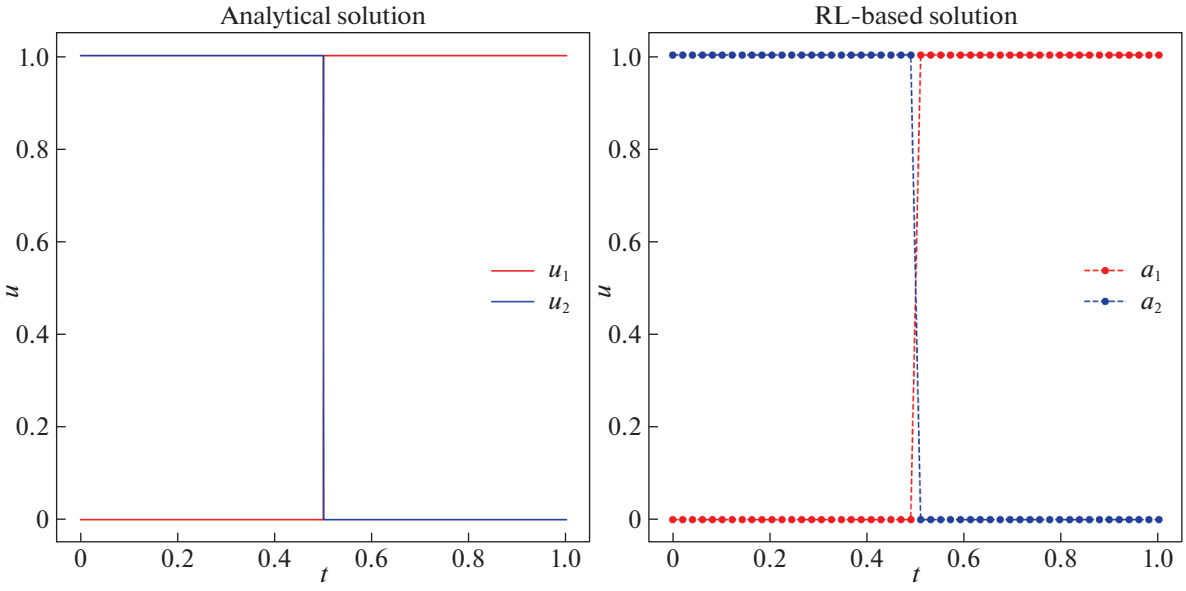


**Рис. 6.** Аналитическое и численное решения линейной задачи оптимального управления при $a_1(t) = t$, $a_2(t) = 1 - t$

Запишем мгновенные вознаграждения в случае действия $[u_t^1, u_t^2]^{\mathrm{T}}$ в момент времени $t$:

$$r_t^1 = u_t^1 a_t^1, \quad r_t^2 = u_t^2 a_t^2,$$
$$r_t = r_t^1 + r_t^2.$$

Рассмотрим пространство действий

$$u = [u^1, u^2]^{\mathrm{T}},$$
$$u^i \in [\epsilon, +\infty).$$

Для реализации ограничений на допустимые значения управления $u_t^1 + u_t^2 \le W$ существует два основных приема, а именно маскирование заведомо нерелевантных действий или с помощью штрафа за нарушение ограничений. В данном случае нарушение условия $u_t^1 + u_t^2 \le W$ влечет за собой наложение штрафа и завершение эпизода. Штраф полагается равным максимально возможной награде за эпизод. Подобный сценарий форсирует агента на использование значений действия из допустимого множества и позволяет ему избегать нарушения поставленных ограничений.

Рассмотрим дисконтирующий фактор. Среда эпизодическая, поэтому необходимости в явном использовании $\gamma$ нет. Здесь и далее $y = 1$, т.е. дисконтирование отсутствует. Динамика среды представляет собой детерминированную функцию, которая не зависит от принятого агентом действия:

$$s_{t+1} = (a_{t+1}^1, a_{t+1}^2, t+1, T-t-1).$$

Результаты работы PPO таковы. Рассмотрим задачу распределения ресурсов. Согласно принципу максимума Понтрягина, аналитическое решение имеет вид ступенчатых функций:

$$u_1^* = \begin{cases} 1, & t \ge 0.5, \\ 0, & t < 0.5, \end{cases}$$
$$u_2^* = \begin{cases} 0, & t \ge 0.5, \\ 1, & t < 0.5. \end{cases}$$

На рис. 6 получено совпадение с этим решением, когда применяется обучение с подкреплением, на рис. 7 – то же самое для трех точек переключения.

