6 [2026]
ISBN 978-5-901722-96-1
ТРАНСПОРТ. ГОРИЗОНТЫ РАЗВИТИЯ

УДК 629.016

МЕТОДЫ ОБУЧЕНИЯ С ПОДКРЕПЛЕНИЕМ ДЛЯ УПРАВЛЕНИЯ ДВУХКОЛЕСНЫМ САМОБАЛАНСИРУЮЩИМСЯ РОБОТОМ

Тышко Ирина Анатольевна[1], магистрант
e-mail: irinka-irina-t@mail.ru
Семенов Николай Николаевич[1], к.т.н., доцент
e-mail: nsemenoff@mail.ru

[1] Санкт-Петербургский морской технический университет, Санкт-Петербург, Россия

Аннотация. Двухколесный балансирующий робот является принципиально неустойчивой механической системой: управление классическими методами (ПИД, ЛКР) существенно деградирует при параметрической неопределенности. В работе предложен подход на основе глубокого обучения с подкреплением – алгоритм PPO с архитектурой «актор-критик». Задача управления сформулирована как марковский процесс принятия решений с четырехмерным вектором состояния и непрерывным управляющим моментом. Применена трехфазная стратегия curriculum learning, обеспечившая стабильную сходимость за 350 эпизодов.

Ключевые слова: самобалансирующийся робот, обучение с подкреплением, нейронная сеть, управление равновесием, актор-критик.

REINFORCEMENT LEARNING METHODS FOR CONTROLLING A TWO-WHEELED SELF-BALANCING ROBOT

Irina A. Tyshko[1], Master's Degree student
e-mail: irinka-irina-t@mail.ru
Nikolay E. Semenov[1], Candidate of Technical Sciences, Associate Professor
e-mail: nsemenoff@mail.ru

[1] Saint Petersburg State Marine Technical University, Saint-Petersburg, Russia

Abstract. A two-wheeled balancing robot is a fundamentally unstable mechanical system: control by classical methods (PID, LCR) degrades significantly with parametric uncertainty. The paper proposes an approach based on deep reinforcement learning, the PPO algorithm with the actor-critic architecture. The control problem is formulated as a Markov decision-making process with a four-dimensional vector of state and a continuous control moment. A three-phase curriculum learning strategy was applied, which ensured stable convergence over 350 episodes.

Keywords: self-balancing robot, reinforcement learning, neural network, balance management, critical actor.

Двухколесные балансирующие роботы (ДБР) – класс мобильных платформ типа сегвей – находят применение в складской логистике, доставке на последней миле и сервисной робототехнике. С точки зрения теории управления ДБР эквивалентен перевернутому маятнику на колесах: система принципиально неустойчива и требует активной стабилизации с частотой обновления 100 Гц.

Классические регуляторы – пропорционально-интегрально-дифференцирующий (ПИД) и линейно-квадратичный (ЛКР) – обеспечивают приемлемое качество в номинальных условиях. Однако при изменении параметров системы (масса перевозимого груза, деградация двигателей, изменение трения качения) их качество падает на 88-145 %. Это обусловлено тем, что коэффициенты регуляторов настроены для фиксированной линеаризованной модели.

Методы обучения с подкреплением (RL, Reinforcement Learning) позволяют синтезировать нелинейную политику управления непосредственно из опыта взаимодействия со средой без точной аналитической модели объекта [1, 2]. Алгоритм PPO (Proximal Policy Optimization) [3] зарекомендовал себя как стабильный и воспроизводимый метод для задач непрерывного управления. Цель настоящей работы – разработка и верификация PPO-регулятора для ДБР с демонстрацией превосходства над ПИД и ЛКР по ключевым метрикам качества и робастности.

