

УДК 004.896

DOI: [10.26102/2310-6018/2026.61.10.011](https://doi.org/10.26102/2310-6018/2026.61.10.011)

Методы обучения с подкреплением для автоматической настройки и онлайн-адаптации регуляторов трудно настраиваемых динамических объектов: аналитический обзор

С.С. Воробей✉

*Московский технический университет связи и информатики, Москва,
Российская Федерация*

Резюме. Рассмотрена задача автоматической настройки и подстройки в процессе работы регуляторов сложных динамических объектов, для которых однократно рассчитанный набор параметров не сохраняет требуемого качества управления на всем сроке эксплуатации. Цель обзора состоит в том, чтобы установить границы применимости существующих методов настройки по единому критерию и обосновать направления дальнейших исследований по нерешенной научной проблеме. Поиск велся по реферативным базам и охватывал классические первоисточники и работы прикладного периода; отбирались рецензируемые источники, относящиеся к настройке регуляторов. Прослежено развитие подходов от эмпирических правил настройки пропорционально-интегрально-дифференциальных регуляторов и управления с активным подавлением возмущений до обучаемых правил настройки, построенных методами обучения с подкреплением. Для каждого семейства методов выделены достоинства и пределы применимости: классические правила рассчитывают фиксированные коэффициенты по одному эксперименту и теряют качество при дрейфе параметров объекта; обучаемые методы работают с непрерывными параметрами, но не гарантируют устойчивости замкнутого контура в процессе обучения и требуют большого числа взаимодействий с объектом. Разрозненные подходы сведены к единому критерию пригодности, а именно работе с непрерывным вектором параметров и выборочной эффективности в реальном времени, и систематизированы в четыре нерешенные проблемы: сохранение устойчивости при обучении, выборочную эффективность, перенос с модели на объект и дефицит подтверждения на натурных стендах. Каждая проблема сопоставлена с направлением дальнейших исследований. Полученное сопоставление показывает, что заявленная проблема остается открытой для всего класса трудно настраиваемых объектов и требует одновременного решения вопросов устойчивости, выборочной эффективности и натурной валидации.

Ключевые слова: обучение с подкреплением, автоматическая настройка регуляторов, онлайн-адаптация, устойчивость замкнутого контура, выборочная эффективность, популяционное обучение, перенос с модели на объект, натурная валидация.

Для цитирования: Воробей С.С. Методы обучения с подкреплением для автоматической настройки и онлайн-адаптации регуляторов трудно настраиваемых динамических объектов: аналитический обзор. *Моделирование, оптимизация и информационные технологии*. 2026;14(10). URL: <https://moitvvt.ru/ru/journal/article?id=2555> DOI: 10.26102/2310-6018/2026.61.10.011

Reinforcement learning methods for automatic tuning and online adaptation of controllers for hard-to-tune dynamic plants: an analytical review

S.S. Vorobey✉

*Moscow Technical University of Communications and Informatics, Moscow,
the Russian Federation*

Abstract. This review addresses the problem of automatic tuning and online adaptation of controllers for complex dynamic plants for which a single computed parameter set does not preserve the required control quality over the whole service interval. The aim is to establish, under a single criterion, the applicability limits of the existing tuning methods and to substantiate directions for further research on an open scientific problem. The search covered classical primary sources and applied-period works selected from abstract databases. The development of approaches is traced from empirical tuning rules for proportional-integral-derivative controllers and active disturbance rejection control to learned tuning rules built by reinforcement learning. For each family of methods the strengths and the applicability limits are identified: classical rules compute fixed coefficients from a single experiment and lose quality under plant parameter drift, whereas learned methods operate with continuous parameters but do not guarantee closed-loop stability during learning and require a large number of interactions with the plant. The scattered approaches are reduced to a single suitability criterion, namely operation with a continuous parameter vector and sample efficiency in real time, and are systematised into four open problems: preserving stability during learning, sample efficiency, transfer from a model to the plant, and the shortage of validation on physical test benches. Each problem is matched to a direction for further research. The mapping shows that the stated problem remains open for the whole class of hard-to-tune plants and requires the simultaneous resolution of stability, sample efficiency, and physical validation.

Keywords: reinforcement learning, automatic controller tuning, online adaptation, closed-loop stability, sample efficiency, population-based training, sim-to-real transfer, physical validation.

For citation: Vorobey S.S. Reinforcement learning methods for automatic tuning and online adaptation of controllers for hard-to-tune dynamic plants: an analytical review. *Modeling, Optimization and Information Technology*. 2026;14(10). (In Russ.). URL: <https://moitvvt.ru/ru/journal/article?id=2555> DOI: 10.26102/2310-6018/2026.61.10.011

Введение

Пропорционально-интегрально-дифференцирующий (ПИД) регулятор остается наиболее распространенным законом управления с обратной связью уже более ста лет. Н. Минорский в работе 1922 года по автоматическому рулевому управлению судами ввел управляющий закон с пропорциональной, интегральной и дифференциальной составляющими по ошибке курса и связал их роли с физикой движения корабля [1]. Систематические правила выбора коэффициентов появились двумя десятилетиями позже в работе Дж. Циглера и Н. Николса [2]. По результатам промышленных исследований ПИД-структура использовалась примерно в 97 % регуляторов, а приемлемое качество демонстрировала лишь около трети контуров; причины неудовлетворительной работы включали настройку, состояние исполнительных механизмов, измерительные и эксплуатационные факторы [3, 4]. Причины такого доминирования известны инженерному сообществу: простая структура с тремя настраиваемыми параметрами, дешевая аппаратная реализация и проработанный инструментарий настройки.

Практика, однако, показывает, что значительная часть промышленных контуров работает неоптимально. Такая ситуация создает потребность в методах автоматической настройки (autotuning) и адаптации коэффициентов, способных поддерживать заданное качество регулирования без постоянного участия квалифицированного инженера-наладчика.

Потребность обостряется на объектах, для которых однократно рассчитанный набор параметров не обеспечивает требуемого качества на всем сроке эксплуатации и службы оборудования. Такое поведение объясняется сочетанием различных свойств объекта: нестационарностью параметров (деградация оборудования со временем, изменение нагрузки и смена режима, дрейф рабочей точки), при которой рассчитанные ранее коэффициенты со временем перестают быть оптимальными, а в частности,

приводят к потере устойчивости [5]; транспортным запаздыванием в канале управления, ограничивающим достижимую полосу пропускания контура; насыщением исполнительного механизма, при котором линейная модель перестает описывать систему; высокой частотой контура, оставляющей на каждый шаг управления единицы миллисекунд и исключающей длительную офлайн-перенастройку в работе. Объект, обладающий таким сочетанием свойств, далее называется трудно настраиваемым.

Именно требование подстройки коэффициентов непосредственно в процессе работы обнажает недостаточность классических методов. Эмпирические правила [2], эвристики, и модельно-ориентированные процедуры рассчитывают фиксированный набор параметров для одного режима по результату однократного эксперимента; при дрейфе параметров объекта этот набор, как правило, перестает минимизировать критерий качества управления. Метаэвристические и оптимизационные методы выполняют глобальный поиск и могут находить высококачественные решения в многомерном пространстве параметров, однако обычно требуют многократных оценок критерия/прогонов, что ограничивает их применение непосредственно на работающем объекте [6, 7]. Поэтому возникает естественная мысль строить правило выбора параметров не аналитически из модели объекта, а по результатам взаимодействия с ним, то есть методами обучения с подкреплением [8].

