ВЫЧИСЛИТЕЛЬНЫЕ И ИНФОРМАЦИОННО-ИЗМЕРИТЕЛЬНЫЕ СИСТЕМЫ

УДК 004.942, 004.896

ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ В СИСТЕМАХ УПРАВЛЕНИЯ ОБЪЕКТАМИ С ТРАНСПОРТНЫМ ЗАПАЗДЫВАНИЕМ

© **В. С. Боровик**[1], **С. В. Шидловский**[1, 2]

[1] *Национальный исследовательский Томский государственный университет, 634050, г. Томск, просп. Ленина, 36*
[2] *Национальный исследовательский Томский политехнический университет, 634050, г. Томск, просп. Ленина, 30*
E-mail: borovik_vasiliy@mail.ru

Обсуждается возможность применения систем обучения с подкреплением для решения задач регулирования в условиях недостатка априорной информации об объекте управления. Представлено решение проблемы обучения системы методом Deep Deterministic Policy Gradient для объектов с транспортным запаздыванием, а также сравнение эффективности предлагаемого решения с классическим методом на основе ПИД-регулирования, параметры которого рассчитаны с применением метода расширенных амплитудно-фазочастотных характеристик и метода Циглера — Никольса.

Ключевые слова: обучение с подкреплением, DDPG, система управления, моделирование, ПИД-регулятор, формирование управляющих воздействий, управление в условиях недостатка априорной информации.

DOI: 10.15372/AUT20210306

Введение. Основной задачей любой автоматизированной системы управления является процесс поддержания постоянства величин, которые характеризуют протекание технологического процесса в соответствии с принятым законом регулирования [1]. Задача может осложняться как проявлением внешних и параметрических воздействий в системе, так и недостаточностью априорной информации об объекте управления (наличие транспортного запаздывания), что негативно сказывается на качестве регулирования из-за невозможности учёта этих факторов на этапе конфигурирования системы. Поэтому к таким системам применение типовых регуляторов (П, И, ПИ, ПИД) оказывается неэффективным [2].

Возможным решением данной проблемы может стать использование систем регулирования, основанных на нейронных сетях, что позволит добиться оптимального уровня контроля технологического процесса, а также повысить показатели надёжности и экономичности систем.

Благодаря разнообразию архитектур и подходов к обучению, возможности анализа текущего состояния системы и выбору наилучшей стратегии поведения, выработанной в ходе процесса обучения, нейросети уже зарекомендовали себя в решении задач классификации, прогнозирования и управления. Такие системы способны к обработке большого потока данных в реальном времени с сохранением требуемой скорости работы, которая определяется сложностью архитектурной модели сети [3–5].

Выделяют такие свойства нейронных сетей, как обобщение и абстрагирование. Обобщение даёт возможность сети анализировать информацию даже при наличии шумов или нечувствительности окружающей среды. Абстрагирование дополняет это свойство, позволяя извлечь сущность из входных сигналов даже при их искажении. Всё это делает нейросеть потенциально эффективной регулирующей системой, способной адаптироваться к

изменению условий протекания процесса в условиях недостатка априорной информации о параметрах объекта [6].

Существует несколько разновидностей машинного обучения. В рамках данного исследования выбран алгоритм обучения с подкреплением, в основе которого лежит процесс взаимодействия обучаемой системы с моделируемой средой посредством формирования сигналов подкрепления на отклики среды. Данный метод из-за особенностей своей архитектуры успешно применяется в робототехнике, телекоммуникации и игровой индустрии [7].

Постановка задачи. В качестве объекта управления рассматриваются объекты с передаточной функцией вида

$$W_{\text{об}}(p) = \frac{k\,\mathrm{e}^{-\tau p}}{Tp+1}, \tag{1}$$

где k — коэффициент передачи объекта, T — постоянная времени, τ — величина запаздывания.

Предполагается, что на объект управления действует параметрическое n и внешнее координатное возмущение f. Координатное возмущение — это неизвестная величина со стороны нагрузки на объект управления, которая проявляется в виде неконтролируемых произвольных изменений технологических параметров и по характеру изменения во времени может быть импульсной и медленно меняющейся. Параметрическое возмущение есть неизвестная величина из некоторого ограниченного множества, в результате действия которой происходит медленное изменение параметров объекта управления.

Ставится задача выбора такого управления u, при котором выходное значение объекта управления совпадало бы с задающим значением или их разница была в допустимых пределах при изменении воздействий f и n. При этом ограничение на управляющее воздействие не накладывается. Под влиянием внешних возмущений, информации о которых часто недостаточно, взаимосвязь между входом и выходом объекта становится неоднозначной и неопределённой, что сильно затрудняет решение задачи [8, 9].

