ГИБРИДНЫЙ ФРЕЙМВОРК ДЛЯ РОБАСТНОГО УПРАВЛЕНИЯ РЕКТИФИКАЦИОННЫМИ КОЛОННАМИ В УСЛОВИЯХ НЕОПРЕДЕЛЁННОСТИ РЕЖИМНЫХ ПАРАМЕТРОВ

Дошанова Малика Юлдашовна,

ТУИТ имени Мухаммада ал-Хоразмий, доцент
E-mail: yuldashevna12@gmail.com

K E Y W O R D S

цифровой двойник, обучение с подкреплением, ректификационная колонна, робастное управление, неопределённость, MPC, control-informed RL, domain randomization

A B S T R A C T

В работе предложен гибридный фреймворк, объединяющий цифровой двойник колонны с политикой обучения с подкреплением, обучаемой офлайн на двойнике при явной рандомизации параметров неопределённости и дообучаемой онлайн по данным реального объекта. Разработана динамическая модель колонны по тарелкам, формализация задачи управления как марковского процесса принятия решений и его робастного/распределительно-робастного расширения, схема байесовской ассимиляции данных для актуализации двойника, а также структура, сочетающая интегральную обратную связь с нейросетевой политикой.

Введение

Ректификация обеспечивает разделение до 90–95 % продуктов нефтехимии и основного органического синтеза и потребляет, по разным оценкам, от 40 до 50 % энергии всей отрасли разделения смесей. Малое отклонение от оптимального режима приводит к заметным потерям энергии либо к выходу состава продукта за спецификацию. Управление осложняется рядом факторов:

1. **Сильная многосвязность** каналов «расход орошения – состав дистиллята» и «теплоподвод в кубе – состав кубового продукта», усугубляемая изменением относительной матрицы усиления (RGA) при смене режима;

2. **Значительное транспортное запаздывание**, обусловленное гидравлическим и тепловым временем пробега по тарелкам;

3. **Существенная параметрическая и структурная неопределённость**: колебания состава и расхода питания, загрязнение и деградация насадки/тарелок, дрейф КПД тарелок, погрешности физико-химических моделей равновесия пар–жидкость.

Классические подходы – децентрализованный ПИД с расстройкой для учёта взаимодействия контуров, а также линейный или нелинейный MPC на основе идентифицированной модели, – обеспечивают удовлетворительное качество регулирования вблизи точки линеаризации, но требуют либо консервативной настройки, либо частой переидентификации модели при значимом изменении режима. Это стимулирует интерес к data-driven методам, в первую очередь к обучению с подкреплением (RL), которое способно синтезировать нелинейную политику управления напрямую по данным взаимодействия со средой.

Ключевое препятствие для промышленного внедрения RL – низкая эффективность по данным и отсутствие гарантий безопасности при обучении «в реальном времени» на действующей установке. Естественным решением является перенос основной массы обучения на **цифровой двойник** – высокоточную модель объекта, синхронизируемую с реальным процессом по измеряемым данным, – с последующим ограниченным дообучением на реальном объекте.

(cc) BY

Систематизация применения RL в промышленном управлении процессами дана в обзоре Nian, Liu и Huang [1], где показано смещение от табличных методов к методам глубокого RL для непрерывных процессов. Значимым продвижением в направлении учёта ограничений стали методы штрафной функции с адаптивным весом для обработки ограничений на состояние в глубоком RL [2, 3], а также вероятностно-ограниченная оптимизация политики, позволяющая формулировать требования по спецификации продукта в вероятностной, а не жёсткой форме.

Отдельное направление связано с попытками объединить преимущества классической теории управления и RL. В частности, предложена структура control-informed reinforcement learning (CIRL), в которой ПИД-компоненты встраиваются непосредственно в архитектуру политики глубокого RL, что позволяет сочетать способность ПИД подавлять немоделируемые возмущения по измеренной ошибке с нелинейной моделирующей способностью RL [4].