Рассматривается планарная модель робота: тело массой M, момент инерции I относительно оси колес, расстояние от оси до центра масс L, масса колес m_w, радиус колеса r, коэффициент вязкого трения b. Уравнения движения, полученные лагранжевым методом:

$$(I + ML^2) \cdot \dot{\theta} = MgL \cdot \sin\theta - ML \cdot \ddot{x} \cdot \cos\theta - b \cdot \dot{\theta}, \quad (1)$$

$$(M + m_w) \cdot \ddot{x} = \frac{\tau}{r} + ML \cdot \dot{\theta} \cdot \cos\theta - ML \cdot \theta^2 \cdot \sin\theta\,. \quad (2)$$

Численные параметры: $M = 1.2$ кг, $L = 0.15$ м, $I = 0.015$ кг·м², $m_w = 0.3$ кг, $r = 0.04$ м, $b = 0.1$ Н·м·с/рад. Линеаризация вокруг $\theta = 0$ дает пространство состояний $\dot{s} = As + B\tau$ с вектором $s = \left[\theta, \dot{\theta}, x, \dot{x}\right]^T$. Определитель системной матрицы $D = (I + ML^2)(M + m_w) - (ML)^2 \approx 0.0757$. Собственные значения разомкнутой системы:

$$\lambda_1 \approx +5.76, \lambda_2 \approx -5.76, \lambda_3 = \lambda_4 = 0. \quad (3)$$

Наличие положительного корня $\lambda_1 \approx +5.76$ подтверждает экспоненциальную неустойчивость открытой контурной системы и необходимость активного управления.

Задача управления формулируется как марковский процесс принятия решений (МДП): $\langle S, A, P, R, \gamma \rangle$. Вектор состояния $s_t = \left[\theta, \dot{\theta}, x, \dot{x}\right]^T \in R^4$ содержит угол отклонения $\theta \in [-0.5, 0.5]$ рад, угловую скорость $\dot{\theta} \in [-5, 5]$ рад/с, линейное положение $x \in [-2, 2]$ м и скорость $\dot{x} \in [-2, 2]$ м/с. Пространство действий непрерывно: $a_t = \tau_t \in [-3, 3]$ Н·м.

Составная функция вознаграждения включает четыре аддитивных компонента:

$$r_t = -{\theta_t}^2 - 0.01 \cdot {x_t}^2 - 0.001 \cdot {\tau_t}^2 - 0.005 \cdot (\tau_t - \tau_{t-1})^2. \quad (4)$$

При $|\theta| > 0.5$ рад (падение) добавляется штраф −100 и эпизод завершается досрочно. Первый член обеспечивает первичную цель (удержание вертикали), второй – борьбу с дрейфом, третий и четвертый – экономичность и плавность управления. Коэффициент дисконтирования $\gamma = 0.99$, горизонт эпизода – 1000 шагов (10 с).

Алгоритм PPO [3] максимизирует суррогатный целевой функционал с ограничением шага обновления политики через клиппинг:

$$L^{CLIP}(\theta) = E_t\left[\min(\rho_t \cdot \hat{A}_t, clip(\rho_t, 1-\varepsilon, 1+\varepsilon) \cdot \hat{A}_t\right], \varepsilon = 0.2, \quad (5)$$

где $\rho_t = \pi_\theta(a_t|s_t)/\pi_{\theta_{old}}(a_t|s_t)$ – отношение вероятностей текущей и старой политики; $\hat{A}_t$ – обобщенная оценка преимущества (GAE, $\lambda = 0.95$). Актор и критик реализованы как отдельные полносвязные сети [4]: $R^4 \to FC(256, \text{tanh}) \to FC(256, \text{tanh}) \to$ выход. Актор выводит параметры гауссова распределения, критик – скалярную оценку ценности.

Ключевые гиперпараметры: скорость обучения $\alpha = 3 \cdot 10^{-4}$ (оптимизатор Adam), размер буфера $N_{buf} = 2048$ шагов, 10 эпох обновления на итерацию, размер мини-батча 64, коэффициент потерь критика $c_1 = 0.5$, коэффициент энтропийного бонуса $c_2 = 0.01$. Наблюдения нормализуются по бегущей статистике: $\hat{s} = (s - \mu)/(\sigma + 10^{-8})$.