Обучение с подкреплением. Основой данного метода является изучение взаимодействия агента и окружения при условии стохастичности параметров процесса. Конечная цель обучения — выработка стратегии, позволяющей достичь объектом управления требуемого состояния.

Алгоритм обучения описывается множеством состояний окружения S и множеством возможных действий агента A. Поведение системы определяется функцией вероятности переходов $\pi(s)$: $S \times A \to [0, 1]$.

В момент времени t агент имеет информацию о состоянии среды $s_t \in S$ и в соответствии с выбранной стратегией поведения $\pi(s)$, которую также можно представить в параметрическом виде:

$$\pi(\theta) = P(a \mid s, \theta), \tag{2}$$

совершает действие $a_t \in A$, в результате чего состояние меняется на s_{t+1}, а агент получает награду r_t.

Алгоритм повторяется до тех пор, пока не найдётся оптимальная стратегия управления, при которой происходит максимизация итоговой величины награждения с учётом параметра дисконтирования $\gamma^k \in [0; 1)$ [10]:

$$J(\theta) = E_{\pi(\theta)}\Big[\sum_{k=0}^{\infty} \gamma^k r_{t+k}\Big] \to \max. \tag{3}$$

При агентном подходе обучения вводится функция полезности (Q-функция) [10]:

$$Q^{\pi}(s,a) = E_{\pi}\Big[\sum_{k=0}^{\infty}\gamma^{k} r_{t+k+1} \mid s_t = s,\ a_t = a\Big]. \tag{4}$$

Оптимальная стратегия поведения системы $\pi(s_t)$ определяет, насколько выгодно агенту выбрать действие a_t на основании состояния s_t [10]:

$$\pi(s_t) = \operatorname{argmax}_{a_t} Q(s_t, a_t).$$

В рамках данной работы выбран алгоритм Deep Deterministic Policy Gradient (DDPG), в основе которого лежит обучение Q-функции (4) критиком, а обучение $\pi(\theta)$ (2) — агентом.

В общем случае алгоритм имеет следующую структуру [10, 11]:

1. Инициализация критика $Q_{\theta^Q}(s,a)$ с весом θ^Q и агента $\pi_{\theta^\pi}(s)$ с весом θ^π.
2. Инициализация Q' с весом $\theta^{Q'} = \theta^Q$ и π' с весом $\theta^{\pi'} = \theta^\pi$.
3. Инициализация буфера B.
4. Инициализация случайного процесса P_t из конечного множества P.
5. Выбор действия $a_t = \pi(s_t) + P_t$.
6. Реализация действия a_t, получение вознаграждения r_t, переход в s_{t+1}.
7. Сохранение параметров (s_t, a_t, r_t, s_{t+1}) в B.
8. Реализация N выборок (s_i, a_i, r_i, s_{i+1}) из B.
9. Вычисление y_i:

$$y_i = r_i + \gamma Q'(s_{i+1}, \pi'(s_i + 1)).$$

10. Обновление критика минимизацией потерь:

$$L = \frac{1}{N}\sum_{i}^{N}(y_i - Q_{\theta^Q}(s_i, a_i))^2.$$

11. Обновление агента с использованием выборочного градиента политики:

$$\nabla_{\theta^\pi} J \approx \frac{1}{N}\sum_{i}^{N} \nabla_a Q(s,a) \mid_{s=s_i,\ a=\pi(s_i)} \nabla_{\theta^\pi}\pi(s) \mid_{s=s_i}.$$

12. Обновление весов с использованием коэффициента сглаживания τ:

$$\theta^{Q'} = \tau\theta^Q + (1-\tau)\theta^{Q'}, \qquad \theta^{\pi'} = \tau\theta^\pi + (1-\tau)\theta^{Q\pi'}.$$

13. Возврат на шаг 4.

Структурная схема, описывающая применяемую модель обучения с подкреплением, представлена на рис. 1, где v — задающее воздействие, e — ошибка регулирования, s_t — состояние объекта, u — управляющее воздействие, w — сигнал на выходе упредителя Смита, f — возмущающее воздействие, y — выходная регулируемая величина.

