



Сравнение классических алгоритмов контроля и методов обучения с подкреплением на примере промышленной эталонной модели в контексте адаптивных цифровых двойников

В.А. Щелов, ORCID: 0009-0005-2773-2038 <shchelov.va@phystech.edu>

*НИУ Московский физико-технический институт (МФТИ),
Физтех-школа прикладной математики и информатики (ФПМИ),
Россия, 141701, Московская область, г. Долгопрудный, Институтский переулок, д. 7.*

Аннотация. Представлены результаты исследования нескольких архитектурных подходов к созданию адаптивных контроллеров на базе обучения с подкреплением (RL) для задачи управления бинарной дистилляционной колонной. Проведен анализ влияния формы функции вознаграждения и структуры нейронной сети на эффективность обучения, что ранее не получало достаточного освещения в литературе. Показано, что класс нейронных сетей на базе радиальных базисных функций, может быть расширен за счет использования полносвязных нейронных сетей с несколькими слоями и значительным числом нейронов в скрытом слое. Результаты исследования свидетельствуют, что архитектуры, в которых RL подбирает параметры конвенционального регулятора, демонстрируют лучшие показатели и требуют значительно меньшего количества итераций для достижения оптимальной производительности по сравнению с использованием чистого подкрепления.

Ключевые слова: обучение с подкреплением; адаптивные контроллеры; гибридные системы управления.

Для цитирования: Щелов В.А. Сравнение классических алгоритмов контроля и методов обучения с подкреплением на примере промышленной эталонной модели в контексте адаптивных цифровых двойников. Труды ИСП РАН, 2026, том 38, вып. 6, часть 1, стр. 161–176. DOI: 10.15514/ISPRAS-2026-38(6)-10.

Благодарности: Автор выражает благодарность команде RocketControl, представленной Павлом Приходько и Сергеем Ковтуном, а также Роману Бельскому из компании «Норильский никель», за предоставление описаний, алгоритмов и данных, относящихся к некоторым симуляторам, использованным в данном исследовании.

Comparison of classical control algorithms and reinforcement learning methods using an industrial benchmark in the context of adaptive digital twins

V.A. Shchelov, ORCID: 0009-0005-2773-2038 <shchelov.va@phystech.edu>

*Moscow Institute of Physics and Technology (National Research University MIPT),
Phystech School of Applied Mathematics and Informatics (FPMI),
Bldg. 7, Institutsky Lane, Dolgoprudny, Moscow Region, 141701, Russia.*

Abstract. The results of a study of several architectural approaches to the creation of adaptive controllers based on reinforcement learning (RL) for the task of controlling a binary distillation column are presented. The analysis of the influence of the form of the reward function and the structure of the neural network on the effectiveness of learning, which previously had not received sufficient coverage in the literature, was carried out. It is shown that the class of neural networks based on radial basis functions (RBF) can be expanded by using fully connected neural networks with several layers and a significant number of neurons in the hidden layer. The results of the study indicate that architectures in which RL selects the parameters of a conventional regulator demonstrate better performance and require significantly fewer iterations to achieve optimal performance compared to using pure RL.

Keywords: reinforcement learning; adaptive controllers; hybrid control systems.

For citation: Shchelov V.A. Comparison of classical control algorithms and reinforcement learning methods using an industrial benchmark in the context of adaptive digital twins. *Trudy ISP RAN/Proc. ISP RAS*, 2026, vol. 38, issue 6, part 1, pp. 161-176 (in Russian). DOI:10.15514/ISPRAS-2026-38(6)-10.

Acknowledgements. We gratefully acknowledge the RocketControl team, represented by Pavel Prikhodko and Sergey Kovtun, as well as Roman Belsky from Norilsk Nickel, for supplying descriptions, algorithms, and data related to certain simulators employed in this research.

1. Введение

Прогресс в вычислительных, сенсорных и коммуникационных технологиях обусловил значительное повышение уровня автоматизации промышленных систем. Парадигма четвертой промышленной революции (Индустрия 4.0) предполагает переход к полностью цифровизированному и автономному производству, в котором интеллектуальные системы осуществляют взаимодействие с окружающей средой в режиме реального времени, осуществляют интеграцию информации и формируют связанные промышленные экосистемы, объединяющие устройства и сервисы. В данном контексте концепция цифрового двойника (ЦД), сформулированная в работе [1], приобрела статус одного из центральных элементов. Цифровой двойник представляет собой виртуальную модель устройства или процесса, динамически отображающую текущее состояние соответствующего физического объекта. Ключевая особенность концепции ЦД заключается в обеспечении двустороннего обмена данными, что позволяет реализовать мониторинг, прогнозирование и оптимизацию в реальном времени. Кроме того, во многих прикладных областях цифровые двойники используются в качестве виртуальных прототипов, способствующих принятию решений в условиях высокой сложности и неопределенности функционирования систем [2].

Цифровые двойники (ЦД) могут быть разработаны для широкого круга объектов – от отдельных физических установок, таких как турбины, двигатели и ветроэнергетические системы, до комплексных промышленных процессов в таких отраслях, как химическое производство, нефтепереработка и газовая промышленность. Для подобных объектов к числу приоритетных задач относятся раннее выявление неисправностей, обеспечение безопасности и оптимизация эксплуатационных характеристик [3]. Применение ЦД позволяет реализовывать архитектуры управления, обеспечивающие мониторинг и коррекцию параметров технологического процесса в реальном времени.

В настоящее время в промышленной автоматизации широко используются классические регуляторы: пропорционально-интегрально-дифференциальный (PID) и модельно-предиктивный контроллер (MPC). PID-регуляторы отличаются простотой формализма и эффективностью вычислений, что обеспечивает их повсеместное применение, а также включение во многие промышленные стандарты. Однако они требуют регулярной ручной калибровки и имеют ограниченную применимость для управления нелинейными процессами. В свою очередь, MPC-контроллеры используют предиктивные модели для расчета управляющего воздействия, что позволяет учитывать ограничения на управление и прогнозировать поведение системы. Основными ограничениями MPC являются высокая вычислительная сложность и сильная зависимость от точности математической модели и оценки состояния объекта управления.