С точки зрения информационных процессов задача настройки регулятора представляет собой задачу обработки информации в контуре управления: сигналы (ошибка, выход объекта, текущие коэффициенты) образуют поток данных, по которому вычислительный метод формирует и оптимизирует вектор параметров θ в работе контура. Нейросетевой агент обучения с подкреплением выступает здесь технологией анализа разнородной информации контура и выработки управляющего решения без явной аналитической модели объекта. Поскольку требование реального времени и стоимость взаимодействия с работающим объектом выводят реализацию метода на уровень распределенного программно-аппаратного комплекса сбора, передачи и обработки информации контура, постановка задачи относится к информатике и информационным процессам, а не исключительно к теории управления. Именно в этой парадигме – метод обработки сигналов контура, нейросетевая технология и архитектура комплекса реального времени – и ведется дальнейший разбор.

Перенос обучения с подкреплением в контур управления, однако сам наталкивается на ряд препятствий, систематический разбор которых составляет содержание настоящего обзора. Научная проблема состоит в следующем: отсутствует метод и его программно-аппаратная реализация, которые одновременно настраивают и адаптируют регуляторы трудно настраиваемых объектов в процессе работы, сохраняют устойчивость замкнутого контура на всех этапах обучения, выборочно эффективны в условиях реального времени и подтверждены натурным экспериментом.

Цель обзора – установить по единому критерию пределы применимости существующих методов настройки регуляторов трудно настраиваемых объектов и обосновать направления дальнейших исследований по выделенной проблеме. Задачи обзора: 1) проследить эволюцию методов настройки от классических правил до обучаемых и для каждого семейства выделить достоинства и пределы применимости; 2) ввести единый критерий сопоставления семейств, а именно пригодность для непрерывного вектора параметров θ и выборочную эффективность при ограниченном бюджете взаимодействия; 3) систематизировать разрозненную литературу в перечень фальсифицируемых нерешенных проблем; 4) сопоставить каждую проблему с направлением дальнейших исследований.

Новизна обзора состоит в том, что разрозненные подходы к настройке регуляторов сведены к единому критерию (непрерывность θ и выборочная эффективность в реальном времени), тогда как существующие обзоры рассматривают семейства методов изолированно, и систематизированы в четыре фальсифицируемые нерешенные проблемы. Практическая польза – обоснование направлений дальнейших исследований: для каждой проблемы указано, какое требование к методу настройки она задает. Изложение строится от общей задачи настройки трудно настраиваемого объекта через критический разбор классических и обучаемых методов к перечню нерешённых проблем и их сопоставлению с направлениями дальнейших исследований.

Материалы и методы

Обзор подготовлен как аналитический с элементами систематического анализа и отбора. Поиск источников велся по базам данных научного цитирования (Scopus, eLIBRARY.RU), электронной библиотеке IEEE Xplore и архиву препринтов arXiv. Временной охват: классические первоисточники по теории ПИД-регуляторов, ADRC и базовым алгоритмам обучения с подкреплением привлекались независимо от года издания, а прикладные работы по настройке регуляторов методами обучения с подкреплением отбирались за период 2017–2026 гг., чтобы отразить и становление направления, и его современный срез. Использовались поисковые запросы по сочетаниям ключевых терминов: reinforcement learning PID tuning, ADRC tuning, safe reinforcement learning control, sim-to-real transfer control, а также их русскоязычные эквиваленты.

Критерии включения определены следующими: рецензируемые статьи и монографии, прямо относящиеся к настройке или онлайн-адаптации параметров регуляторов либо к методам обучения с подкреплением, применимым к этой задаче. Критерии исключения: работы без отношения к параметрической настройке регуляторов, схожие публикации и учебные пособия. Всего на этапе анализа рассмотрено более 150 работ; после применения критериев в обзор отобрано 57 источников. Каждое семейство методов разбиралось по единому шаблону: достоинство, границы применимости, вывод о пригодности для непрерывного вектора параметров θ при ограниченном бюджете взаимодействия с объектом. Этот критерий и служил основанием для сопоставления семейств между собой.

Математическая постановка задачи. Динамику линеаризованного в окрестности рабочей точки одномерного объекта удобно описывать передаточной функцией с транспортным запаздыванием:

$$W_{об}(s) = \frac{Y(s)}{U(s)} = \frac{b_m s^m + \dots + b_1 s + b_0}{a_n s^n + \dots + a_1 s + a_0} e^{-\tau s}, \quad (1)$$

где $U(s)$, $Y(s)$ – изображения по Лапласу управляющего воздействия и выхода объекта; a_i , b_j – коэффициенты, зависящие от рабочей точки и изменяющиеся во времени для нестационарного объекта; $\tau \geq 0$ – транспортное запаздывание; $n \geq m$ – порядок объекта. Зависимость коэффициентов от времени и наличие звена $e^{-\tau s}$ составляют формальное содержание понятия трудно настраиваемого объекта: модель (1) не является фиксированной, а ее параметры заранее не известны с достаточной точностью.

На объект (1) замыкается регулятор. В простейшем случае это ПИД-регулятор, формирующий управление по ошибке $e(t) = r(t) - y(t)$ как $u(t) = K_P e(t) + K_I \int_0^t e(\xi) d\xi + K_D \dot{e}(t)$, где $r(t)$ – задающее воздействие; K_P , K_I , K_D – настраиваемые коэффициенты [2]. Качество замкнутого контура определяется вектором параметров регулятора θ , который для ПИД-регулятора совпадает с тройкой коэффициентов,

$\theta = (K_P, K_I, K_D)^T$. Задача автоматической настройки и онлайн-адаптации состоит в построении правила выбора θ непосредственно по наблюдаемым сигналам контура, минимизирующего функционал качества управления:

$$J(\theta) = \int_0^T [q e^2(t, \theta) + \rho u^2(t, \theta)] dt \rightarrow \min_{\theta \in \Theta}, \quad (2)$$

где $e(t, \theta)$ – ошибка регулирования при текущем векторе параметров; $u(t, \theta)$ – соответствующее управление; $q > 0$, $\rho \geq 0$ – весовые коэффициенты, задающие компромисс между точностью наведения и энергией управления; T – горизонт оценивания; Θ – допустимое множество параметров, на котором замкнутый контур сохраняет устойчивость. Частными случаями (2) являются интегральные критерии ISE, IAE и ITAE [5], систематически рассматриваемые в теории автоматического управления техническими системами¹.

Принципиальное отличие постановки от классической состоит в требовании онлайн-адаптации: правило выбора θ должно отслеживать изменение параметров объекта (1) в процессе работы, а не рассчитывать θ однократно. Именно это требование и делает задачу (2) открытой для всего класса трудно настраиваемых объектов. Далее задача конкретизируется для двух классов регуляторов – ПИД и ADRC – при рассмотрении классических и адаптивных методов настройки.

Классические и адаптивные методы настройки

ПИД-регулятор: от Циглера-Николса к автонастройке и адаптивным схемам.

Трехчленный закон управления получил практическое инженерное применение ранее своего систематического теоретического оформления. Работа Минорского 1922 года закрепила пропорционально-интегрально-дифференциальную структуру [1], которая к настоящему времени остается наиболее распространенной в промышленных контурах и излагается в классических курсах теории автоматического регулирования [9]. Систематические правила выбора коэффициентов появились в 1942 году: Дж. Циглер и Н. Николс предложили два эмпирических метода, не требующих явной модели объекта [2]. В методе незатухающих колебаний интегральная и дифференциальная составляющие отключаются, а пропорциональный коэффициент увеличивается до возникновения устойчивых автоколебаний; по их параметрам коэффициенты вычисляются как:

$$K_P = 0,6 K_u, \quad K_I = \frac{2K_P}{T_u}, \quad K_D = \frac{K_P T_u}{8}, \quad (3)$$

где K_u – критический коэффициент усиления, при котором контур выходит на границу устойчивости; T_u – период критических колебаний на этой границе. Второй метод опирается на аппроксимацию переходной характеристики разомкнутого объекта моделью первого порядка с запаздыванием. Несмотря на почти восьмидесятилетнюю давность, правила (3) сохраняются как отправная точка настройки. Прозрачность и независимость от детальной модели делают метод Циглера-Николса естественным опорным (базовым) методом для экспериментального сравнения.