Система охвачена отрицательной обратной связью для учёта текущего отклонения регулируемой величины блоком формирования состояния окружения, который также принимает на вход задающее воздействие v. Как было показано в [12], из-за отсутствия начального отклика системы применение алгоритма DDPG для управления объектами с

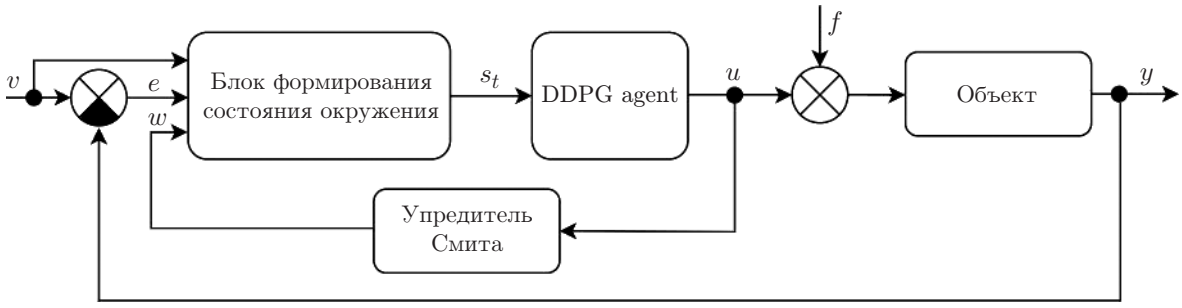

Рис. 1. Структурная схема модели обучения с подкреплением

запаздыванием является затруднительным и требует корректировки алгоритма обучения или внесения в модель структурных изменений [13, 14]. Важный фактор при работе с нейронными сетями — формирование уникальных признаков, на основе которых возможно было бы обучение системы. Поэтому в работе предлагается ввести в блок формирования состояния окружения дополнительную компоненту w. Сама же компонента w формируется на выходе упредителя Смита с передаточной функцией вида

$$W_{\text{УС}}(p) = W^*_{\text{об}}(p)(1 - \mathrm{e}^{-\tau p}), \tag{5}$$

где $W^*_{\text{об}}(p)$ — передаточная функция объекта без запаздывания.

Таким образом, величина s_t формируется благодаря наблюдению за ошибкой, её интегралом, задающим воздействием и выходным значением упредителя Смита, или, если представить это в векторной форме, $\left[\int e\,dt \;\; e \;\; v \;\; w\right]^\top$. В дальнейшем данная величина учитывается в блоке DDPG agent, предназначенном для формирования управляющего воздействия u согласно выработанной стратегии с учётом наличия в канале управления возмущающего воздействия f.

Величину вознаграждения r_t можно описать следующим соотношением:

$$r_t = 10(|e| < 0,1) - 1(|e| < 0,1) - 100(v \leqslant 0 \;||\; v \geqslant 15).$$

Подробная реализация DDPG-модели представлена в [11].

Моделирование системы регулирования. Для иллюстрации описанного подхода рассмотрим объект с передаточной функцией вида (1), имеющий следующие параметры: $k = 6$; $T = 97,7$ с; $\tau = 5$ с. Параметрическое возмущение n вызывает изменение постоянной времени T, при этом в качестве её возможных значений с учётом неизменности величины запаздывания примем диапазон $9 < T/\tau < 30$.

Для оценки эффективности предлагаемого решения была создана модель классической системы управления (рис. 2, *a*) на основе ПИД-регулятора при двух различных вариациях настроечных параметров, а также модель системы управления на основе ПИД-регулятора с применением упредителя Смита (5) (рис. 2, *b*), где v — задающее воздействие, e — ошибка регулирования, u — управляющее воздействие, f — возмущающее воздействие, y — выходная регулируемая величина.

Настройка ПИД-регулятора. В первом случае расчёт параметров производился с помощью корневого метода параметрического синтеза, основанного на понятии расширенных амплитудно-фазочастотных характеристик (РАФЧХ). Во втором случае использовался метод незатухающих колебаний (Циглера — Никольса) [15, 16]. При определении