Важным свидетельством практической применимости RL к реальным ректификационным колоннам является промышленный пилотный проект Yokogawa и JSR, в котором алгоритм Factorial Kernel Dynamic Policy Programming (FKDPP) использовался для управления декантером и ректификационной колонной при производстве винилацетат-мономера в течение продолжительного периода эксплуатации [13]. Отдельно стоит отметить применение RL не только к оперативному управлению, но и к задачам проектирования: показано использование алгоритма soft actor–critic для автоматизации выбора конструктивных параметров тарельчатой ректификационной колонны в среде AspenPlus®/RadFrac [5].

Обзор по искусственному интеллекту в цифровых двойниках [6] систематизирует методы машинного обучения, используемые на разных этапах жизненного цикла двойника – от построения суррогатной модели до принятия решений [7]. Значительное развитие получило направление робастного/учитывающего неопределённость предиктивного управления на основе цифрового двойника. В работе [8] предложен фреймворк робастного MPC с многошаговым прогнозом на основе модели Time-Series Dense Encoder (TiDE) и квантильной регрессии для количественной оценки неопределённости предсказаний двойника, формулирующий задачу робастного MPC.

Отдельно развивается направление физически информированных нейросетевых двойников непосредственно для динамики ректификационных колонн. В работе [9] предложена **physics-informed neural network (PINN)** модель для динамического, потарелочного моделирования бинарной колонны в переходных режимах, где ограничения фазового равновесия, материального и теплового баланса встраиваются непосредственно в функцию потерь при обучении.

В работе [10] показано ускорение сходимости политики при обучении на двойнике по сравнению с обучением непосредственно на физическом объекте. Систематический обзор этого направления с точки зрения RL выполнен в [11, 12], где digital twin рассматривается как средство одновременного решения задач эффективности данных, безопасности исследования и переноса модели.

Математический аппарат

Общая архитектура предложенного фреймворка – взаимодействие реального объекта, цифрового двойника и контура обучения с подкреплением с рандомизацией домена – показана на рисунке 1. Цифровой двойник обучает control-informed политику офлайн при рандомизации параметров φ, после чего политика передаёт управляющие воздействия на реальный объект, а измерения объекта используются для непрерывной актуализации двойника.

(cc) BY

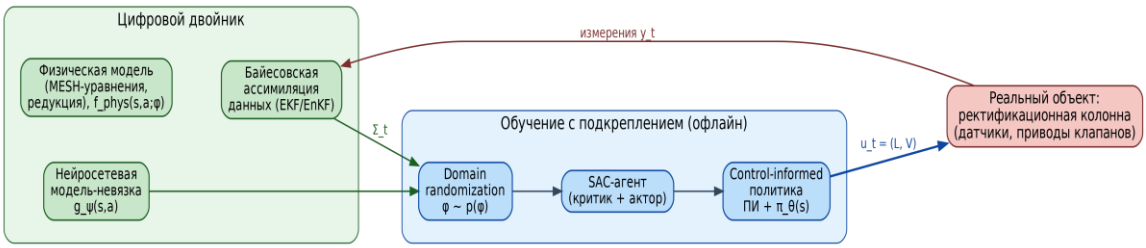


Рисунок 1. Архитектура гибридного фреймворка

Рассматривается бинарная ректификационная колонна с N теоретическими тарелками, нумеруемыми снизу вверх, с питанием на тарелку N_F. Для каждой тарелки $i = 1, \dots, N$ динамика описывается системой уравнений материального баланса (M), фазового равновесия (E), суммирования долей (S) и теплового баланса (H) – так называемых **MESH-уравнений**.

Материальный баланс по компоненту:

$$\frac{d(M_i x_i)}{dt} = L_{i+1}x_{i+1} + V_{i-1}y_{i-1} - L_i x_i - V_i y_i + F_i z_i - S_i x_i, \quad (1)$$

где M_i – мольная задержка жидкости на тарелке i; L_i, V_i – мольные расходы жидкости и пара; x_i, y_i – мольные доли лёгкого компонента в жидкой и паровой фазах; F_i, z_i – расход и состав питания на тарелке i; S_i – боковой отбор.