Классические методы управления в значительной степени зависят от экспертных эвристик, ручного подбора параметров и корректности предположений, положенных в основу модели, что затрудняет использование таких подходов в динамичных и слабо формализуемых производственных условиях [4]. В тоже время, все возрастающий объем данных промышленных производств и развитие методов машинного обучения, создает условия для применения подходов, основанных на данных (data-driven), позволяющих создавать адаптивные системы мониторинга и управления. Наиболее многообещающими представляются гибридные методы, при которых интеграция data-driven подходов с традиционными системами управления открывает перспективы для повышения эффективности и адаптивности промышленных процессов, снижая присущие классическим методам ограничения [5-10].

Среди методов машинного обучения значительное внимание привлекло обучение с подкреплением (reinforcement learning, RL) благодаря его способности обучаться стратегии управления на основе взаимодействия с виртуальной средой, будь то симуляция или игра [11]. RL все чаще используется для управления промышленными процессами, где симуляторы служат как средами обучения, так и тестовыми площадками для оценки изученных стратегий. Недавние работы продемонстрировали потенциал управления на основе RL для сложных промышленных систем [12-15]. Однако существующие решения RL часто тесно связаны со специфическими симуляторами процессов, не имеют реализаций с открытым исходным кодом и требуют существенной адаптации к конкретной предметной области, что препятствует воспроизводимости и обобщению.

В рамках концепции цифровых двойников предполагается, что на начальных этапах разработки, когда прототип цифрового двойника (digital twin prototype) еще не интегрирован с физическим объектом, его данные формируются посредством моделирования в виртуальной среде [1]. После установления связи с соответствующим физическим компонентом цифровой двойник должен перейти к работе с реальными данными объекта. Предполагается, что методы машинного обучения, в частности обучение с подкреплением, могут служить технологической основой для осуществления этого перехода бесшовно в автоматическом режиме [16].

2. Связанные работы

Работы по исследованию применения RL для управления промышленными процессами можно разделить на несколько категорий: непосредственное применение RL для контроля процесса, гибридные методы, сочетающие PID и RL, гибридные методы, сочетающие MPC и RL. Ранние исследования, такие как [16], подчеркивают способность RL справляться с нелинейной динамикой, стохастичностью и оптимизацией на долгосрочном горизонте. Ма и др. [12] применили DDPG для управления полупериодической полимеризацией, показав конкурентоспособные результаты с гибридными стратегиями PI/MPC, но полагаясь на симулятор, специфичный для процесса и не предназначенный для повторного использования.

В работе [13] представлена схема оптимизации в реальном времени на основе RL, которая заменяет стационарный RTO архитектурой «актор-критик», однако их реализация остается привязанной к одной нелинейной модели реактора. Ли и др. [14] разработали контроллер DRL для процесса периодического культивирования биофармацевтического препарата с использованием IndPenSim, добившись значительного повышения выхода продукции за счет ограничения пространства действий рецептами действий человека в специализированной среде моделирования.

Существует несколько стратегий адаптивного PID-контроля, включая адаптивное PID-регулирование на основе модели [17-19] и адаптивное PID-регулирование на основе нейронной сети [20, 21]. Было отмечено, что адаптивное PID-регулирование на основе модели требует предположения, что созданная модель может точно представлять динамику реального объекта [22]. PID-регулирование на основе нейронных сетей использующее обучение с учителем имеет ряд ограничений, такие как сложность получения обучающего сигнала и трудности с прогнозированием значений для немаркированных данных. Обучение с подкреплением не имеет указанных недостатков. Поэтому, рядом авторов были предложены архитектуры адаптивных PID-контроллеров, на основании RBF сети [23, 24], однако систематического сравнения с сетями другой архитектуры не проводилось.

Комбинация MPC и RL – быстро развивающееся направление, обещающее повышение производительности в изменяющихся, стохастических и сложных средах [25, 26]. Несмотря на теоретические достижения и растущий объем академической литературы в этой области, остается заметный пробел в предоставлении программных инструментов с открытым исходным кодом для RL на основе MPC. Оба подхода базируются на формализме Марковского процесс принятия решений и стремятся найти оптимальную политику π для максимизации заданного функционала, однако существенно различаются в подходе к решению задачи. RL обучается методом проб и ошибок через взаимодействие со средой, сходясь к оптимальной стратегии на отдаленном, в идеале бесконечном, временном горизонте. MPC напротив, базируется на точной модели математической системы, решая оптимизационную задачу на каждом временном шаге для конечного временного горизонта. Основными слабостями RL применительно к контролю процессов являются необходимость большого количества итераций для обучения, трудности обобщения на неизвестные состояния, а также малая интерпретируемость. MPC зависит от точности используемой модели процесса, получить которую может быть трудно, и точности оценки состояния системы на момент решения оптимизационной задачи. Различные комбинации MPC и RL предлагаются для решения широкого круга проблем, от увеличения вычислительной эффективности при решении задачи оптимизации, до достижения адаптивности MPC за счет RL [27].

Настоящая работа в качестве отправной точки фокусируется на рассмотрении PID+RL в их различных сочетаниях, поскольку PID регулирование предоставляет удобный и простой формализм для контроля системы. Одновременно с этим, как было указано ранее систематического сравнения подходов и используемых структур сети не проводилось. Сравнение с MPC выполняется, но сочетание MPC+RL рассматривается только в общих чертах, поскольку варианты реализации адаптивного MPC требуют еще больших вычислительных затрат по сравнению с базовым MPC и затруднительны для дальнейшего практического применения.