Главным достоинством правил (3) является простота: они не требуют идентификации модели и реализуются практически на любом промышленном контроллере. Главный минус прямо вытекает из постановки (2): вектор θ рассчитывается однократно для одного режима и далее не изменяется, поэтому при дрейфе параметров объекта он, как правило, перестает быть оптимальным. Прямым развитием правил стали уточненные таблицы Коэна-Куна для объектов с выраженным запаздыванием [10] и

¹ Бесекерский В.А., Попов Е.П. *Теория систем автоматического управления*. Санкт-Петербург: Профессия; 2003. 752 с.

методы, основанные на идентификации модели первого порядка по переходной характеристике. Все они сохраняют общую черту: коэффициенты считаются один раз по одному эксперименту.

Существенный шаг к автоматизации сделан в методе релейной автонастройки К. Острёма и Т. Хёгглунда [11]. Вместо ручного вывода контура на границу устойчивости в обратную связь временно включается релейный элемент, переключающий сигнал между двумя уровнями; возникающие устойчивые автоколебания позволяют определить критические параметры K_u и T_u из эксперимента, после чего коэффициенты вычисляются по модифицированным соотношениям типа (3). Эксперимент проводится в замкнутом контуре, амплитуда автоколебаний ограничена уровнем релейного сигнала, а процедура выполняется автоматически; благодаря этим свойствам метод реализован в большинстве промышленных ПИД-контроллеров с функцией автонастройки. Однако и он рассчитывает фиксированный вектор по одному эксперименту.

Задача подстройки коэффициентов при изменении характеристик объекта решается в рамках самонастраивающихся (self-tuning) и адаптивных ПИД-регуляторов [12]. Самонастраивающийся регулятор на каждом шаге рекуррентно идентифицирует модель объекта и пересчитывает θ по текущей оценке (например, методом размещения полюсов); модельно-эталонное адаптивное управление непрерывно подстраивает коэффициенты так, чтобы поведение контура приближалось к заданной эталонной модели [13]. Эти методы снимают ограничение однократного расчета, но опираются на явную идентификацию модели заданной структуры, качество которой падает на нелинейных и нестационарных объектах, и настраиваются на минимизацию мгновенной ошибки $e(t)$, а не интегрального функционала (2) на горизонте T . Тем самым задача (2) в ее онлайн-постановке остается для ПИД-регуляторов открытой.

ADRC и проблема настройки расширенного наблюдателя состояния. Один из путей ослабить зависимость качества от точности модели состоит в том, чтобы перенести компенсацию неопределенностей внутрь структуры регулятора. Этот путь реализован в управлении с активным подавлением возмущений (Active Disturbance Rejection Control, ADRC), предложенном Хань Цзинцином (J. Han) [14]. Совокупное действие внутренних неопределенностей модели и внешних возмущений трактуется как единое обобщенное возмущение f , которое оценивается в реальном времени и компенсируется в законе управления. Для объекта n -го порядка, записанного в виде $y^{(n)} = f(\cdot) + b_0 u$, где b_0 – приближенно известный коэффициент усиления по управлению, возмущение f вводится как дополнительная переменная состояния и оценивается расширенным наблюдателем состояния (PHC, Extended State Observer, или ESO):

$$\dot{\hat{x}}_i = \hat{x}_{i+1} + \beta_i (y - \hat{x}_1), \quad i = 1, \dots, n, \quad \dot{\hat{x}}_{n+1} = \beta_{n+1} (y - \hat{x}_1), \quad (4)$$

где $\hat{x}_1, \dots, \hat{x}_n$ – оценки координат состояния; $\hat{x}_{n+1} \approx f$ – оценка обобщенного возмущения; β_i – коэффициенты усиления наблюдателя; $y - \hat{x}_1$ – ошибка наблюдения по выходу. Оцененное возмущение вычитается из управления, что в идеальном случае приводит контур к цепочке интеграторов и делает закон управления формально независимым от детальной модели объекта.

В исходной нелинейной формулировке Хань Цзинцин функции усиления РНС нелинейны, а число подлежащих заданию коэффициентов β_i велико, что сдерживало практическое применение. Решающий шаг к инженерной настройке сделан З. Гао (Z. Gao), который заменил нелинейные функции линейными и связал все коэффициенты наблюдателя с единственным параметром – полосой пропускания наблюдателя ω_o [15, 16]:

$$\beta_i = \binom{n+1}{i} \omega_o^i, \quad i = 1, \dots, n+1, \quad (5)$$

где $\omega_o > 0$ – полоса пропускания наблюдателя; $\binom{n+1}{i}$ – биномиальный коэффициент, размещающий все полюсы РНС в точке $-\omega_o$. Коэффициенты обратной связи по состоянию аналогично параметризуются полосой пропускания контроллера ω_c . Линейный вариант (Линейный активный регулятор с компенсацией возмущений, LADRC) второго порядка имеет всего три настраиваемых параметра: ω_o , ω_c и b_0 .

Достоинство ADRC состоит в том, что правильно настроенный регулятор, как правило, превосходит ПИД-регулятор по робастности к неопределенностям и по подавлению возмущений [17]. Однако, несмотря на параметризацию (5), ADRC настраивается труднее ПИД-регулятора, и тому есть несколько причин. Во-первых, даже минимальный вектор $\theta = (\omega_o, \omega_c, b_0)^T$ содержит больше связанных компонентов, чем тройка (K_P, K_I, K_D) , а для нелинейного ADRC и многомерных объектов их число возрастает до десятка и более. Во-вторых, выбор ω_o задает прямой компромисс: увеличение полосы наблюдателя ускоряет и уточняет оценку возмущения, но усиливает чувствительность к шуму измерения и провоцирует пиковый эффект (peaking) в переходном режиме РНС, характерный для наблюдателей с высоким усилением [18]; ограничения настройки LADRC только по двум полосам пропускания и способы использования априорной информации об объекте исследованы в [19]. В-третьих, неудачный выбор параметров РНС способен сделать регулятор неработоспособным вплоть до потери устойчивости контура, тогда как ПИД-регулятор при умеренной расстройке сохраняет работоспособность. Наконец, для ADRC отсутствуют простые эмпирические правила уровня (3): полосы ω_o и ω_c подбираются итеративно по результатам моделирования или натурных испытаний.

Исходя из этого, переход от ПИД-регулятора к ADRC означает, что минимизация (2) ведется по вектору с большим числом связанных компонентов, а допустимое множество θ заведомо уже и хуже описано аналитически. Реализовать преимущество ADRC ручной настройкой на трудно настраиваемом объекте практически не удастся, а при дрейфе его параметров фиксированный θ перестает минимизировать (2) ровно так же, как и для ПИД-регулятора, но с большей ценой ошибки. Поэтому автоматический подбор полос пропускания ω_o и ω_c снимает главное препятствие к применению ADRC: отсутствие инженерных правил настройки. И для ПИД-регулятора, и для ADRC задача (2) в ее онлайн-постановке остается открытой, а ее решение требует правила выбора θ , не опирающегося на явную модель объекта.