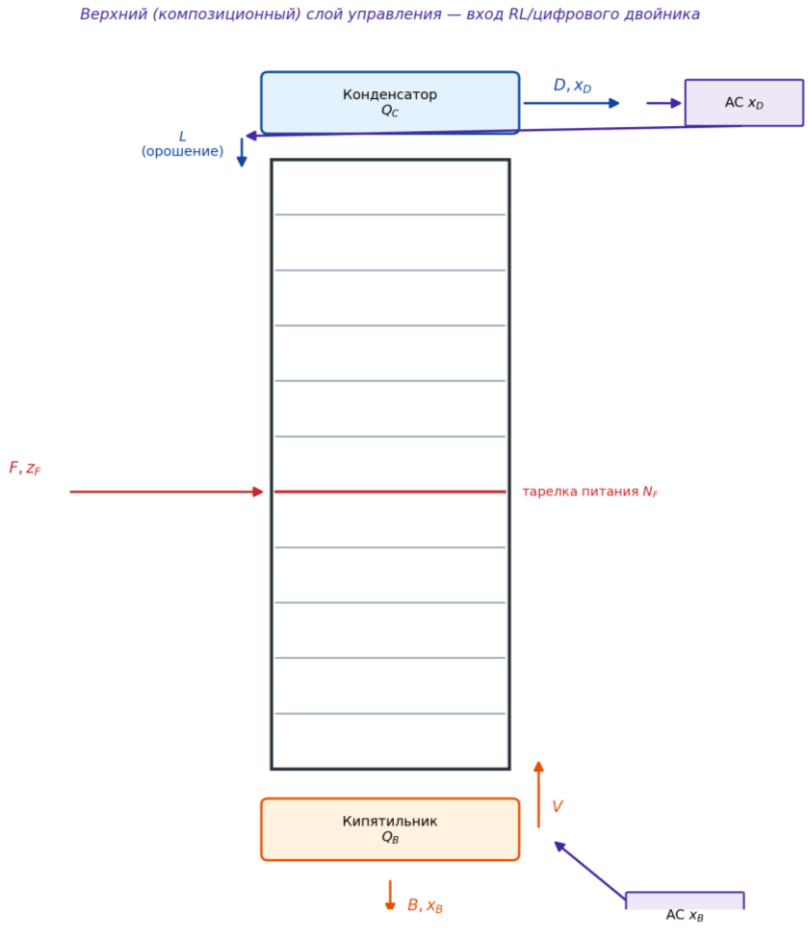


Рисунок 2. Принципиальная схема ректификационной колонны

Фазовое равновесие с коэффициентом относительной летучести α_i:

$$y_i = \frac{\alpha_i x_i}{1+(\alpha_i - 1)x_i}. \quad (2)$$

(cc) BY

Суммирование долей: $\sum x_{i,k} = 1$, $\sum y_{i,k} = 1$. Тепловой баланс:

$$\frac{d(M_i h_i^L)}{dt} = L_{i+1}h_{i+1}^L + V_{i-1}h_{i-1}^V - L_i h_i^L - V_i h_i^V + F_i h_i^F - Q_i, \quad (3)$$

где h^L, h^V, h^F – энтальпии жидкости, пара и питания, Q_i – подвод/отвод тепла. Гидравлика тарелки обычно задаётся эмпирическим соотношением Francis weir или аналогом.

Управляющими воздействиями традиционно выступают расход орошения L_0 и тепловой поток в кипятильнике Q_B; регулируемые переменные – составы дистиллята x_D и кубового продукта x_B. Уровни в кубе и флегмовой ёмкости стабилизируются вспомогательными контурами и в данной работе рассматриваются как часть базового слоя регулирования, на который опирается верхний слой управления.

Полная MESH-модель представляет собой систему из O(N) жёстких дифференциально-алгебраических уравнений (ДАУ) и в силу вычислительной трудоёмкости используется как «эталонная» часть цифрового двойника, тогда как для целей обучения политики в реальном времени применяется её редукция – линеаризованная либо идентифицированная модель пониженного порядка.