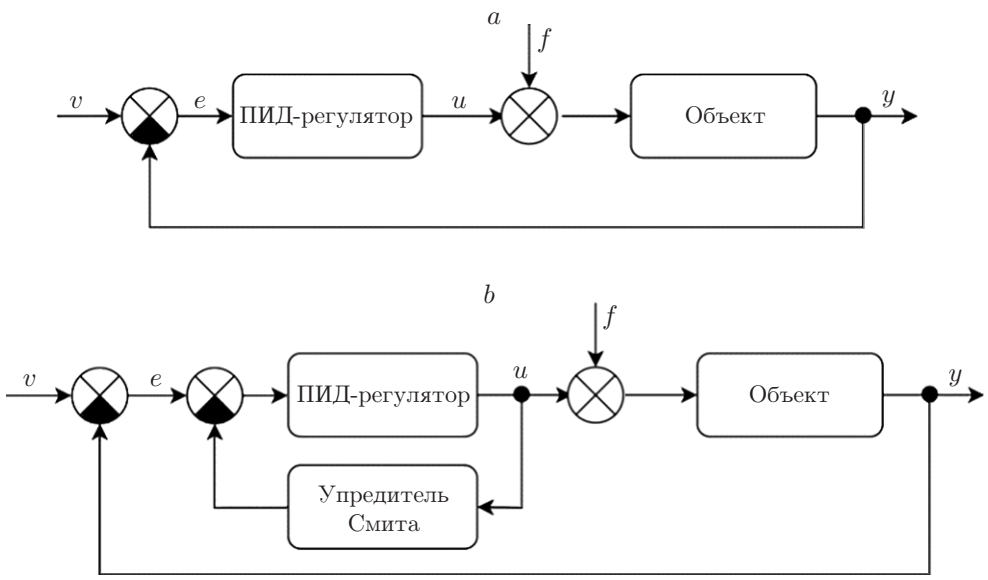

Рис. 2. Структурная схема системы управления с ПИД-регулятором по каналу $f(t)-y(t)$: классическая система управления (a), система управления с применением упредителя Смита (b)

настроек регулятора в качестве показателя оптимальности системы регулирования был выбран второй интегральный критерий качества

$$I_2 = \int_0^\infty |\varepsilon(t)|\, dt, \tag{6}$$

где $\varepsilon(t)$ — ошибка регулирования.

Для нахождения значения параметров регулятора корневым методом необходимо выразить оператор p в передаточной функции объекта $W_{\text{об}}(p)$ как $p = (i-m)\omega$ или $p = -\eta+i\omega$, где m — степень колебательности, η — степень устойчивости, ω — частота. Затем применяются следующие расчётные формулы для ПИД-регулятора [15]:

$$C_0 = \frac{K_p}{T_{\text{и}}} = \omega(m^2+1)\Big[\omega C_2 - \frac{\text{Im}_{\text{об}}(m,\omega)}{A^2_{\text{об}}(m,\omega)}\Big],$$

$$C_1 = -\frac{m\,\text{Im}_{\text{об}}(m,\omega) + \text{Re}_{\text{об}}(m,\omega)}{A^2_{\text{об}}(m,\omega)} + 2\omega m C_2, \qquad C_2 = K_p T_{\text{д}},$$

где K_p — коэффициент пропорциональности; $T_{\text{и}}$ — постоянная времени интегрирования; $A_{\text{об}}(m,\omega)$ — амплитуда; $T_{\text{д}}$ — постоянная времени дифференцирования; C_0 — интегральная составляющая; C_1 — пропорциональная составляющая; C_2 — дифференциальная составляющая.

Следующим шагом является поиск значений параметров на границе заданного запаса устойчивости, минимизирующих принятый критерий качества работы системы. Так, второму интегральному критерию соответствует точка 0,95max (C_0) в сторону большего значения частоты (рис. 3, a). Переходная характеристика системы представлена на рис. 3, b.