Методы обучения с подкреплением для настройки регуляторов. Установленная открытость задачи (2) требует правила выбора θ , не опирающегося на явную модель объекта (1). Такое правило естественно строить бихевиористическими методами обучения с подкреплением (reinforcement learning, RL), в которых правило выработки управления настраивается по результатам взаимодействия с объектом, а не выводится аналитически из его модели [8]. Математической основой служит марковский процесс принятия решений, задаваемый кортежем $(\mathcal{S}, \mathcal{A}, P, R, \gamma)$, где $\mathcal{S}$ – пространство состояний; $\mathcal{A}$ – пространство действий; $P(s' | s, a)$ – функция переходов; $R(s, a)$ – функция вознаграждения; $\gamma \in [0, 1)$ – коэффициент дисконтирования. Агент действует по политике π , максимизируя ожидаемое дисконтированное суммарное вознаграждение, а оптимальная функция ценности действия удовлетворяет уравнению Беллмана:

$$Q^*(s, a) = \mathbb{E}_{s' \sim P} \left[R(s, a) + \gamma \max_{a' \in \mathcal{A}} Q^*(s', a') \right], \quad (6)$$

где $Q^*(s, a)$ – оптимальная ценность действия a в состоянии s ; s' – следующее состояние; a' – действие в следующем состоянии; γ – коэффициент дисконтирования.

Применительно к настройке регулятора состояние s агрегирует наблюдаемые сигналы контура (ошибку и ее историю, выход объекта, текущие коэффициенты), действие a соответствует выбору или коррекции вектора θ , а вознаграждение R строится как убывающая функция мгновенного вклада в (2). Существенно, что непрерывность вектора θ влечет непрерывность пространства действий $\mathcal{A}$, и именно это требование разграничивает применимые семейства алгоритмов.

Приведенная постановка предполагает постоянный шаг дискретизации. Распределенная реализация контура в реальном времени вносит дополнительную особенность: при разнесении агента и объекта по телекоммуникационной сети интервал между последовательными переходами не постоянен, поскольку зависит от сетевой задержки и потерь пакетов. Время пребывания агента в состоянии становится случайной величиной Δt , а марковская постановка обобщается до полумарковской [20, 21]. Дисконтирование при этом входит в уравнение Беллмана степенью по времени пребывания:

$$Q^*(s, a) = \mathbb{E}_{s' \sim P, \Delta t \sim F} \left[R(s, a) + \gamma^{\Delta t} \max_{a' \in \mathcal{A}} Q^*(s', a') \right], \quad (7)$$

где Δt – время пребывания в состоянии, то есть случайная длительность управляющего шага, определяемая сетевой задержкой в распределенном контуре; $F(\Delta t | s, a)$ – функция распределения времени пребывания; $\gamma^{\Delta t}$ – дисконтирующий множитель, зависящий от фактической длительности перехода (для целого числа примитивных тактов k он принимает вид γ^k). При постоянном единичном шаге $\Delta t \equiv 1$ выражение (7) переходит в обычное уравнение (6). Полумарковское дисконтирование согласует ценность переходов разной длительности и оказывается необходимым при сетевой реализации контура реального времени.

Ценностные методы: Q-learning и DQN. Исторически первыми были табличные методы, представляющие функцию ценности в виде таблицы и обновляющие ее по уравнению (6); классический Q-learning сходится к оптимуму при конечных $\mathcal{S}$ и $\mathcal{A}$ [22]. Замена таблицы нейросетевым аппроксиматором привела к алгоритму DQN (Deep Q-Network), достигшему результата человеческого уровня в дискретных средах [23]. Достоинство этих методов состоит в простоте и доказанной сходимости на конечных пространствах. Однако для настройки регулятора их ограничение оказывается принципиальным: и Q-learning, и DQN извлекают политику через операцию $\max_a Q(s, a)$, выполнимую лишь при дискретном множестве действий. Вектор θ непрерывен, и применение DQN требует дискретизации каждой компоненты на сетку. Грубая сетка ограничивает достижимую точность, а мелкая порождает экспоненциальный рост числа действий по числу настраиваемых параметров. Ценностные подходы к RL-настройке ПИД-регуляторов использовали дискретное пространство действий: в табличном Q-learning обученные Q-таблицы применялись для изменения коэффициентов ПИД-регулятора в зависимости от состояния системы, а в DQN-PID ценностный метод использовался для динамической настройки параметров ПИД-регулятора в дискретном пространстве [24, 25]. При переносе таких подходов на непрерывный вектор θ каждую его компоненту приходится дискретизировать, что ограничивает точность и увеличивает размер пространства действий. Для непрерывной настройки поэтому требуются методы, работающие с θ напрямую. В непрерывных actor-critic-подходах, например, коэффициенты адаптируются DDPG-агентом [26].

Методы градиента политики. Альтернативу образуют методы градиента политики (policy gradient), непосредственно оптимизирующие параметрическую политику по критерию суммарного вознаграждения без операции $\max_a$ и потому естественно работающие с непрерывным $\mathcal{A}$ [27]. Их базовый вариант REINFORCE

обладает высокой дисперсией оценок градиента [28]; методы с ограничением шага обновления TRPO [29] и PPO [30] снижают эту дисперсию ограничением расхождения между последовательными версиями политики. Достоинство семейства состоит в прямой работе с непрерывными действиями и стохастической политикой, удобной для исследования пространства. Минус состоит в том, что чистые on-policy методы требуют большого числа взаимодействий с объектом, поскольку каждое обновление использует только свежесобранные траектории. На объекте управления, где взаимодействие стоит времени стенда и сопряжено с риском, это ограничение делает on-policy подход выборочно невыгодным, поэтому в задачах настройки регуляторов чистые policy-gradient методы применяются реже actor-critic-схем.

Actor-critic для непрерывных действий: DDPG, TD3, SAC. Семейство actor-critic объединяет оба подхода: актер (актор) параметризует политику $\pi_\phi(s)$, а критик обучает функцию ценности, снижающую дисперсию градиента актера. Именно это семейство доминирует в современных работах по RL-настройке ПИД-регулятора и ADRC и наиболее пригодно для непрерывного вектора θ . Алгоритм DDPG (Deep Deterministic Policy Gradient) [31] представляет собой off-policy actor-critic-метод для непрерывных действий с детерминированной политикой, целевыми сетями для стабилизации обучения и буфером опыта для повторного использования собранных данных. Критик обучается минимизацией невязки уравнения (6), а актер обновляется по градиенту ценности с целевым значением

$$y_t = r_t + \gamma Q_{\psi'}(s_{t+1}, \mu_{\phi'}(s_{t+1})), \quad (8)$$

где y_t – целевое значение для обновления критика на шаге t ; r_t – полученное вознаграждение; γ – коэффициент дисконтирования; $\mu_{\phi'}$, $Q_{\psi'}$ – целевые сети актера и критика с медленно обновляемыми параметрами; s_{t+1} – следующее состояние. Детерминированность политики и переиспользование данных из буфера делают DDPG выборочно более экономным, чем on-policy методы, что существенно при ограниченном бюджете взаимодействия с реальным объектом. Поэтому DDPG стал наиболее распространенным в RL-настройке ПИД-регуляторов [32, 33]. Другие варианты его применения представлены в [34, 35]. Этот подход остается востребованным и в относительно новых работах: алгоритм применяют к настройке ПИД-регуляторов для нелинейных многомерных объектов с переменной во времени неопределенностью [36] и к адаптивной настройке в задачах промышленного управления процессами [37].