Задача управления колонной формализуется как дискретный по времени марковский процесс принятия решений (MDP), задаваемый кортежем:

$$\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma, \rho_0 \rangle. \quad (4)$$

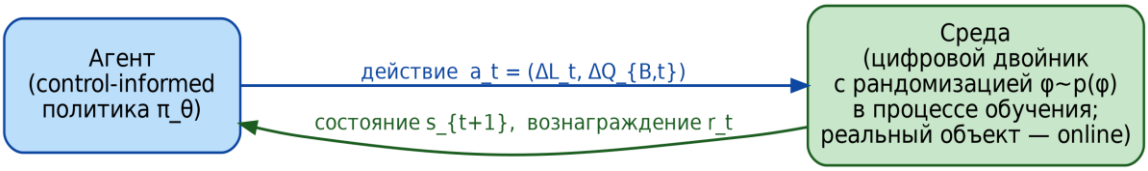


Рисунок 3. Цикл взаимодействия «агент–среда»: агент выдаёт действие a_t, среда возвращает новое состояние s_{t+1} и вознаграждение r_t .

Состояние s_t включает измеряемые/оцениваемые переменные: составы на выбранных тарелках, расходы, уровни, а также – что существенно для учёта запаздывания – скользящее окно предыстории измерений и управляющих воздействий:

$$s_t = [y_{t-\tau_{max}:t}, u_{t-\tau_{max}:t-1}, \hat{d}_t], \quad (5)$$

где d_t – оценка ненаблюдаемого возмущения, формируемая цифровым двойником.

Действие – приращения управляющих воздействий, физически ограниченные пределами клапанов и скоростью их перемещения:

$$a_t = u_t = [\Delta L_t, \Delta Q_{B,t}]^\top \in \mathcal{A} \subset \mathbb{R}^{n_a}, u_{min} \leq u_t \leq u_{max}. \quad (6)$$

Переходная динамика $P(s_{t+1}|s_t, a_t)$ задаётся дискретизированной моделью двойника с добавлением стохастического возмущения, отражающего измерительный шум и немоделируемую динамику.

Функция вознаграждения формулируется как отрицательное значение многокритериального штрафа, типичного для задач слежения:

$$r(s_t, a_t) = -\begin{bmatrix} w_1 \left\| x_D^{sp} - x_D(t) \right\|^2 + w_2 \left\| x_B^{sp} - x_B(t) \right\|^2 \\ + w_3 \|\Delta u_t\|^2 + \\ + w_4 \mathbb{I}\{u_t \notin [u_{min}, u_{max}]\} c_{viol} \end{bmatrix} \quad (7)$$

где первые два слагаемых штрафуют отклонение составов от задания, третье – «резкость» управления, четвёртое – мягкий штраф за нарушение ограничений [2]. Задача обучения состоит в поиске политики $\pi_\theta(a_t, s_t)$, максимизирующей ожидаемую дисконтированную сумму вознаграждений:

$$\theta^* = arg\max_{\theta} \mathbb{E}_{\tau\sim\pi_\theta,P}[\sum_{t=0}^{T}\gamma^t r(s_t,a_t)] \quad (8)$$

Поскольку истинная динамика объекта P неизвестна и отличается от модели двойника P_φ, обучение только на номинальной модели приводит к политике, переобученной под конкретную реализацию модели и хрупкой при реальном рассогласовании «модель–объект». Для устранения этого используется формализация робастного MDP* [14]: ϕϕ

$$\theta^*_{robust} = arg\max_{\theta}\min_{\phi\in\Phi} \mathbb{E}_{\tau\sim\pi_\theta,P_\phi}[\sum_t \gamma^t r(s_t,a_t)], \quad (9)$$

где Φ – множество неопределённости параметров двойника. Максиминная задача решается приближённо через рандомизацию домена: на каждом эпизоде обучения параметр φ сэмплируется из распределения p(φ) на Φ, и политика оптимизируется по ожиданию, а не по наихудшему случаю, что даёт достаточно робастные политики при значительно меньшей вычислительной стоимости:

$$\theta^*_{robust} \approx arg\max_{\theta} \mathbb{E}_{\phi\sim p(\phi)}\mathbb{E}_{\tau\sim\pi_\theta,P_\phi}[\sum_t \gamma^t r(s_t,a_t)]. \quad (10)$$

Для дополнительной устойчивости к действиям используется идея action-robust RL [15]: с вероятностью α на исполнение подставляется наихудшее возмущённое действие

$$\bar{a}_t = a_t + \delta_t, \quad (11)$$

где δ_t выбирается второй, «противоборствующей» политикой, обучаемой одновременно с основной по принципу минимакс-игры. Это моделирует, в частности, эффекты насыщения и запаздывания исполнительных механизмов сверх номинальных допущений.