Также, часто в работах, демонстрирующих применение RL или выполняющих сравнение алгоритмов, используются метрики характерные для RL такие как среднее вознаграждение и дисперсия во время обучения, что затрудняет сравнение алгоритмов. В настоящей работе метрики вознаграждения используются на предварительном шаге отбора моделей. Финальное сравнение выполняется с использованием традиционных для теории контроля метрик: интеграл квадратичной ошибки (ISE), интеграл абсолютной ошибки (IAE),

взвешенной по времени квадратичной ошибки (ITSE) и взвешенной по времени абсолютной ошибки (ITAE), формулы которых представлены ниже.

$$ISE = \int e^2(t) dt, \quad (1)$$

$$IAE = \int |e(t)| dt, \quad (2)$$

$$ITSE = \int t e^2(t) dt, \quad (3)$$

$$ITAE = \int t |e(t)| dt. \quad (4)$$

где t – время, а $e(t) = y_d - y(t)$ – невязка между целевым значением и наблюдаемым значением на момент времени t .

3. Описание модели промышленного процесса

В качестве объекта для сравнения алгоритмов была выбрана модель бинарной дистилляционной колонны (БДК), которая служит эталоном, предоставленным нашими промышленными партнерами, для оценки альтернативных стратегий управления. Система включает в себя одну управляемую переменную, два управляемых выхода и вектор состояния, состоящий из 32 переменных (концентрации на тарелках ДК). Внутри БДК не происходит химической реакции, вместо этого осуществляется термодинамическое разделение смеси на дистиллят и остаток. Схематично система представлена на рис.1.

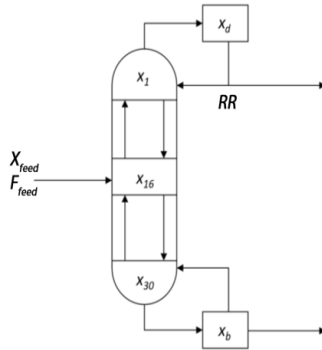


Рис.1. Схематическое изображение модели БДК.
Fig. 1. Schematic representation of the BDK model.

Основные динамические зависимости, управляющие работой колонны, приведены ниже.

$$\begin{aligned} F_d &= \frac{1}{2} F_{feed}, & F_l &= R F_f, \\ F_v &= F_l + F_d, & F_b &= F_f + F_l. \end{aligned} \quad (5)$$

где F_{feed} – поток входной смеси (моль/мин), F_d – поток дистиллята (моль/мин), F_l – поток жидкости в ректификационной секции колонны (моль/мин), F_v – поток пара в колонне (моль/мин), F_b – поток жидкости в секции отгонки (моль/мин).

Уравнение равновесия жидкость-пар описывается уравнением:

$$y_i = \frac{x_i \alpha}{1 + x_i (\alpha - 1)}. \quad (6)$$

где, x_i содержание компонента на i тарелке, α относительная летучесть.

В конденсаторе выполняется соотношение:

$$\frac{dx_0}{dt} = \frac{F_v}{N_c} (y_1 - x_0), \quad (7)$$

где x_0 содержание компонента в конденсаторе, N_c мольное содержание компонента в конденсаторе.

В ректификационной и отгоночной секциях колонны выполняются аналогичные выражения:

$$\frac{dx_i}{dt} = \frac{F_l(x_{i-1} - x_i) - F_v(y_i - y_{i+1})}{N_t}, \quad (8)$$

$$\frac{dx_i}{dt} = \frac{F_L(x_{i-1} - x_i) - F_v(y_i - y_{i+1})}{N_t}, \quad (9)$$

где, N_t – мольное содержание компонента на тарелке.

Для подающей тарелки и ребойлера соответственно

$$\frac{dx_{16}}{dt} = \frac{F_{feed} \bar{x}_{feed} + F_l x_{15} - F_L x_{16}}{N_t} - \frac{F_v (y_{16} - y_{17})}{N_t}, \quad (10)$$

$$\frac{dx_{31}}{dt} = \frac{F_L x_{30} - (F_f - F_d) x_{31} - F_v y_{31}}{N_b}, \quad (11)$$

где, x_{16} и x_{31} – содержание компонента на подающей тарелке и в ребойлере соответственно, N_b – мольное содержание компонента в ребойлере.

Касательно состава компонента верны следующие соотношения:

$$x_d = x_0, \quad x_b = x_{31}, \quad (12)$$

$$\bar{x}_{feed} = x_{feed} + d(t). \quad (13)$$

Возмущение вводится через состав подаваемого сырья x_{feed} , который подвергается воздействию гауссова шума $d(t)$, отфильтрованного фильтром Баттерворта, с параметрами $\mu = 0$, $\sigma = 0,05$, порядком фильтра $n = 2$ и частотой среза $\omega_c = 0,05$.

Типичная картина возмущения x_{feed} представлена на рис.2.

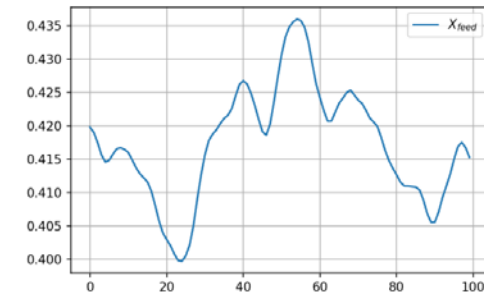


Рис. 2. Характерная форма возмущения x_{feed} (содержание компонента во входной смеси) в системе.
Fig. 2. Characteristic form of disturbance x_{feed} (content of the component in the input mixture) in the system.

Номинальные параметры эксплуатации БДК составляют $\alpha = 1.6$, $N_t = 0.25$ (моль), $N_c = 0.5$ (моль), $N_b = 1.0$ (моль), $F_{feed} = 1.0$ (моль/мин), $x_{feed} = 0.42$ (без шума). Начальное состояние системы представлено вектором $x_0 \dots x_{31} = (0.935, 0.900, 0.862, 0.821, 0.779, 0.738, 0.698, 0.661, 0.628, 0.599, 0.574, 0.553, 0.535, 0.521, 0.510, 0.501, 0.494, 0.485, 0.474, 0.459, 0.441, 0.419, 0.392, 0.360, 0.324, 0.284, 0.243, 0.201, 0.161, 0.125, 0.092, 0.064)$.