Развитием DDPG служат два алгоритма, устраняющие его известные дефекты. Алгоритм TD3 (Twin Delayed DDPG) [38], который снижает систематическую переоценку Q -функции, применяя два критика с выбором минимальной оценки, задержанное обновление актера и сглаживание целевой политики шумом; это повышает устойчивость обучения. Алгоритм SAC (Soft Actor-Critic) [39] вводит в целевую функцию энтропийный бонус за стохастичность политики, что усиливает исследование пространства параметров и робастность к локальным минимумам, особенно на ранних этапах обучения. Все три алгоритма работают с непрерывным θ напрямую и потому пригодны для настройки регуляторов; различие между ними сводится к компромиссу между простотой реализации (DDPG), устойчивостью оценки ценности (TD3) и полнотой исследования (SAC). Те же алгоритмы переносят и на настройку ADRC: в недавней работе глубокое RL применялось к адаптивной настройке ADRC для PMSM-привода; в [40] используется TD3-агент. Общий недостаток семейства состоит в требовании большого числа взаимодействий со средой, с объектом настройки, и в отсутствии гарантий устойчивости замкнутого контура на ранних этапах обучения [31].

Модельно-ориентированный RL. Перечисленные выше методы не привязаны к модели (model-free): они не строят модель динамики объекта, расходуя на это много шагов взаимодействия. Модельно-ориентированный RL (model-based RL) обучает приближенную модель переходов и использует ее для виртуальных шагов взаимодействия с моделью объекта при оптимизации политики, что повышает выборочную эффективность. Алгоритм PILCO аппроксимирует динамику гауссовскими процессами и аналитически распространяет неопределенность по траектории, достигая рекордной выборочной эффективности [41]; его расширение применено и к настройке многомерных ПИД-регуляторов [42]. Более поздний алгоритм PETS заменяет гауссовские процессы ансамблем нейронных сетей, что улучшает масштабируемость на объекты большей размерности [43]. Достоинство этих методов состоит в радикальном сокращении числа взаимодействий с объектом. Недостаток логически выводится из их устройства: на трудно настраиваемом нестационарном объекте (1) с насыщением и запаздыванием обучаемая модель сама становится источником погрешности, ошибки модели накапливаются вдоль виртуальной траектории и смещают оптимизацию политики. Эта же причина ограничивает рассмотренные выше модельно-ориентированные адаптивные регуляторы. Поэтому выигрыш в числе взаимодействий частично обесценивается сложностью и хрупкостью модели.

Безопасное обучение с подкреплением. Отдельное направление образует безопасное (safe), или ограниченное (constrained), обучение с подкреплением, критически важное для систем управления, где неудачная политика на ранних этапах способна вывести контур из устойчивости. Формально задача ставится как марковский процесс принятия решений с ограничениями [44]. При переменной длительности переходов постановка обобщается до полумарковской (constrained semi-Markov decision process, CSMDP) [20, 21]: максимизируется суммарное вознаграждение при условии неперевышения ожидаемых стоимостей по заданным ограничениям. Методы решения включают лагранжевы подходы [45], проецирование на множество допустимых политик (CPO) [46] и подходы на основе функций Ляпунова, сертифицирующих устойчивость в процессе обучения [47]. Развитие лагранжевых схем продолжается: в работе множитель Лагранжа сам управляется регулятором с активным подавлением возмущений, что повышает устойчивость соблюдения ограничений при обучении [48]. Достоинство направления состоит в том, что оно прямо отвечает на проблему сохранения устойчивости при обучении. Минус состоит в том, что ни один из перечисленных механизмов не апробирован применительно к параметрической настройке ПИД-регуляторов и ADRC в натурных испытаниях и не дает гарантии устойчивости именно в процессе настройки коэффициентов, а не настройки самой политики. Эта проблема для настройки регуляторов является центральной, и ее развернутый разбор приведен далее в разделе «Нерешенные проблемы».

Популяционное обучение и автоматизация подбора гиперпараметров. Качество любого из перечисленных алгоритмов существенно зависит от гиперпараметров самого RL-агента: скорости обучения, размера буфера опыта, коэффициента дисконтирования γ , параметров шума исследования, весов в функции вознаграждения. Перебор по сетке или случайный поиск [49] требуют последовательного обучения множества агентов и при высокой стоимости каждого эксперимента с реальным объектом управления становятся неприемлемыми. Метод популяционного обучения (Population Based Training, PBT) [50] совмещает параллельный поиск гиперпараметров с эволюционным отбором: поддерживается популяция агентов, и периодически отстающий агент перенимает веса более успешного (этап эксплуатации), после чего его гиперпараметры случайно возмущаются (этап исследования). В отличие от статического поиска, PBT обнаруживает динамические расписания гиперпараметров, то есть их зависимость от

стадии обучения, которые невозможно найти однократным подбором, и делает это в пределах одного параллельного прогона. Развитием подхода служит вариант FIRE PBT с гибким управлением темпом обновления [51]. Достоинство PBT для настройки регуляторов состоит в возможности одновременно обучать популяцию агентов на серии моделей и автоматически адаптировать не только гиперпараметры алгоритма, но и весовые коэффициенты функции вознаграждения. Ограничением остается вычислительная стоимость поддержания популяции, требующая параллельных ресурсов, либо времени при последовательном исполнении, что выводит вопрос реализации на уровень программно-аппаратного комплекса.

Результаты

По результатам сопоставления рассмотренных семейств методов по критериям работы с непрерывным вектором параметров и выборочной эффективности обоснован выбор метода для последующего исследования. Непрерывность вектора исключает табличные методы и DQN, требующие дискретизации с потерей точности. Ограниченный бюджет взаимодействия с трудно настраиваемым объектом (1) делает невыгодными выборочно затратные on-policy методы, а модельно-ориентированный RL, выигрывая в числе взаимодействий, переносит риск на качество обучаемой модели динамики, которое на нестационарном нелинейном объекте заведомо ограничено. На этом фоне off-policy actor-critic-методы образуют рациональный компромисс: они работают с непрерывным θ напрямую и переиспользуют собранные данные из буфера опыта. Из этого семейства в качестве опорного метода естественно выбрать DDPG как простейший off-policy actor-critic-алгоритм для непрерывных действий, ставший опорным в литературе по RL-настройке ПИД-регуляторов [31, 32]; устранение его переоценки Q -функции и недостаточного исследования (приемы TD3 и SAC) и интеграция ограничений устойчивости (safe RL) образуют отдельное направление развития. Для подбора гиперпараметров RL-агента и весов функции вознаграждения естественно выбрать PBT [50]. Выбранные методы работают с непрерывным вектором параметров и выборочно экономны, однако сами по себе не гарантируют устойчивости замкнутого контура на всех этапах обучения и не решают вопрос переноса с модели на реальный объект. Совокупность выявленных ограничений образует четыре нерешенные проблемы, каждая из которых остается открытой для всего класса трудно настраиваемых объектов (1) и задает требования к дальнейшим исследованиям.

Проблема 1: сохранение устойчивости замкнутого контура на всех этапах обучения. Обучение с подкреплением вырабатывает политику по результатам взаимодействия с объектом, а не выводит ее из модели, поэтому на ранних этапах агент применяет к контуру заведомо неоптимальные, а порой и дестабилизирующие значения θ . Для трудно настраиваемого объекта с насыщением и запаздыванием выход за границу допустимого множества θ означает не временную потерю качества, а реальную потерю устойчивости работающего контура. Безопасное обучение с подкреплением предлагает частные механизмы (лагранжевы схемы [45], проецирование на допустимое множество политик [46], ограничение исследования областью притяжения по функции Ляпунова [47]), однако ни один из них натурно не апробирован применительно к параметрической настройке ПИД-регуляторов и ADRC и не дает гарантии устойчивости именно в процессе настройки коэффициентов. Эта проблема остается открытой, потому что контур на реальном объекте нельзя выводить из устойчивости даже на время обучения, а существующие механизмы безопасности проверены на иных постановках.