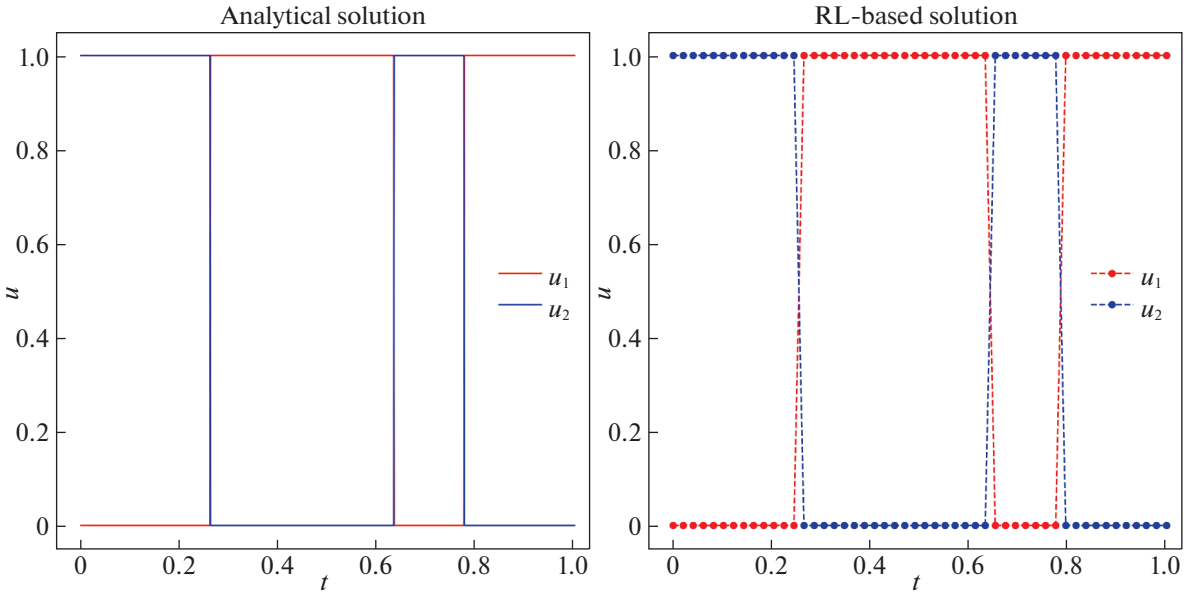


**Рис. 7.** Аналитическое и численное решения линейной задачи оптимального управления при $a_1(t) = t$, $a_2(t) = \max\{\sin(3.5\pi t), 0\}$

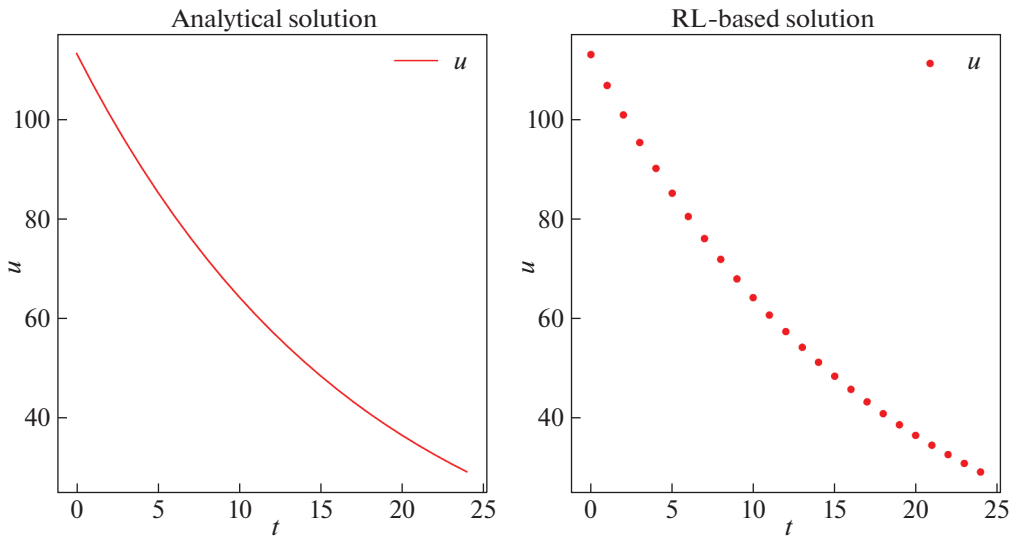


**Рис. 8.** Управление для задачи оптимального потребления

Проанализировав полученные решения, приходим к выводу, что решения задач оптимального управления, найденные с помощью обучения с подкреплением, совпадают с аналитическими с точностью до дискретизации.