Регулируемой переменной является коэффициент рефлюкса R , ограниченный значениями $0,1 \leq R \leq 10$. Наблюдаемыми переменными являются составы дистиллята и остатка, x_d и x_b .

Временной шаг составляет 1 мин. Основная задача управления – поддержание состава дистиллята на целевом уровне путем регулирования R . В настоящей работе в качестве целевого состояния системы рассматривается состояние, при котором содержание системы достигает целевого значения $x_d = 0.8$ и поддерживается постоянным.

Особенностью системы является то, что в описанных условиях регулирование системы может быть достигнуто наивным Р-регулятором с $K_p = 1$, однако остается еще большой запас для повышения эффективности регулирования. Одновременно с этим, применение стандартной процедуры подбора параметров PID-регулятора, такой как метод Циглера-Никольса приводит к значительной осцилляции наблюдаемой переменной. Общее выражение для PID регулятора может быть представлено формулой

$$u(t) = K_p e(t) + K_i \int_0^t e(\tau) d\tau + K_d \frac{de}{dt}, \quad (14)$$

где $e(t)$ имеет то же значение что и в выражениях (1-4), а K_p, K_i, K_d – коэффициенты усиления пропорциональной, интегрирующей и дифференцирующей составляющих регулятора.

В настоящей работе в качестве базовых методов подбора параметров используются метод Циглера-Никольса, а также метод подбора параметров методом минимизации функционала за счет применения программирования методом последовательных наименьших квадратов (SLSQP). В качестве функционала для минимизации методом SLSQP была выбрана интегральная абсолютная погрешность по времени (ITAE). Основным преимуществом ITAE является сильный штраф на последующих временных шагах, что заставляет функцию как можно раньше достигать целевого значения, при этом подавляя возможные осцилляции. Значения параметров регулятора приведены в табл. 1, а отклик представлен на рис. 3.

Табл. 1. Параметры PID-регуляторов, полученных конвенциональными методами.

Table 1. Parameters of PID controllers obtained by conventional methods.

Метод	K_p	K_i	K_d
Наивный	1	0	0
Циглера-Никольса	19.2	5.5	8.7
SLSQP	28.1	0	42.1

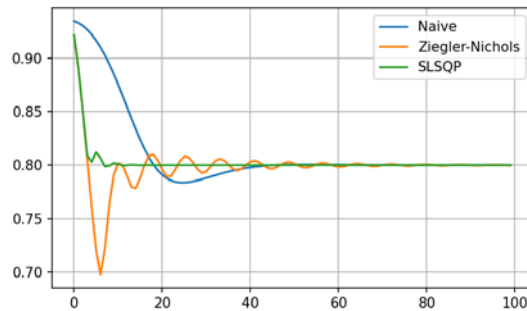


Рис. 3. Отклик PID-регуляторов, полученных конвенциональными методами (без возмущений, $d(t)$ в выражении 13 равен 0).

Fig. 3. Response of PID controllers obtained using conventional methods (without disturbances, $d(t)$ in expression 13 is equal to 0).

4. Описание экспериментов

В исследовании рассматривались подходы к созданию адаптивных контроллеров в рамках концепции цифровых двойников с применением методов глубинного обучения и усиленного обучения. Все обучение производилось на симуляторе системы, описанной в разделе 3, без возмущений ($d(t)$ в выражении 13 равно 0) для того, чтобы оценить обобщающую способность и эмулировать переход от идеальной системы к реальной. Финальное сравнение осуществлялось на системе с участием возмущения, описанного в разделе 3. Отбор параметров моделей и алгоритмов осуществлялся по среднему вознаграждению на обучении. Лучшие контроллеры, полученные применением RL, сравнивались с конвенциональными регуляторами на метриках ISE IAE ITSE ITAE (выражения 1-4).

В рамках работы исследовались три различных архитектурных подхода с использованием нейронных сетей в качестве аппроксиматоров управляющей функции, и выполнялся сравнительный анализ с классическими методами управления (PID контроллерами, настроенными по методам, описанным в разделе 3 и MPC). Первый подход основан на использовании входного вектора, содержащего значение ошибки, её первые и вторые дифференциалы (e, de, d^2e), выходными сигналами являются параметры PID-контроллера (K_p, K_i, K_d). Второй подход предполагает использование вектора контекста системы, включающего текущие значение наблюдаемых переменных, целевое значение и предыдущее значение управляемой переменной, в качестве выходного вектора также выступают параметры PID-контроллера (K_p, K_i, K_d). Третий подход также использует входной вектор, аналогичный подходу два, но вместо предсказания параметров PID осуществляется прямое предсказание значения управляющей переменной. Условно эти подходы можно обозначить как RL-pid, RL-semi-pid и чистый RL. Схематично это изображено на рис. 4.

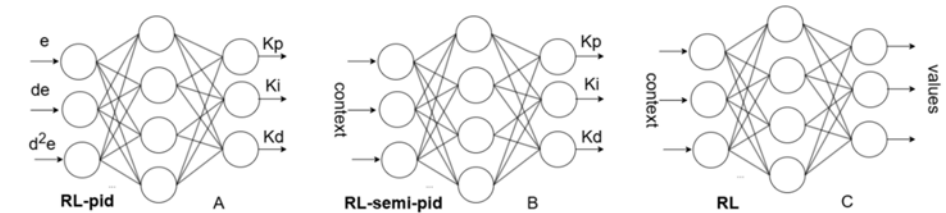


Рис. 4. Исследованные подходы: А) RL-pid входной вектор (e, de, d^2e) и выходной (K_p, K_i, K_d)

В) RL-semi-pid вектор context = (y_{t-1}, y_d, u_{t-1}) выходной вектор (K_p, K_i, K_d)

С) чистый RL вектор context = (y_{t-1}, y_d, u_{t-1}) и выходной (y_t).