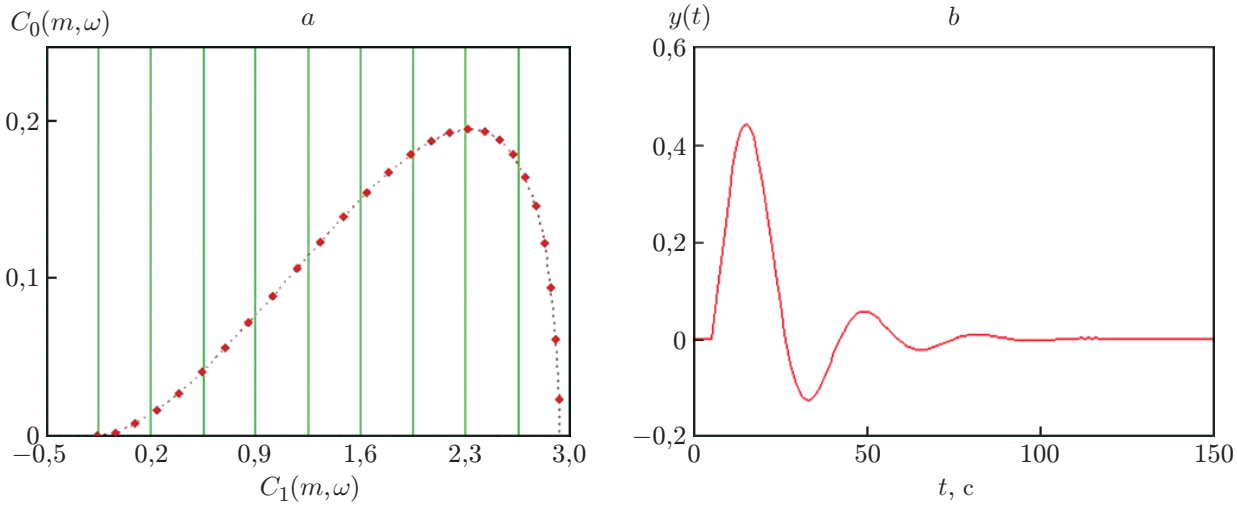

Рис. 3. Характеристики ПИД-регулятора: область параметров C_0, C_1 для корневого метода (a), переходный процесс системы (b)

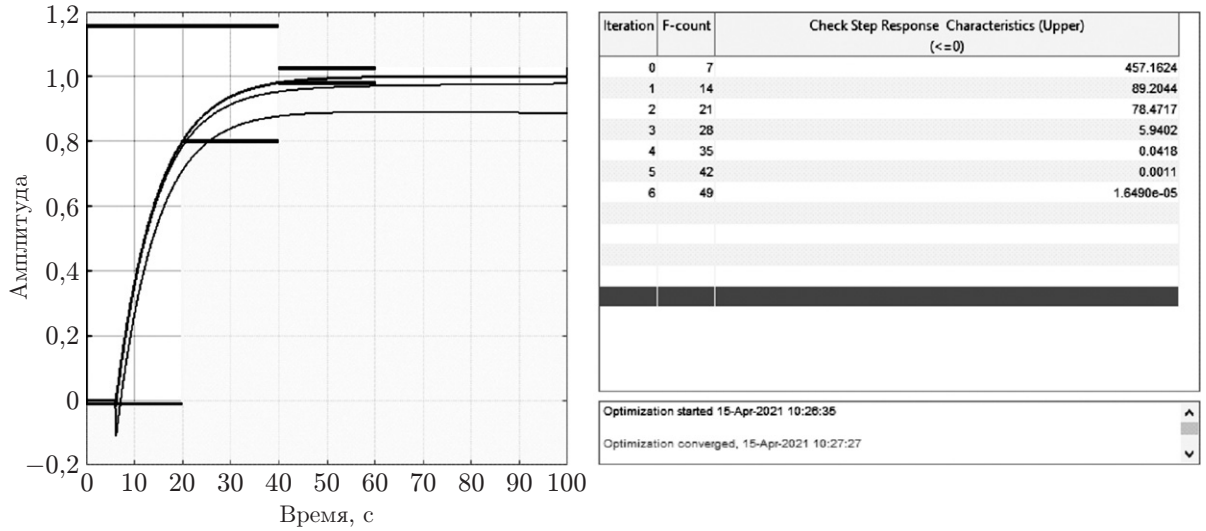

Рис. 4. Семейство переходных процессов, получаемых при процедуре оптимизации

В основе метода незатухающих колебаний (Циглера — Никольса) лежат расчёт критической настройки пропорциональной составляющей K_u, при которой система управления находится на границе устойчивости, и период колебаний T_u. Затем находятся остальные настроечные параметры, выбранные в соответствии с требованиями к процессу регулирования:

$$C_0 = 0{,}4K_u/T_u, \qquad C_1 = 0{,}2K_u, \qquad C_2 = 0{,}066K_uT_u.$$

Таким образом, согласно приведённым процедурам расчёта настроечные параметры ПИД-регулятора, полученные с помощью корневого метода, составляют: $C_0 = 0{,}193$, $C_1 = 2{,}408$, $C_2 = 0{,}5$; методом Циглера — Никольса: $C_0 = 0{,}108$, $C_1 = 1{,}069$, $C_2 = 0{,}018$.