Двойник строится как аддитивная комбинация физической модели пониженного порядка $f_{phys}(\cdot;\varphi)$ и нейросетевой модели-невязки g_ψ(·), компенсирующей систематическое расхождение физической модели с данными объекта:

$$\hat{s}_{t+1} = f_{phys}(s_t,a_t,\phi) + g_\psi(s_t,a_t). \quad (12)$$

Параметры физической части φ и весов невязки ψ актуализируются в реальном времени по мере поступления новых измерений y_t схемой байесовской ассимиляции данных – в простейшем случае расширенным или ансамблевым фильтром Калмана (EKF/EnKF):

$$\phi_{t+1} = \phi_t + K_t(y_t - h(\hat{s}_t;\,\phi_t)),\ K_t = \sum_t H_t^\top (H_t \sum_t H_t^\top + R_v)^{-1}, \quad (13)$$

где H_t – якобиан оператора наблюдения, R_v – ковариация измерительного шума, Σ_t – ковариация оценки состояния/параметров, обновляемая на каждом шаге ассимиляции. Апостериорная ковариация Σ_t одновременно служит количественной оценкой текущей неопределённости двойника, которая может использоваться для сужения/смещения множества Φ в задаче робастного RL по мере накопления данных об объекте и для формирования доверительных интервалов прогноза, аналогично квантильной оценке неопределённости, используемой в робастном MPC на основе двойника [8].

По аналогии с CIRL [4], итоговое управляющее воздействие формируется как сумма двух компонент – классической интегрально-пропорциональной обратной связи и нейросетевой поправки π_θ, использующей полное состояние для компенсации многосвязности, запаздывания и известных возмущений:

$$u_t = (K_p e_t + K_i \sum_{k\le t} e_k \Delta t)_{PI} + (\pi_\theta(s_t))_{RL,\ twin-informed}, \quad (14)$$

где

$$e_t = y^{sp} - y_t. \quad (15)$$

Такая параметризация обеспечивает устойчивость к немоделируемым возмущениям через измеренную ошибку, упреждающую компенсацию известных/оцениваемых возмущений и взаимосвязи каналов через π_θ, обучаемую на данных двойника. Частным случаем данной структуры, используемым для аналитически прозрачного вычислительного эксперимента, является линейная feed-forward компенсация возмущения по составу питания на основе

(cc) BY

модели двойника, которая может рассматриваться как редуцированная аппроксимация обучаемого компонента π_θ вблизи рабочей точки – по аналогии с известной связью между линейно-квадратичным регулированием (LQR) и решением, к которому сходится RL-политика в линейно-квадратичной постановке задачи управления [16].

Полная (нелинейная) RL-компонента π_θ обучается алгоритмом типа soft actor–critic (SAC) – выбор обусловлен устойчивостью к переобучению за счёт энтропийной регуляризации и хорошей выборочной эффективностью для непрерывных действий, что подтверждено, в частности, применением SAC к смежным задачам оптимизации конструктивных параметров колонн [5]. Функция потерь критика:

$$\mathcal{L}_Q(\omega) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{B}}\Big[\big(Q_\omega(s,a) - \big(r + \gamma\Big(\min_{i=1,2} Q_{\bar{\omega}_i}(s',a') - \beta\, log\, \pi_\theta(a'|s')\Big)\big)\big)^2\Big] \quad (16)$$

функция потерь актора:

$$\mathcal{L}_\pi(\theta) = \mathbb{E}_{s\sim\mathcal{B},\ a\sim\pi_\theta}\Big[\beta\, log\, \pi_\theta(a|s) - \min_{i=1,2} Q_{\omega_i}(s,a)\Big], \quad (17)$$

где $\mathcal{B}$ – буфер воспроизведения, наполняемый эпизодами взаимодействия с рандомизированным по $\varphi\sim p(\varphi)$ двойником, β – коэффициент температуры энтропии. Двухэтапная процедура обучения:

1. **Офлайн-этап на двойнике.** Политика обучается на симуляциях двойника с рандомизацией параметров φ до сходимости по критерию средней суммарной награды на валидационном наборе реализаций φ, не пересекающемся с обучающим.