Fig. 4. Studied architectural approaches А) RL-pid input vector (e, de, d^2e) output (K_p, K_i, K_d)

В) RL-semi-pid vector context = (y_{t-1}, y_d, u_{t-1}) output vector (K_p, K_i, K_d)

С) pure RL vector context = (y_{t-1}, y_d, u_{t-1}) output (y_t).

Обучение реализовывалось посредством алгоритмов Proximal Policy Optimization (PPO), Soft Actor-Critic (SAC).

Для PPO функция ценности состояния, обозначаемая $v_\pi(s)$, представляет собой ожидаемое будущее вознаграждение, которое будет получено, начиная с состояния s , измеренное, начиная с этого состояния s , и следуя политике π (in-policy алгоритм). Алгоритм обновляет политики следующим образом:

$$\theta_{k+1} = \arg \max_{\theta} \mathbb{E} [L(s, a, \theta_k, \theta)] \quad (15)$$

Как правило, для максимизации целевой функции выполняется несколько шагов стохастического градиентного спуска (SGD). L задается следующим образом:

$$L(s, a, \theta_k, \theta) = \min (r(\theta)(a|s) A^{\pi_{\theta_k}}(s, a), \text{clip}(r(\theta)(a|s), 1 - \epsilon, 1 + \epsilon) A^{\pi_{\theta_k}}(s, a)) \quad (16)$$

где

$$r(\theta)(a|s) = \frac{\pi_{\theta}(a|s)}{\pi_{\theta_k}(a|s)}, \tag{17}$$

$$A^{\pi_{\theta_k}}(s,a) = Q^{\pi_{\theta_k}}(s,a) - V^{\pi_{\theta_k}}(s) \tag{18}$$

SAC это метод обучения с подкреплением, предназначенный для вычисления оптимальной политики, которая максимизирует ожидаемое долгосрочное вознаграждение и энтропию политики. Энтропия политики служит мерой неопределенности политики в зависимости от текущего состояния. Более высокое значение энтропии побуждает алгоритм к более глубокому исследованию, предотвращая чрезмерную концентрацию на использовании известных вознаграждений.

Отличные от стандартных параметры алгоритмов обучения приведены в табл. 2.

Табл. 2. Параметры использованных алгоритмов обучения.

Table 2. Parameters of the used learning algorithms.

Параметр	PPO	SAC
Скорость обучения	0.0003	0.0003
Дисконт (γ)	0.99	0.99
Размер буфера воспроизведения	--	10 ⁶
Целевая энтропия (α)	--	Авто
Критик	True	--
Generalized Advantage Estimator	True	--

Структура использованных нейронных сетей представляла собой полносвязную 4-х слойную нейронную сеть с 2-мя скрытыми слоями с количеством нейронов от 16 до 256 по степеням двойки. Функция активации для нейронов использовалась SiLU:

$$SiLU(x) = x \frac{1}{1 + e^{-x}} \tag{19}$$

Все обучение велось на эпизодах длиной 100 шагов, за которые алгоритм должен был достичь целевого значения. Допустимый диапазон изменения коэффициентов усиления K_p , K_b , K_d от 0 до 100.

Также исследовалось влияние формы функции вознаграждения на обучение. Рассматривались 2 различные формы функции вознаграждения с нормализацией и без нее. Функция вознаграждения без нормализации рассчитывалась согласно выражению

$$r = -ITAE \tag{20}$$

Нормализованная функция вознаграждения рассчитывалась как,

$$r = \begin{cases} 0 & |e(t)| < |e(t-1)| \\ \frac{1}{1+ITAE} & |e(t)| > |e(t-1)| \end{cases} \tag{21}$$

Для обоих случаев ITAE рассчитывалась по ближайшему горизонту шагов симулятора (4 шага). Нормализованная функция вознаграждения удобна в смысле легко интерпретируемой визуализации успеха обучения на эпизоде. Поскольку вознаграждение на шаге, в этом случае, располагается в диапазоне [0,1], результат будет показывать процент действий с максимальной эффективностью на эпизоде. Гипотетический идеальный алгоритм управления, достигающий целевого значения на первом шаге и удерживающий его в течение всего эпизода, получит вознаграждение 100. Исследование влияние формы функции вознаграждения на обучение проводилось на единственной архитектуре сети с количеством

нейронов в скрытом слое равно 16 на единственном алгоритме PPO. Для каждой из форм функции вознаграждения обучался контроллер в течении 800 итераций, после чего вычислялись его метрики ISE IAE ITSE ITAE.

Последовательность экспериментов может быть представлена следующим образом:

1. Анализ влияния структуры функции награды на параметры обучения и финальную эффективность.
2. Исследование влияния размера NN на параметры обучения и качество финального контроллера.
3. Обучение контроллеров алгоритмами PPO, SAC в рамках 3-х описанных подходов RL-pid, RL-semi-pid и RL.
4. Выбор лучших моделей для финального сравнения с классическим контроллером (PID, настроенные по методикам Циглера-Никольса, и минимизации SLSQP и MPC), с использованием стандартных метрик: ISF, IAE, ITAE, ITSH. В рамках финального теста в выражении $13\ d(t)=f(\mu, \sigma, n, \omega_c)$ согласно описанию в разделе 3, количество шагов, на которых выполняется сравнение, равняется 300.

5. Результаты

Результаты применения нормализованной и ненормализованной функций вознаграждения при обучении PPO на сети с 16 нейронами в каждом из скрытых слоев приведены в табл 3.

Табл. 3. Метрики контроллера после обучения для каждой из функции вознаграждения.

Table 3. Controller metrics after training for each reward function.

Функция вознаграждения	ISE	IAE	ITSE	ITAE
Нормализованная	0.05	1.22	0.45	38.2
Не нормализованная	0.05	1.31	0.48	39.9

Как видно из таблицы, для обеих форм функции вознаграждения результаты обучения близки. В дальнейшем, ввиду большего удобства нормализованной функции вознаграждения, используется нормализованная функция вознаграждения.

Влияние архитектуры сети на скорость и результативность обучения представлены на рис. 5 в координатах номер учебной итерации – средняя награда на итерации.