Для расчёта параметров ПИД-регулятора с упредителем Смита воспользуемся пакетом расширения системы MATLAB — Simulink Design Optimization, позволяющим решать задачу синтеза регулирующих систем путём поиска оптимальных настроечных парамет-

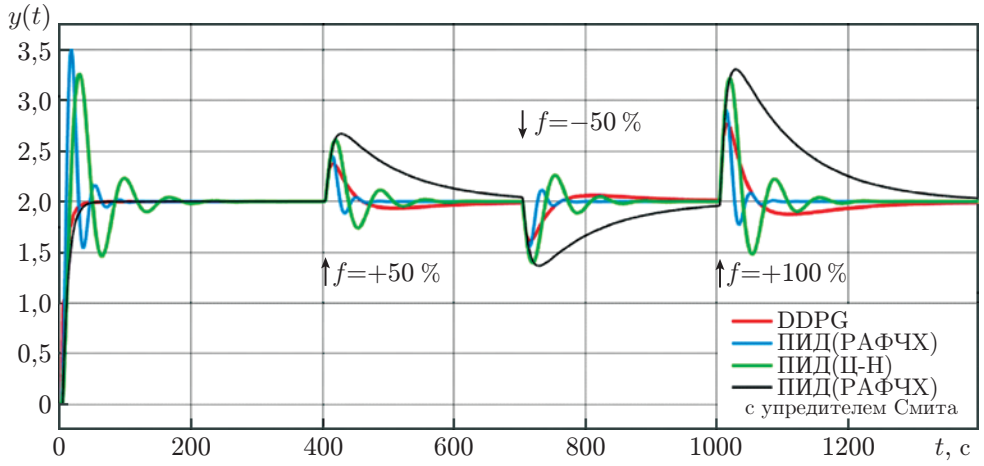

Рис. 5. Динамические характеристики рассматриваемых систем регулирования

Таблица 1

Показатель качества	ПИД-регулятор (корневой метод)			ПИД-регулятор (корневой метод) с упредителем Смита			ПИД-регулятор (метод незатухающих колебаний)			Система с алгоритмом DDPG		
	T, с											
	47,7	97,7	147,7	47,7	97,7	147,7	47,7	97,7	147,7	47,7	97,7	147,7
$y_м$	—	3,47	3,17	2,22	0	0	3,53	3,238	3,254	2,204	0	2,144
σ, %	—	73,6	58,95	11	0	0	76,5	61,9	62,7	10,2	0	7,2
$t_р$, с	—	57,9	65,8	34,23	30,98	43,2	77,1	111,8	184,5	50,44	24,42	65,22
I_2	—	39,2	45,94	23,85	27,63	40,22	46,7	65,03	86,34	22,56	20,43	31,18

ров на основе требований к качеству процесса регулирования: времени нарастания, времени перехода в установившийся режим, величины перерегулирования, величины статической ошибки. Процесс поиска представляет собой итерационное моделирование переходного процесса и подстройку параметров регулятора на основе отклика системы и выбранного метода оптимизации (градиентный спуск) (рис. 4). Определены следующие оптимальные настроечные параметры: $C_0 = 0{,}0188$, $C_1 = 1{,}8536$, $C_2 = -0{,}3552$.

Моделирование обучения системы управления с подкреплением производилось в пакете Reinforcement Learning Toolbox MATLAB 2020b со следующими параметрами: шаг обучения 0,1 с, максимальное количество эпизодов 5000, максимальное значение накопленного вознаграждения 1000 (3). Конфигурация используемого оборудования: 4-ядерный 8-поточный процессор Intel® Core™ i5-8300H, 16 ГБ ОЗУ. Отметим, что обучение исследуемой системы заняло около 180 мин.

На рис. 5 представлено сравнение переходных характеристик систем с ПИД-регулированием и системы регулирования на основе алгоритма DDPG при воздействии внешних возмущений f, имеющих значения +50 %, −50 %, +100 % от значения задающего воздействия 2 и $T = 97{,}7$ с.

Оценить качество процесса регулирования можно с помощью следующих показателей: максимальное динамическое отклонение $y_м$, перерегулирование σ, время регулирования $t_р$, второй интегральный критерий I_2 (6). Табл. 1 содержит качественное сравнение процесса