4.2. Задача оптимального потребления (1.2). Аналитическое решение задачи (1.2) задается формулой (2.7). Под состоянием будем понимать вектор из трех компонент:

$$s = (c, t, T - t),$$

где $c$ – значение доступного капитала в текущий момент времени, $t$ – текущее время с момента начала эпизода и $T - t$ – оставшееся время до конца эпизода.

Определим дисконтирующий фактор как $\gamma = \beta$, а вознаграждение – как полезность потребления в текущий момент времени.

$$r_t = \log(u_t).$$

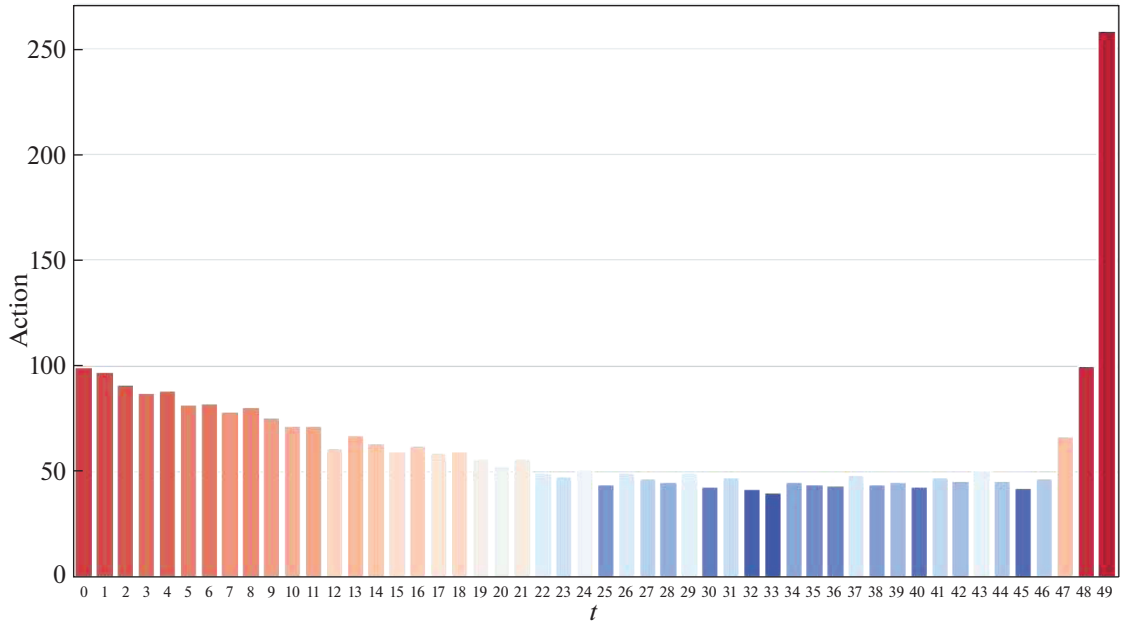


**Рис. 9.** Управление для задачи оптимального потребления, где коэффициент прироста капитала представлен как дискретное распределение вероятностей