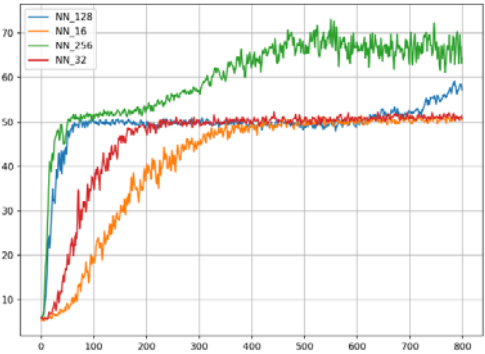


Рис. 5. Влияние количества нейронов в скрытых слоях нейронной сети на скорость и глубину обучения (PPO, RL-semi-pid).

Fig. 5. The influence of the number of neurons in the hidden layers of a neural network on the speed and depth of learning (PPO, RL-semi-pid).

Дальнейшее обучение для каждого из вариантов архитектуры RL-pid, RL-semi-pid и RL обучение выполнялось с нормализованной функцией потерь и 256-тью нейронами в скрытых слоях для каждого алгоритма. Результаты обучения каждого из контроллеров представлены в табл. 4.

Из табл. 4 видно, наибольшее среднее вознаграждение набрали алгоритмы для архитектур RL-pid и RL-semi-pid. Вариант RL-pid кажется, более универсальным, но поскольку оба подхода показывают близкие значения, было принято решения сравнить оба подхода в финальном тесте. Таким образом, в финальном сравнении принимали участие контроллеры PID_{Naive} , PID_{ZN} , PID_{SLSQP} , как референсные PID регуляторы, MPC регулятор и два data-driven регулятора построенных по описанным выше архитектурам с использованием PPO: RL-pid_{PPO} и RL-semi-pid_{PPO}. Метрики финального сравнения алгоритмов приведены в табл. 5.

Табл. 4. Среднее вознаграждение после обучения для каждой архитектуры и алгоритма для сети с 256 нейронами в скрытых слоях.

Table 4. Average reward after training for each architecture and algorithm for a network with 256 neurons in hidden layers.

	RL-pid	RL-semi-pid	RL
PPO	71.2	69.7	29.1
SAC	67.2	63.2	26,3

Табл. 5. Метрики контроллера после обучения для каждой из функции вознаграждения.

Table 5. Controller metrics after training for each reward function.

	ISE	IAE	ITSE	ITAE
PID_{Naive}	0.32	4.32	7.08	348.3
PID_{ZN}	0.06	1.76	0.81	85.08
PID_{SLSQP}	0.04	0.59	0.07	28.70
RL-pid _{PPO}	0.05	0.82	0.07	26.53
RL-semi-pid _{PPO}	0.05	0.93	0.06	13.33
MPC	0.03	0.46	0.07	28.56

6. Обсуждение результатов

Данные табл. 4 показывают, что варианты архитектуры, в которых RL не напрямую управляет процессом, а подбирает параметры для контроллера показывают большую эффективность в процессе обучения и достигают более высоких результатов, чем прямое применение RL. Одновременно с этим, для варианта чистого RL нужно значительно большее количество учебных итераций и предъявления большого количества разнообразных траекторий для достижения вознаграждения сопоставимого с тем, которое достигается для вариантов RL-pid или RL-semi-pid. Так в вариантах RL-pid или RL-semi-pid среднее вознаграждение в 50 единиц досчитается за 40-60 итераций, в то время как в варианте чистого RL для достижения среднего вознаграждения в 25 и более единиц требуется более 5000 итераций. Небольшая разница в результатах PPO и SAC, скорее всего, обусловлена необходимостью более тщательного подбора параметров обучения для SAC.

Данные табл. 5 показывают, что контроллеры, полученные применением RL-pid_{PPO}, RL-semi-pid_{PPO}, демонстрируют сходные метрики с референсными алгоритмами PID_{SLSQP} и MPC и превосходят их по целевой характеристике ITAE. Анализируя траектории изменения выходного сигнала под действием каждого из контроллеров, можно видеть разницу в стратегиях поведения в рамках каждого из контроллеров. На рис. 6 представлены результаты

контроля для каждого из регуляторов на первых и последних 50 шагах финального испытания. Из рисунка видно, что MPC и PID_{SLSQP} представляют траекторию близкую к оптимальной на начальном участке, причем PID_{SLSQP} делает это даже более жестко чем MPC. При этом RL-pid_{PPO} и RL-semi-pid_{PPO} показывают худшие результаты на начальном участке траектории, но зато компенсируют их в дальнейшем, более тесно прижимая значение наблюдаемой переменной к целевому значению (минимизация ITAE). Интересно, что RL-semi-pid_{PPO}, который показал минимальный ITAE, жертвует качеством на начальном участке траектории, где штраф за ITAE не так силен, отдавая предпочтение близости значения наблюдаемой переменной к целевому значению. При этом RL-pid_{PPO} старается более плавно достичь целевого значения. Референсные алгоритмы MPC и PID_{SLSQP} показывают хорошие результаты на начальном участке траектории, но чуть хуже справляются с удержанием наблюдаемой переменной вблизи целевого значения ввиду стохастичности возмущения, описанного в разделе 3.

На рис. 7 показана разница в стратегиях манипулирования параметрами $K_p K_i K_d$ PID-регулятора в случаях RL-pid и RL-semi-pid. Из рисунка видно, что найденная RL-pid стратегия держать K_d максимально высоким, K_i максимально низким, при этом основное управление осуществлять за счет манипулирования K_p , в ситуациях, когда возмущение становится достаточно сильным.