Проблема 2: выборочная эффективность в условиях реального времени. Модельно-свободные actor-critic-методы требуют от тысяч до миллионов шагов

взаимодействия со средой [29]. Высокая частота контура (1) оставляет на каждый шаг управления единицы миллисекунд, а каждое взаимодействие с реальным объектом стоит времени стенда и сопряжено с риском, поэтому требуемый объем данных на работающем объекте либо недоступен, либо недопустим. Модельно-ориентированный RL повышает выборочную эффективность ценой построения модели динамики, качество которой на нестационарном нелинейном объекте заведомо ограничено, а перебор гиперпараметров по сетке требует последовательного обучения множества агентов и при финансово и технически дорогом эксперименте невыгоден. Проблема остается открытой, потому что выборочная эффективность не отделена от требования реального времени: метод должен укладываться в частоту контура и одновременно обходиться доступным числом взаимодействий.

Проблема 3: перенос с модели на объект без гарантий сохранения качества. Естественный способ обойти проблемы 1 и 2 состоит в том, чтобы обучать агента на модели объекта, а затем переносить найденную политику на реальный контур. Однако модель (1) редко известна неточно, и расхождение между моделью и объектом (reality gap) обесценивает перенос: политика, оптимальная в симуляции, на реальном объекте может не минимизировать (2). Методы доменной рандомизации [52] и обзоры переноса [53] повышают робастность агента к неопределенности модели, но гарантий отсутствия деградации качества при переносе, как правило, не дают. Проблема остается открытой, потому что для трудно настраиваемого объекта обучение на модели не освобождает от настройки на самом объекте.

Проблема 4: дефицит натурной валидации. Большинство работ по RL-настройке регуляторов ограничивается экспериментом в симуляции; полноценная натурная валидация на реальных объектах представлена лишь в немногих исследованиях [54], что отражает сохраняющийся разрыв между теорией и практикой. Низкая воспроизводимость глубокого RL, т. е. чувствительность результата к гиперпараметрам, случайной инициализации и деталям реализации [55], усугубляет этот разрыв: без натурального эксперимента со статистической обработкой заявленное преимущество метода нельзя считать объективным с достаточной степенью достоверности результатов. Проблема остается открытой, потому что натурный эксперимент со статистическим подтверждением остается редкостью именно там, где он наиболее необходим.

Совокупность четырех проблем позволяет уточнить формулировку научной проблемы, поставленной во введении. Отсутствует метод и его программно-аппаратная реализация автоматической настройки и онлайн-адаптации регуляторов трудно настраиваемых объектов, который одновременно сохраняет устойчивость замкнутого контура в процессе обучения, выборочно эффективен в условиях реального времени и подтвержден натурным экспериментом. Каждое из трех требований фальсифицируемо: сохранение устойчивости проверяется тем, что на всем интервале обучения значения θ не выводят контур из допустимого множества Θ ; выборочная эффективность – достижением минимума (2) за бюджет взаимодействия, совместимый с частотой контура (1); подтвержденность – снижением функционала (2) на реальном стенде относительно классической настройки при статистически значимом результате.

Обсуждение

Полученные результаты показывают, что четыре выделенные проблемы взаимосвязаны: каждая из них задает отдельное требование к искомому методу настройки и соответствующее направление дальнейших исследований. Соответствие выявленных проблем требованиям к методу и направлениям дальнейшей работы представлено в Таблице 1.

Таблица 1 – Соответствие нерешенных проблем направлениям дальнейших исследований
Table 1 – Mapping of open problems to directions for further research

Нерешенная проблема	Требование к методу настройки	Направление дальнейших исследований
Анализ методов и выявление разрывов (задача обзора)	Единый критерий сопоставления семейств методов	Систематический анализ методов RL-настройки и выделение нерешённых проблем
Проблема 1: устойчивость замкнутого контура при обучении	Сохранение устойчивости на всех этапах обучения	Разработка метода настройки с механизмом самосохранения устойчивости
Проблемы 2 и 3: выборочная эффективность в реальном времени и перенос с модели на объект	Выборочная эффективность в темпе контура без модели-посредника	Построение программно-аппаратного комплекса реального времени с популяционным обучением
Проблема 4: дефицит натурной валидации	Подтверждение на реальном объекте со статистической обработкой	Экспериментальная проверка метода на натурных стендах

Первое направление образует сам систематический анализ методов настройки, итог которого изложен в настоящей статье: проведен разбор классических, адаптивных и обучаемых методов и выделены четыре нерешенные проблемы. Обзор не решает проблему, а определяет ее и вводит единый критерий, по которому сопоставляются семейства методов, занимая тем самым первую ступень в дальнейшей работе.

Проблема 1 (устойчивость при обучении) задает направление разработки метода настройки, сохраняющего устойчивость. Задачу настройки естественно формализовать как полумарковский процесс принятия решений с ограничениями (CSMDP), а устойчивость обеспечивать функцией вознаграждения с механизмом самосохранения (self-preserving penalty), штрафующим выход качества за уровень опорной настройки, так что на всем интервале обучения значения θ не выводят контур из допустимого множества Θ . Перенос подхода на ADRC, для которого автоматический подбор полос ω_o и ω_c снимает главное препятствие к применению, остается открытой перспективой: ни один из существующих механизмов безопасного обучения, как показано выше при рассмотрении проблемы сохранения устойчивости, пока не апробирован именно для параметрической настройки ПИД-регуляторов и ADRC.

Проблемы 2 и 3 (выборочная эффективность в реальном времени и перенос с модели на объект) задают направление построения программно-аппаратного комплекса реального времени. Распределенная платформа с обобщенным популяционным обучением гетерогенной популяции моделей способна обеспечить работу контура управления и определение гиперпараметров в реальном времени при обучении непосредственно на объекте. Обучение на объекте снимает зависимость от точности модели и способно решить обе проблемы сразу: выборочная эффективность достигается параллельным популяционным поиском, а проблема переноса решается тем, что модель-посредник не используется. На наш взгляд, именно вынесение реализации на уровень

комплекса, а не отдельного алгоритма, отличает это направление от большинства известных работ, ограниченных одиночным агентом в симуляции.

Проблема 4 (дефицит натурной валидации) задает направление экспериментальной проверки метода. Достоверность и эффективность метода имеет смысл подтверждать на натурных стендах с количественным сравнением с настройкой по Циглеру-Николсу (3) и статистической обработкой результатов по нескольким прогонам; без такого подтверждения, как отмечено выше при рассмотрении дефицита натурной валидации, заявленное преимущество метода нельзя считать установленным. Четыре направления в совокупности показывают, что заявленная научная проблема не может быть решена частично: метод обязан одновременно сохранять устойчивость, быть выборочно эффективным в реальном времени и опираться на натурное подтверждение.

Заключение

Главный вывод обзора состоит в том, что задача автоматической настройки и онлайн-адаптации регуляторов трудно настраиваемых объектов не решается ни одним семейством методов по отдельности, поскольку требует одновременного выполнения трех фальсифицируемых условий: сохранения устойчивости замкнутого контура на всех этапах обучения, выборочной эффективности в темпе работы контура и подтверждения на реальном объекте. Если разбирать методы по единому критерию (пригодность для непрерывного вектора θ и выборочная эффективность при ограниченном бюджете взаимодействия), то классические правила и адаптивные схемы оказываются неэффективными из-за однократного расчета или зависимости от явной идентификации модели, табличные RL-методы и DQN – из-за дискретизации, on-policy методы – из-за расхода взаимодействий, а модельно-ориентированный RL переносит риск на качество обучаемой модели динамики. Рациональным компромиссом остаются off-policy actor-critic-методы (DDPG с приемами TD3 и SAC), дополненные популяционным обучением для подбора гиперпараметров; именно к этому выбору приводит сквозной критерий обзора.