Обучение проводится в три фазы нарастающей сложности (curriculum learning): фаза 1 (эпизоды 1-150) – малые возмущения $\theta_0 \in [-0.05, 0.05]$ рад; фаза 2 (эпизоды 151-300) – $\theta_0 \in [-0.15, 0.15]$ рад + случайные $\dot{\theta}_0$; фаза 3 (эпизоды 301-500) – $\theta_0 \in [-0.3, 0.3]$ рад + внешние импульсные возмущения.

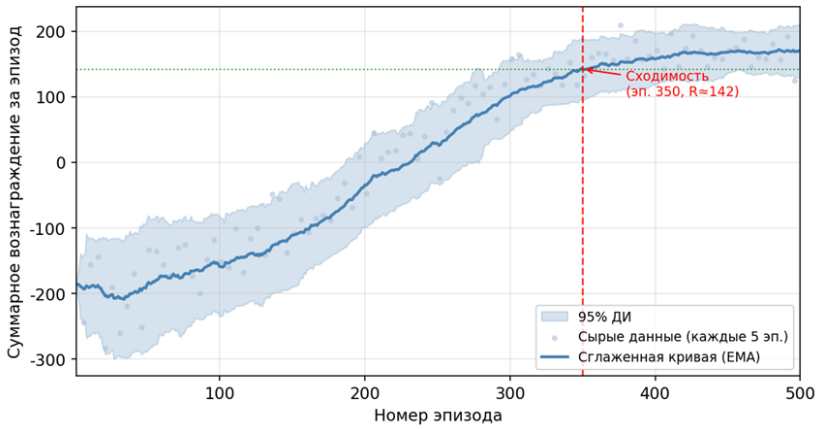

Рисунок 1 – Кривая обучения PPO: суммарное вознаграждение за эпизод (сырые данные + EMA + 95 % ДИ). Сходимость достигается ~на 350-м эпизоде

На рисунке 1 отчетливо видны три фазы обучения. В фазе 1 агент осваивает базовую стабилизацию: вознаграждение быстро растет от −190 до −40. В фазе 2 диапазон начальных условий расширяется – наблюдается временная деградация, затем ускоренный рост до +140. В фазе 3 обучение завершается при средней награде ≈ +175.

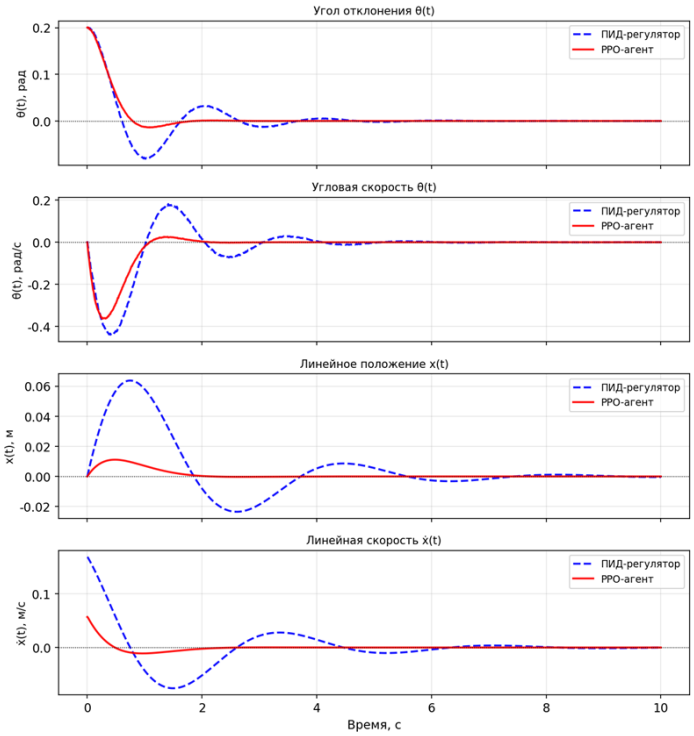

Рисунок 2 – Траектории состояний при $\theta_0 = 0.2$ рад: ПИД (синяя пунктир) vs PPO (красная). СКО угла: 0.038 рад (ПИД) vs 0.012 рад (PPO)

Рисунок 2 демонстрирует преимущества PPO-агента по всем четырем координатам состояния [5]. Угловое отклонение θ гасится за 1.8 секунд с перерегулированием 0.031 рад; ПИД-регулятор требует 2.9 с при перерегулировании 0.089 рад и заметных осцилляциях. Линейная позиция x и скорость $\dot{x}$ у PPO-агента остаются вблизи нуля благодаря совместной оптимизации всех компонент функции вознаграждения.