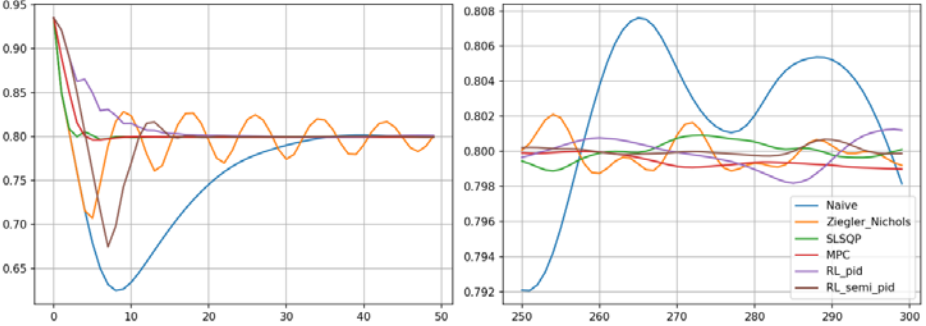


Рис. 6. Траектория управления на начальном и заключительном этапах (50 итераций) финального теста.

Fig. 6. The control trajectory at the initial and final stages (50 iterations) of the final test.

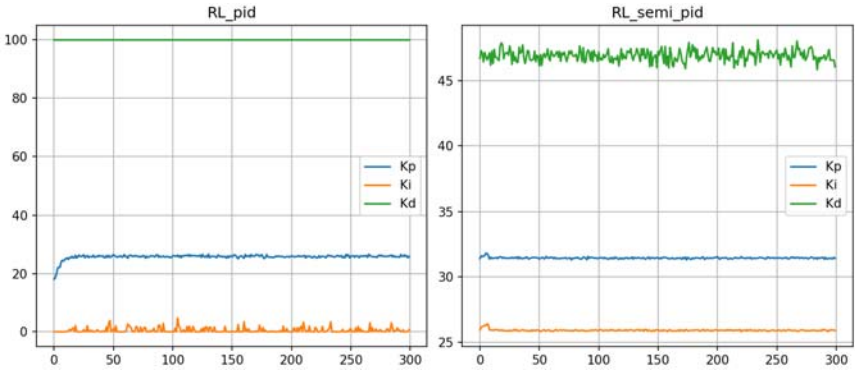


Рис. 7. Стратегия управления параметрами зависимого PID-регулятора со стороны нейронных сетей RL-pid и RL-semi-pid.

Fig. 7. The strategy for controlling the parameters of a dependent PID controller using RL-pid and RL-semi-pid neural networks.

Параметр K_p при этом примерно соответствует параметру K_p определенному при минимизации методом SLSQP как описано в разделе 3. При этом RL-semi-pid наоборот удерживает K_p и K_i почти постоянными на относительно высоких значениях, а управление осуществляет через параметр K_d .

7. Выводы

Было проведено исследование нескольких архитектурных подходов к созданию адаптивных контроллеров на базе обучения с подкреплением RL-pid RL-semi-pid RL на примере модели бинарной дистилляционной колонны. Было проанализировано влияние формы функции вознаграждения на обучение, а также исследовано влияние структуры сети на обучение для рассматриваемой задачи контроля, которое ранее не рассматривалось в литературе.

Было показано, что класс нейронных сетей на базе RBF, предложенный ранее [23, 24], может быть расширен. Для указанного типа регуляторов возможно применение также полносвязная нейронных сетей с достаточным количеством слоев и нейронов в скрытом слое. По результатам исследования, можно заключить, что варианты архитектуры, где RL подбирает параметры для PID-регулятора (RL-pid и RL-semi-pid), показывают лучшие результаты и требуют значительно меньшего количества итераций для достижения оптимальной производительности по сравнению с применением чистого RL. Адаптивные контроллеры, созданные с помощью RL, демонстрируют сходные или лучшие метрики, особенно для выбранной в качестве целевой в рамках исследования метрики ITAE. Было выявлено, что в рамках применения RL для оптимального обучения и функционирования алгоритмов необходим тщательный подбор параметров обучения. Одновременно с этим, совмещение RL с контроллером на основе PID действительно позволяет находить эффективные стратегии регулирования в случае неизвестных на момент обучения возмущений. Хорошие начальные предположения о возможных значениях параметров управляемого регулятора, такие как K_d , K_i , K_d и их дисперсии, могут приводить к еще более быстрому поиску оптимальной стратегии регулирования.

Одновременно с этим нужно отметить, что регуляторы получающиеся совмещением обучения с подкреплением с конвенциональными методами контроля нельзя назвать полностью автоматическими. Поскольку в этом случае вся сложность настройки таких контроллеров перетекает из области эксплуатации в сферу обучения адаптивных алгоритмов и их оценки, где возможно требуется учет даже большего количество параметров, чем в оригинальном алгоритме.

Дальнейшие направления изучения связаны с увеличением сложности систем и возмущений, увеличением количество одновременно контролируемых параметров, а также переходе к исследованию комбинаций обучения с подкреплениями с другими типами конвенциональных регуляторов, таких как MPC+RL LQR+RL и так далее. Отдельное направление связано с изучением применения адаптивных регуляторов в случае систем с несколькими контурами управления, когда наравне со штатным течением производственного процесса требуется также реагировать на возникающие исключительные ситуации, такие как неисправность сенсоров, актуаторов или другого оборудования.

Список литературы/References

- [1]. Grieves M., Vickers J. Digital twin: Mitigating unpredictable, undesirable emergent behavior in complex systems. *Transdisciplinary perspectives on complex systems: New findings and approaches*, 2017, pp. 85-113.
- [2]. Jones D., Snider C., Nassehi A., Yon J., Hicks B. Characterising the digital twin: A systematic literature review. *CIRP Journal of Manufacturing Science and Technology*, 2020, vol. 29, pp. 36-52. DOI: 10.1016/j.cirpj.2020.02.002. Available at: <https://www.sciencedirect.com/science/article/pii/S1755581720300110>, accessed 25.03.2026.