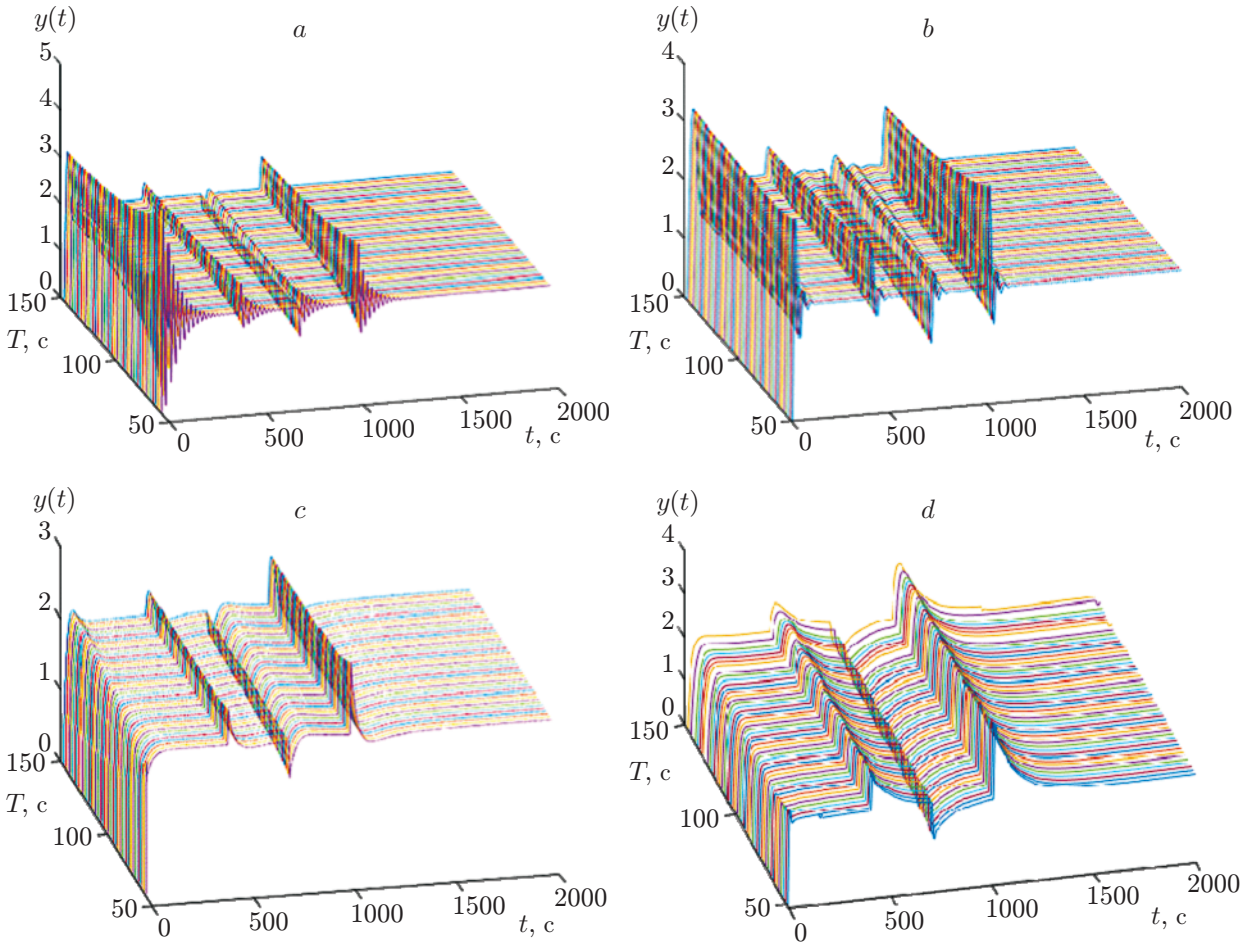

Рис. 6. Динамика изменения переходного процесса при параметрическом и внешнем возмущении: модель с ПИД-регулятором (корневой метод) (a), модель с ПИД-регулятором (метод незатухающих колебаний) (b), модель обучения с подкреплением (c), модель ПИД-регулятора с упредителем Смита (d)

регулирования при различных значениях параметрического возмущения n, представленного изменением значения постоянной времени объекта T.

Классические методы регулирования показали разные результаты. ПИД-регулятор, настроенный с применением корневого метода, имеет лучшие показатели регулирования по сравнению с ПИД-регулятором, настроенным с помощью метода Циглера — Никольса, однако демонстрирует выход на границу устойчивости при достижении величины $T = 53$ с и дальнейшее неустойчивое движение в случае ещё большего понижения величины T. Применение упредителя Смита позволяет нивелировать влияние транспортного запаздывания, что даёт прирост качественных показателей процесса регулирования и устойчивость системы к внешним воздействиям даже при наличии параметрического возмущения. По сравнению с остальными моделями система на основе нейросети эффективнее справляется с регулированием и стабилизацией переходного процесса. Отследить работу системы при различных параметрических воздействиях можно на рис. 6.