Вместе с тем выбранное семейство само по себе не закрывает задачу. Безопасное обучение с подкреплением располагает механизмами сохранения устойчивости (лагранжевы схемы, проецирование на допустимое множество, сертификаты Ляпунова), однако ни один из них пока не апробирован именно для параметрической настройки ПИД-регуляторов и ADRC, а не для настройки самой политики: это и есть наиболее существенный незакрытый пробел, выявленный обзором. Отсюда вытекают и направления дальнейших исследований: формализация задачи настройки как процесса принятия решений с ограничениями и разработка метода с механизмом самосохранения устойчивости; построение программно-аппаратного комплекса реального времени с популяционным обучением, ведущего обучение непосредственно на объекте и потому снимающего проблему переноса с модели; экспериментальная проверка метода на натурных стендах со статистической обработкой. Систематизация разрозненных подходов по единому критерию и сведение их к четырем фальсифицируемым проблемам, на наш взгляд, и составляют практическую пользу обзора: они переводят общую постановку в проверяемые требования к будущему методу.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Minorsky N. Directional stability of automatically steered bodies. *Journal of the American Society for Naval Engineers*. 1922;34(2):280–309. <https://doi.org/10.1111/j.1559-3584.1922.tb04958.x>

2. Ziegler J.G., Nichols N.B. Optimum settings for automatic controllers. *Transactions of the ASME*. 1942;64(8):759–765. <https://doi.org/10.1115/1.4019264>
3. Desborough L., Miller R. Increasing customer value of industrial control performance monitoring – Honeywell's experience. In: *Chemical Process Control-VI: Assessment and New Directions for Research: Proceedings of the Sixth International Conference on Chemical Process Control, 07–12 January 2001, Tucson, AZ, USA*. CACHE, AIChE; 2002. P. 169–189.
4. Bauer M., Horch A., Xie L., Jelali M., Thornhill N. The current state of control loop performance monitoring – A survey of application in industry. *Journal of Process Control*. 2016;38:1–10. <https://doi.org/10.1016/j.jprocont.2015.11.002>
5. Åström K.J., Hägglund T. *Advanced PID Control*. ISA – Instrumentation, Systems, and Automation Society; 2006. 460 p.
6. Gaing Z.-L. A particle swarm optimization approach for optimum design of PID controller in AVR system. *IEEE Transactions on Energy Conversion*. 2004;19(2):384–391. <https://doi.org/10.1109/TEC.2003.821821>
7. Fleming P.J., Purshouse R.C. Evolutionary algorithms in control systems engineering: A survey. *Control Engineering Practice*. 2002;10(11):1223–1241. [https://doi.org/10.1016/S0967-0661\(02\)00081-3](https://doi.org/10.1016/S0967-0661(02)00081-3)
8. Sutton R.S., Barto A.G. *Reinforcement Learning: An Introduction*. Cambridge: MIT Press; 2018. 552 p.
9. Dorf R.C., Bishop R.H. *Modern Control Systems*. Upper Saddle River: Prentice Hall; 2001. 831 p.
10. Cohen G.H., Coon G.A. Theoretical consideration of retarded control. *Transactions of the ASME*. 1953;75:827–834. <https://doi.org/10.1115/1.4015451>
11. Åström K.J., Hägglund T. Automatic tuning of simple regulators with specifications on phase and amplitude margins. *Automatica*. 1984;20(5):645–651. [https://doi.org/10.1016/0005-1098\(84\)90014-1](https://doi.org/10.1016/0005-1098(84)90014-1)
12. Åström K.J., Wittenmark B. *Adaptive Control*. Reading: Addison-Wesley; 1995. 574 p.
13. Ioannou P., Fidan B. *Adaptive Control Tutorial*. Philadelphia: SIAM; 2006. 405 p. <https://doi.org/10.1137/1.9780898718652>
14. Han J. From PID to active disturbance rejection control. *IEEE Transactions on Industrial Electronics*. 2009;56(3):900–906. <https://doi.org/10.1109/TIE.2008.2011621>
15. Gao Zh. Scaling and bandwidth-parameterization based controller tuning. In: *2003 American Control Conference, 04–06 June 2003, Denver, CO, USA*. IEEE; 2003. P. 4989–4996. <https://doi.org/10.1109/ACC.2003.1242516>
16. Gao Zh. Active disturbance rejection control: a paradigm shift in feedback control system design. In: *2006 American Control Conference, 14–16 June 2006, Minneapolis, MN, USA*. IEEE; 2006. <https://doi.org/10.1109/ACC.2006.1656579>
17. Huang Y., Xue W. Active disturbance rejection control: methodology and theoretical analysis. *ISA Transactions*. 2014;53(4):963–976. <https://doi.org/10.1016/j.isatra.2014.03.003>
18. Khalil H.K. *High-Gain Observers in Nonlinear Feedback Control*. Philadelphia: SIAM; 2017. 323 p. <https://doi.org/10.1137/1.9781611974867>
19. Fu C., Tan W. Tuning of linear ADRC with known plant information. *ISA Transactions*. 2016;65:384–393. <https://doi.org/10.1016/j.isatra.2016.06.016>
20. Sutton R.H., Precup D., Singh S. Between MDPs and semi-MDPs: a framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*. 1999;112(1–2):181–211. [https://doi.org/10.1016/S0004-3702\(99\)00052-1](https://doi.org/10.1016/S0004-3702(99)00052-1)
21. Bradtke S., Duff M.O. Reinforcement learning methods for continuous-time Markov decision problems. In: *Advances in Neural Information Processing Systems 7: Annual*

- Conference on Neural Information Processing Systems*, 28 November – 01 December 1994, Denver, CO, USA. Cambridge: MIT Press; 1995. P. 393–400.
22. Watkins Ch.J.C.H., Dayan P. *Q-learning*. *Machine Learning*. 1992;8(3–4):279–292. <https://doi.org/10.1007/BF00992698>
 23. Mnih V., Kavukcuoglu K., Silver D., et al. Human-level control through deep reinforcement learning. *Nature*. 2015;518(7540):529–533. <https://doi.org/10.1038/nature14236>
 24. Shi Q., Lam H.-K., Xiao B., Tsai Sh.-H. Adaptive PID controller based on Q-learning algorithm. *CAAI Transactions on Intelligence Technology*. 2018;3(4):235–244. <https://doi.org/10.1049/trit.2018.1007>
 25. Wang F., Ren B., Liu Y., Cui B. Tracking moving target for 6 degree-of-freedom robot manipulator with adaptive visual servoing based on deep reinforcement learning PID controller. *Review of Scientific Instruments*. 2022;93(4):045108. <https://doi.org/10.1063/5.0087561>
 26. Qin Y., Zhang W., Shi J., Liu J. Improve PID controller through reinforcement learning. In: *2018 IEEE CSAA Guidance, Navigation and Control Conference, 10–12 August 2018, Xiamen, China*. IEEE; 2018. <https://doi.org/10.1109/GNCC42960.2018.9019095>
 27. Sutton R.S., McAllester D.A., Singh S., Mansour Y. Policy gradient methods for reinforcement learning with function approximation. In: *Advances in Neural Information Processing Systems 12: Annual Conference on Neural Information Processing Systems, 29 November – 04 December 1999, Denver, CO, USA*. Cambridge: MIT Press; 2000. P. 1057–1063.
 28. Williams R.J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*. 1992;8(3):229–256. <https://doi.org/10.1007/BF00992696>
 29. Schulman J., Levine S., Abbeel P., Jordan M.I., Moritz Ph. Trust region policy optimization. In: *Proceedings of the 32nd International Conference on Machine Learning, 06–11 July 2015, Lille, France*. JMLR; 2015. P. 1889–1897.
 30. Schulman J., Wolski F., Dhariwal P., Radford A., Klimov O. *Proximal policy optimization algorithms*. arXiv. URL: <https://arxiv.org/abs/1707.06347> [Accessed 13th July 2026].
 31. Lillicrap T.P., Hunt J.J., Pritzel A., et al. Continuous control with deep reinforcement learning. In: *Proceedings of the 4th International Conference on Learning Representations, 02–04 May 2016, San Juan, Puerto Rico*. 2016. URL: <https://arxiv.org/abs/1509.02971>
 32. Lakhani A.I., Chowdhury M.A., Lu Q. Stability-preserving automatic tuning of PID control with reinforcement learning. *Complex Engineering Systems*. 2022;2(1):3. <https://doi.org/10.20517/ces.2021.15>
 33. Arshad J., Khan A., Aftab M., et al. Deep deterministic policy gradient to regulate feedback control systems using reinforcement learning. *Computers, Materials & Continua*. 2022;71(1):1153–1169. <https://doi.org/10.32604/cmc.2022.021917>
 34. Lee D., Lee S.J., Yim S.C. Reinforcement learning-based adaptive PID controller for DPS. *Ocean Engineering*. 2020;216(11). <https://doi.org/10.1016/j.oceaneng.2020.108053>
 35. Lawrence N.P., Forbes M.G., Loewen Ph.D., McClement D.G., Backström J.U., Gopaluni R.Bh. Deep reinforcement learning with shallow controllers: An experimental application to PID tuning. *Control Engineering Practice*. 2022;121(5):105046. <https://doi.org/10.1016/j.conengprac.2021.105046>
 36. Wang H., Ricardez-Sandoval L.A. A deep reinforcement learning-based PID tuning strategy for nonlinear MIMO systems with time-varying uncertainty. *IFAC-PapersOnLine*. 2024;58(14):887–892. <https://doi.org/10.1016/j.ifacol.2024.08.449>