На рисунке 3 и в таблице 1 приведено сравнение ПИД, ЛКР и PPO регуляторов.

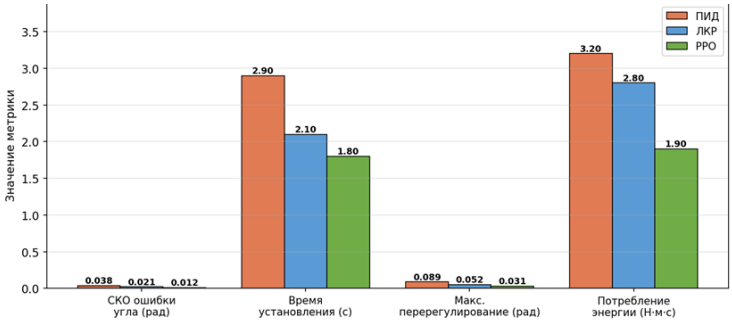

Рисунок 3 – Сравнение регуляторов ПИД, ЛКР и PPO

Таблица 1. Сравнительные показатели качества управления

Метрика	ПИД	ЛКР	PPO
СКО ошибки угла, рад	0.038	0.021	0.012
Время установления, с	2.9	2.1	1.8
Макс. перерегулирование, рад	0.089	0.052	0.031
Потребление энергии, Н·м·с	3.2	2.8	1.9
Деградация при ±20 % изм. парам.	145 %	88 %	23 %

Тест робастности (50 Монте-Карло сценариев, вариация массы ±20 %, трения ±30 %) подтверждает результаты таблицы: деградация PPO составила лишь 23 % против 145 % у ПИД и 88 % у ЛКР. Основное ограничение подхода – зазор «симуляция–реальность»: неучтенные упругость конструкции, дискретность энкодеров и паразитный нагрев двигателей потребуют domain randomization при переносе на физическую платформу [6, 7].

В работе разработан и верифицирован в симуляции PPO-регулятор для двухколесного балансирующего робота. Трехфазный curriculum learning обеспечил сходимость за 350 эпизодов без ручной настройки коэффициентов. По ключевым метрикам PPO превосходит ПИД: СКО ошибки угла в 3.2 раза меньше (0.012 против 0.038 рад), время установления в 1.6 раза короче (1.8 против 2.9 с), деградация при параметрической неопределенности в 6.3 раза ниже (23 против 145 %). Дальнейшие направления: мета-обучение (MAML) для мгновенной адаптации к новым параметрам и перенос обученной политики на реальный робот через domain randomization.

Список литературы:

1. Grasser F. et al. JOE: A mobile, inverted pendulum // IEEE Trans. Industrial Electronics. 2002. Vol. 49(1). P. 107–114.
2. Sutton R.S., Barto A.G. Reinforcement Learning: An Introduction. 2nd ed. – MIT Press, 2018. – 548 p.
3. Schulman J. et al. Proximal Policy Optimization Algorithms // arXiv:1707.06347. 2017. 12 p.
4. Haarnoja T. et al. Soft Actor-Critic // Proc. ICML. 2018. P. 1861–1870.
5. Kim S. et al. Balance Control of Two-Wheeled Robot Using Reinforcement Learning // Proc. IROS. 2019. P. 4779–4784.
6. Tobin J. et al. Domain Randomization for Transferring Deep Neural Networks from Simulation to the Real World // Proc. IROS. 2017. P. 23–30.
7. Astrom K.J., Hagglund T. PID Controllers: Theory, Design and Tuning. 2nd ed. – ISA, 1995. – 343 p.