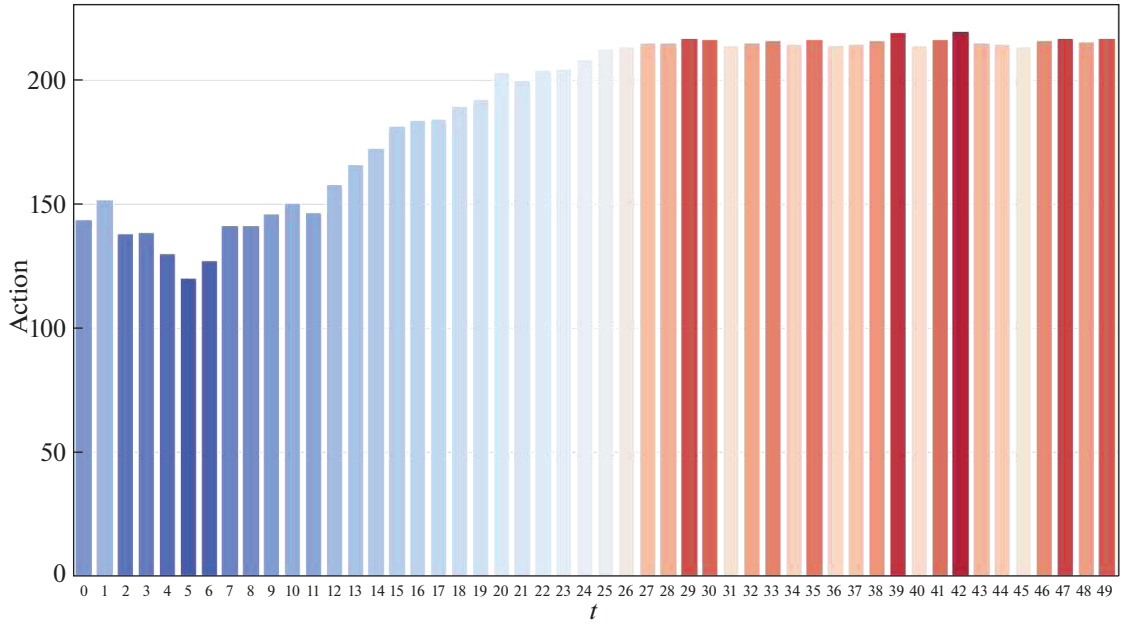


**Рис. 10.** Управление для задачи оптимального потребления, где коэффициент прироста капитала представлен как диффузионный процесс

Складывая все дисконтированные вознаграждения в каждый момент времени, получим итоговое дисконтированное значение вознаграждения за эпизод, которое совпадает со значением функционала:

$$G_0 = \sum_{i=0}^{T-1} \gamma^t r_t = \sum_{i=0}^{T-1} \beta^t \log(u_t).$$

В отличие от линейной задачи оптимального управления динамика среды для задачи оптимального потребления обладает более сложной структурой и зависит от действия,

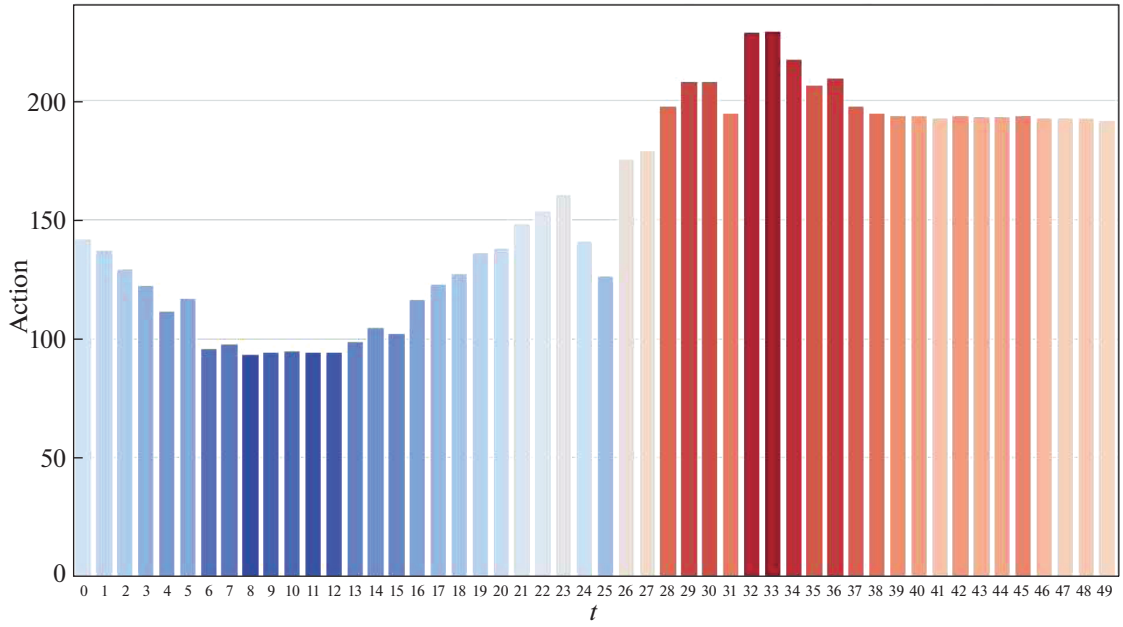


**Рис. 11.** Управление для задачи оптимального потребления, где коэффициент прироста капитала представлен как пуассоновский процесс

предпринятого на выбранном шаге. Тем не менее динамика среды остается детерминированной функцией.