Заключение. Применение методов компенсации инерционности дало возможность адаптировать обучающую модель к наличию транспортного запаздывания у объекта управления и синтезировать систему, которая позволяет эффективнее осуществлять процесс регулирования по сравнению с классическими решениями. Разработанная система

управления на основе нейросети с использованием алгоритма DDPG обладает высокими показателями качества регулирования, устойчивости к параметрическим и внешним возмущениям.

Благодаря множеству архитектур, возможности совмещения классических подходов к синтезу систем автоматического управления и методов обучения, применение искусственного интеллекта в задачах управления технологическими процессами позволяет конфигурировать систему, способную к качественному функционированию в сложных, постоянно меняющихся условиях среды.

Отметим, что подобный подход является универсальным, однако для объектов более высоких порядков может потребоваться модификация наблюдателя системы и ввод дополнительных величин наблюдения для обеспечения высокого качества регулирования. Также, по мнению авторов, необходимо добавить, что в такой реализации, скорее всего, не понадобится вычисление точных значений производных при формировании наблюдателя, поскольку важными особенностями рассматриваемых систем являются (выражаясь терминологией методов машинного обучения) «особые признаки» протекания технологического процесса, а не его точные значения. Этому будут посвящены дальнейшие исследования.

СПИСОК ЛИТЕРАТУРЫ

1. **Шидловский С. В.** Автоматическое управление. Реконфигурируемые системы: Учеб. пособие. Томск: Изд-во Том. ун-та, 2010. 168 с.
2. **Французова Г. А., Востриков А. С.** Особенности синтеза ПИД-регулятора для нелинейного объекта второго порядка // Автометрия. 2019. **55**, № 4. С. 57–64. DOI: 10.15372/AUT20190406.
3. **Медведев В. С., Потемкин В. Г.** Нейронные сети. MATLAB 6 /Под общ. ред. В. Г. Потемкина. М.: ДИАЛОГ-МИФИ, 2002. 496 с.
4. **Уоссермен Ф.** Нейрокомпьютерная техника: теория и практика. М.: Мир, 1992. 184 с.
5. **Пакулич Д. В., Якимов С. А., Алямкин С. А.** Распознавание возраста по изображению лица с использованием свёрточных нейронных сетей // Автометрия. 2019. **55**, № 3. С. 52–61. DOI: 10.15372/AUT20190307.
6. **Васенков Д. В.** Методы обучения искусственных нейронных сетей // Компьютерные инструменты в образовании. 2007. № 1. С. 1–10.
7. **Sutton R. S., Barto A. G.** Reinforcement Learning: An Introduction. London: The MIT Press Cambridge, 2015. 352 p.
8. **Шидловский С. В.** Автоматическое управление. Перестраиваемые структуры. Томск: Томский государственный университет, 2006. 288 с.
9. **Шидловский С. В.** Логическая система с перестраиваемой структурой в задачах управления технологическим процессом // Автометрия. 2005. **41**, № 4. С. 104–113.
10. **Lillicrap T. P., Hunt J. J., Pritzel A. et al.** Continuous Control With Deep Reinforcement Learning. 2019. 14 p. URL: https://arxiv.org/pdf/1509.02971.pdf (дата обращения: 28.01.2021).
11. **Deep** Deterministic Policy Gradient Agents. URL: https://www.mathworks.com/help/reinforcement-learning/ug/ddpg-agents.html (дата обращения: 28.01.2021).
12. **Боровик В. С., Шидловский С. В.** Обучение с подкреплением в задачах управления технологическими процессами // Телекоммуникации. 2020. № 11. С. 36–40.
13. **Mohammadi M., Arefi M. M., Setoodeh P., Kaynak O.** Optimal tracking control based on reinforcement learning value iteration algorithm for time-delayed nonlinear systems with external disturbances and input constraints // Inform. Sciences. 2021. **554**. P. 84–98.

14. **Chen B., Xu M., Liu Z. et al.** Delay-aware multi-agent reinforcement learning for cooperative and competitive environments. 2020. 11 p. URL: https://arxiv.org/pdf/2005.05441v2.pdf (дата обращения: 28.01.2021).
15. **Шидловский С. В.** Теория автоматического управления: Учеб. пособие. Томск: Изд-во НТЛ, 2003. 40 с.
16. **Стефани Е. П.** Основы расчёта настройки регуляторов теплоэнергетических процессов. М.: Энергия, 1972. 377 с.

Поступила в редакцию 28.01.2021
После доработки 15.04.2021
Принята к публикации 26.04.2021