- [3]. Kreuzer T., Papapetrou P., Zdravkovic J. Artificial intelligence in digital twins—a systematic literature review. *Data & Knowledge Engineering*, 2024, vol. 151. DOI: 10.1016/j.datak.2024.102304.
- [4]. Chen J., Luo N. An offline-to-online reinforcement learning framework with trajectory-guided exploration for industrial process control, *Journal of Process Control*, 2025, vol. 154, 103535.
- [5]. Qin S.J. Data-driven fault detection and diagnosis for complex industrial processes. *IFAC Proceedings Volumes*, 2009, vol. 42, no. 8, pp.1115-1125.
- [6]. Lomov I., Lyubimov M., Makarov I., Zhukov L.E. Fault detection in Tennessee eastman process with temporal deep learning models. *Journal of Industrial Information Integration*, 2021, vol. 23, 100216.
- [7]. Golyadkin M., Pozdnyakov V., Zhukov L., Makarov I. Sensorscan: Self-supervised learning and deep clustering for fault diagnosis in chemical processes. *Artificial Intelligence*, 2023, vol. 324, 104012. DOI: 10.1016/j.artint.2023.104012. Available at <https://www.sciencedirect.com/science/article/pii/S0004370223001583>, accessed 25.03.2026.
- [8]. Zhevnenko D., Kazantsev M., Makarov I. Interaction models for remaining useful lifetime estimation. *Journal of Industrial Information Integration*, 2023, vol. 33, 100444.
- [9]. Ahmad A., Kovalenko A., Makarov I. Anomaly detection using graph-based autoencoder with graph structure learning layer. In *2024 IEEE 6th International Symposium on Logistics and Industrial Informatics (LINDI)*, 2024, pp. 89-94. DOI:10.1109/LINDI63813.2024.10820392.
- [10]. Kovalenko A., Pozdnyakov V., Makarov I. Graph neural networks with trainable adjacency matrices for fault diagnosis on multivariate sensor data. *IEEE Access*, 2024.
- [11]. Shakya A.K., Pillai G., Chakrabarty S. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 2023, vol. 231, 120495.
- [12]. Ma Y., Zhu W., Benton M.G., Romagnoli J. Continuous control of a polymerization system with deep reinforcement learning. *Journal of Process Control*, 2019, vol. 75, pp. 40-47.
- [13]. Powell K.M., Machalek D., Quah T. Realtime optimization using reinforcement learning. *Computers & Chemical Engineering*, 2020, vol. 143, 107077.
- [14]. Li H., Qiu T., You F. AI-based optimal control of fed-batch biopharmaceutical process leveraging deep reinforcement learning. *Chemical Engineering Science*, 2024, vol. 292, 119990.
- [15]. Hassanpour H., Mhaskar P., Corbett B. A practically implementable reinforcement learning control approach by leveraging offset-free model predictive control. *Computers & Chemical Engineering*, 2024, vol. 181, 108511.
- [16]. Nian R., Liu J., Huang B. A review on reinforcement learning: Introduction and applications in industrial process control. *Computers & Chemical Engineering*, 2020, vol. 139, 106886.
- [17]. Chang W.D., Hwang R.C., Hsieh J.G. A multivariable on-line adaptive PID controller using auto-tuning neurons. *Engineering Application of Artificial Intelligence*, 2003, no. 16, pp. 57-63.
- [18]. Yamamoto T., Shah S. Design and experimental evaluation of multivariable self-tuning PID controller. *IEE Proceedings-Control Theory and Applications*, 2004, vol. 151, no. 5, pp. 645-652.
- [19]. Yu D.L., Chang T.K., Yu D.W. A stable self-learning PID control for multivariable time varying systems. *Control Engineering Practice*, 2007, vol. 15, no. 12, pp. 1577-1587.
- [20]. Chen J.H., Huang T.C. Applying neural networks to on-line updated PID controllers for nonlinear process control. *Journal of Process Control*, 2004, vol. 14, no. 2, pp. 211-230.
- [21]. Liao Y.T., Koiwai K., Yamamoto T. Design and implementation of a hierarchical-clustering CMAC PID controller. *Asian Journal of Control*, 2019, vol. 21, no. 3, pp. 1077-1087.
- [22]. Hou Z.S., Chi R.H., Gao H.J. An overview of dynamic linearization based data-driven control and applications. *IEEE Transactions on Industrial Electronics*, 2016, vol. 64, no. 5, pp. 4076-4090.
- [23]. Zhe G., Toru Y. Design of a Reinforcement Learning PID Controller. *IEEJ Transactions on Electrical and Electronic Engineering*, 2021, vol. 16. DOI: 10.1002/tee.23430.
- [24]. Jin Z., Li H., Gao H. An intelligent weld control strategy based on reinforcement learning approach. *Int J Adv Manuf Technol*, 2019, vol. 100, pp. 2163-2175. DOI:10.1007/s00170-018-2864-2.
- [25]. Gros S., Zanon M. Data-Driven Economic NMPC Using Reinforcement Learning. *IEEE Transactions on Automatic Control*, Feb. 2020, vol. 65, no. 2, pp. 636-648.
- [26]. Zanon M., Gros S. Safe reinforcement learning using robust MPC. *IEEE Transactions on Automatic Control*, 2021, vol. 66, no. 8, pp. 3638-3652.
- [27]. Reiter R., Hoffmann J., Reinhardt D., Messerer F., Baumgärtner K., Sawant S., Boedecker J., Diehl M., Gros S. Synthesis of Model Predictive Control and Reinforcement Learning: Survey and Classification. *arXiv preprint*, 2025. DOI:10.48550/arXiv.2502.02133.

Информация об авторах / Information about authors

Владимир Александрович ЩЕЛОВ – преподаватель-исследователь Физтех-школы прикладной математики и информатики Московского физико-технического института, сотрудник кафедры корпоративных информационных систем. Область исследований включает: автоматизированные системы управления, методы машинного обучения в промышленной автоматизации, адаптивное управление технологическими процессами, цифровые двойники производственных систем, интеграция ERP и MES-решений.

Vladimir Aleksandrovich SHCHELOV is a research professor at the Phystech School of Applied Mathematics and Informatics at the Moscow Institute of Physics and Technology. He works at the Department of Corporate Information Systems. The research area includes: automated control systems, machine learning methods in industrial automation, adaptive process management, digital twins of production systems, and integration of ERP and MES solutions.