Применяя оптимизацию градиента стратегии, PPO обеспечивает абсолютную сходимость численного решения при фиксированном исходном значении капитала $x$ в (2.4) (см. рис. 8) к аналитическому с точностью до дискретизации.

4.3. Задача стохастического кредитования (1.3). Эта задача сложнее предыдущих из-за недетерминированности среды. Данная задача не имеет аналитического решения, поэтому не удается сравнить с эталонным.

Марковский процесс принятия решений для задачи стохастического кредитования задается аналогично детерминированному случаю. Ниже представлены решения трех типов задач (1.4)–(1.6) стохастического кредитования, полученные с использованием алгоритма оптимизации градиента стратегии. Для наших трех случаев решения представлены на рис. 9–11. В данном случае удалось добиться сходимости алгоритмов обучения с подкреплением.

**Заключение.** Итак, обучение с подкреплением строит решения для детерминированных задач оптимального управления. Они совпадают с найденными посредством классических алгоритмов. Также получаются разрывные управления. Для стохастических задач оптимального управления данный подход дает сходящуюся последовательность. Однако не следует забывать, что количество итераций при обучении может быть очень большим (в нашем случае – несколько миллионов). Это имеет место, в частности, при росте количества точек дискретизации. Метод обучения с подкреплением демонстрирует свою эффективность для широкого класса задач. Тем не менее всегда нужно выбирать, когда он успешно конкурирует с традиционными аналитическими и численными методами.

## СПИСОК ЛИТЕРАТУРЫ

1. *Sewak M.* Deterministic Policy Gradient and the DDPG: Deterministic-Policy-Gradient-Based Approaches. 2019.
2. *Schulman J.* Trust Region Policy Optimization. 2015. https://arxiv.org/abs/1502.05477.
3. *Haarnoja T.* Soft Actor-Critic: Off-Policy Maximum Entropy Deep Rein-forcement Learning with a Stochastic Actor. 2018. https://arxiv.org/abs/1801.01290.
4. *Huang S.* A2C is a special case of PPO. 2022. https://arxiv.org/abs/2205.09123.
5. *Schulman J.* Proximal Policy Optimization Algorithms. 2017. https://arxiv.org/abs/1707.06347.

6. *Zhang L.* Penalized Proximal Policy Optimization for Safe Reinforcement Learning. 2022. https://arxiv.org/abs/2205.11814.
7. *Chen X.* The Sufficiency of Off-policyness: PPO is insufficient according to an Off-policy Measure. 2022. https://arxiv.org/abs/2205.10047.
8. *Ghosh A.* Provably Efficient Model-Free Constrained RL with Linear Function Approximation. 2022. https://arxiv.org/abs/2206.11889.
9. *Song Z.* Safe-FinRL: A Low Bias and Variance Deep Reinforcement Learning Implementation for High-Freq Stock Trading. 2022. https://arxiv.org/abs/2206.05910.
10. *Kaledin M.* Variance Reduction for Policy-Gradient Methods via Empirical Variance Minimization. 2022. https://arxiv.org/abs/2206.06827.
11. *Luo Q.* Finite-Time Analysis of Fully Decentralized Single-Timescale Actor- Critic. 2022. https://arxiv.org/abs/2206.05733.
12. *Deka A.* ARC – Actor Residual Critic for Adversarial Imitation Learning. 2022. https://arxiv.org/abs/2206.02095.
13. *Цурков В.И.* Динамические задачи большой размерности. М.: Наука, 1988. 287 с.
14. *Бекларян Л.А., Флёрова А.Ю., Жукова А.А.* Методы оптимального управления: учеб. пособие. М.: Наука, 2018.
15. *Оксендаль Б.* Стохастические дифференциальные уравнения. Введение в теорию и приложеия. М.: Мир, 2003.
16. *Понтрягин Л.С.* Принцип максимума в оптимальном управлении. М.: Наука, 2004.