37. Veerasamy G., Balaji S., Kadirvelu Th., Ramasamy V. Reinforcement Learning Based Adaptive PID Controller for a Continuous Stirred Tank Heater Process. *Iranian Journal of Chemistry and Chemical Engineering*. 2025;44(1):265–282. <https://doi.org/10.30492/ijcce.2024.2029225.6615>
38. Fujimoto S., van Hoof H., Meger D. Addressing function approximation error in actor-critic methods. In: *Proceedings of the 35th International Conference on Machine Learning, 10–15 July 2018, Stockholm, Sweden*. PMLR; 2018. P. 1582–1591.
39. Haarnoja T., Zhou A., Abbeel P., Levine S. Soft actor-critic: off-policy maximum entropy deep reinforcement learning with a stochastic actor. In: *Proceedings of the 35th International Conference on Machine Learning, 10–15 July 2018, Stockholm, Sweden*. PMLR; 2018. P. 1856–1865.
40. Wang Y., Fang Sh., Hu J., Huang D. A novel active disturbance rejection control of PMSM based on deep reinforcement learning for more electric aircraft. *IEEE Transactions on Energy Conversion*. 2023;38(2):1461–1470. <https://doi.org/10.1109/TEC.2023.3235927>
41. Deisenroth M.P., Rasmussen C.E. PILCO: A model-based and data-efficient approach to policy search. In: *Proceedings of the 28th International Conference on Machine Learning, 28 June – 02 July 2011, Bellevue, WA, USA*. Madison: Omnipress; 2011. P. 465–472.
42. Doerr A., Nguyen-Tuong D., Marco A., Schaal S., Trimpe S. Model-based policy search for automatic tuning of multivariate PID controllers. In: *2017 IEEE International Conference on Robotics and Automation, 29 May – 03 June 2017, Singapore*. IEEE; 2017. P. 5295–5301. <https://doi.org/10.1109/ICRA.2017.7989622>
43. Chua K., Calandra R., McAllister R., Levine S. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In: *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems, 03–08 December 2018, Montréal, Canada*. 2018. P. 4759–4770.
44. Altman E. *Constrained Markov Decision Processes*. Boca Raton: CRC Press; 1999. 256 p.
45. Stooke A., Achiam J., Abbeel P. Responsive safety in reinforcement learning by PID Lagrangian methods. In: *Proceedings of the 37th International Conference on Machine Learning, 13–18 July 2020, Virtual Event*. PMLR; 2020. P. 9133–9143.
46. Achiam J., Held D., Tamar A., Abbeel P. Constrained policy optimization. In: *Proceedings of the 34th International Conference on Machine Learning, 06–11 August 2017, Sydney, NSW, Australia*. PMLR; 2017. P. 22–31.
47. Berkenkamp F., Turchetta M., Schoellig A.P., Krause A. Safe model-based reinforcement learning with stability guarantees. In: *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems, 04–09 December 2017, Long Beach, CA, USA*. 2017. P. 908–918.
48. Zhang M., Zhang H., Ji J., Yang Y., Sun Y. Enhance the safety in reinforcement learning by ADRC Lagrangian methods. arXiv. URL: <https://arxiv.org/abs/2601.18142> [Accessed 13th July 2026].
49. Bergstra J., Bengio Y. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*. 2012;13(2):281–305.
50. Jaderberg M., Dalibard V., Osindero S., et al. Population based training of neural networks. arXiv. URL: <https://arxiv.org/abs/1711.09846> [Accessed 13th July 2026].
51. Dalibard V., Jaderberg M. Faster Improvement Rate Population Based Training. arXiv. URL: <https://arxiv.org/abs/2109.13800> [Accessed 13th July 2026].
52. Tobin J., Fong R., Ray A., Schneider J., Zaremba W., Abbeel P. Domain randomization for transferring deep neural networks from simulation to the real world. In: *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems, 24–28 September*

- 2017, Vancouver, Canada. IEEE; 2017. P. 23–30. <https://doi.org/10.1109/IROS.2017.8202133>
53. Zhao W., Queralta J.P., Westerlund T. Sim-to-real transfer in deep reinforcement learning for robotics: a survey. In: *2020 IEEE Symposium Series on Computational Intelligence, 01–04 December 2020, Canberra, ACT, Australia*. IEEE; 2020. P. 737–744. <https://doi.org/10.1109/SSCI47803.2020.9308468>
54. Reutov A., Vorobey S., Katanskiy A., et al. Optical stabilization for laser communication satellite systems through proportional-integral-derivative (PID) control and reinforcement learning approach. *Review of Scientific Instruments*. 2025;96(3):035121. <https://doi.org/10.1063/5.0239029>
55. Henderson P., Islam R., Bachman Ph., Pineau J., Precup D., Meger D. Deep reinforcement learning that matters. *Proceedings of the AAAI Conference on Artificial Intelligence*. 2018;32(1). <https://doi.org/10.1609/aaai.v32i1.11694>

ИНФОРМАЦИЯ ОБ АВТОРЕ / INFORMATION ABOUT THE AUTHOR

Воробей Сергей Сергеевич, соискатель, **Sergey S. Vorobey**, PhD candidate, Moscow
Московский технический университет связи и Technical University of Communications and
информатики, Москва, Российская Федерация. Informatics, Moscow, the Russian Federation.
e-mail: sv@rootml.com
ORCID: [0000-0002-5259-6841](https://orcid.org/0000-0002-5259-6841)

*Статья поступила в редакцию 03.08.2026; одобрена после рецензирования 06.10.2026;
принята к публикации 16.10.2026.*

*The article was submitted 03.08.2026; approved after reviewing 06.10.2026;
accepted for publication 16.10.2026.*