2. **Онлайн дообучение/адаптация.** После развёртывания на реальном объекте политика дообучается ограниченным числом шагов с консервативным ограничением отклонения от офлайн-политики, чтобы избежать деградации качества управления в переходный период, а двойник продолжает актуализироваться схемой ассимиляции данных, постепенно сужая используемое при последующих дообучениях множество неопределённости Φ.

Методика вычислительного эксперимента

Для количественной проверки центрального тезиса работы – что информированность регулятора моделью цифрового двойника даёт измеримый выигрыш относительно чисто реактивной обратной связи при наличии параметрической неопределённости – использован общепризнанный в литературе по многосвязному управлению бенчмарк Wood–Berry [17]: линеаризованная модель бинарной ректификационной колонны для разделения смеси метанол–вода, заданная передаточными функциями с запаздыванием (FOPDT) для двух выходов (состав дистиллята x_D, состав куба x_B) и двух входов (расход орошения L, расход пара в кипятильнике V), а также каналом возмущения по составу питания z_F. Данная модель – стандартный тестовый полигон в работах по многосвязному и робастному управлению [18] в силу сильного взаимодействия каналов и заметного запаздывания.

Полная модель объекта, использованная как «истинный», неизвестный регулятору объект в вычислительном эксперименте определяется следующим выражением:

$$x_D(s) = \frac{K_{11}e^{-\theta_{11}s}}{\tau_{11}s+1}L(s) + \frac{K_{12}e^{-\theta_{12}s}}{\tau_{12}s+1}V(s) + \frac{K_{d1}e^{-\theta_{d1}s}}{\tau_{d1}s+1}z_F(s), \quad (18a)$$

$$x_B(s) = \frac{K_{21}e^{-\theta_{21}s}}{\tau_{21}s+1}L(s) + \frac{K_{22}e^{-\theta_{22}s}}{\tau_{22}s+1}V(s) + \frac{K_{d2}e^{-\theta_{d2}s}}{\tau_{d2}s+1}z_F(s) \quad (18б)$$

с номинальными параметрами K_{11}=12.8, τ_{11}=16.7, θ_{11}=1.0; K_{12}=−18.9, τ_{12}=21.0, θ_{12}=3.0; K_{21}=6.6, τ_{21}=10.9, θ_{21}=7.0; K_{22}=−19.4, τ_{22}=14.4, θ_{22}=3.0 и параметрами

(cc) BY

канала возмущения K_{d1}=3.8, τ_{d1}=14.9, θ_{d1}=8.1; K_{d2}=4.9, τ_{d2}=13.2, θ_{d2}=3.4.

Сопоставление проведено между двумя схемами управления, использующими одну и ту же базовую децентрализованную ПИ-обратную связь – с целью выделить в чистом виде эффект, вносимый именно информацией цифрового двойника:

ПИД (базовый вариант). Чисто реактивная децентрализованная обратная связь без какой-либо информации о возмущении или модели двойника; воздействие формируется исключительно по измеренной ошибке составов.

Гибридная схема «двойник + прямая связь» (DT+FF). К той же ПИ-обратной связи добавляется компонента прямой связи:

$$u_{ff}(t) = -\hat{G}_{ss}^{-1}\hat{G}_{d,ss}\left[z_F * h_{lag}\right](t) \quad (19)$$

рассчитываемая цифровым двойником: измеряемое изменение состава питания z_F пропускается через фильтр первого порядка h_{lag} с постоянной времени, оцененной двойником, и статический декуплирующий коэффициент, формируя упреждающую компенсацию до появления существенного отклонения состава. Двойник в этом эксперименте намеренно откалиброван неидеально: его оценки параметров канала возмущения содержат случайное рассогласование до ±15 % относительно истинных значений объекта – что моделирует реалистичный сценарий несовершенной идентификации.

Такая редуцированная реализация используется как интерпретируемое приближение к полной нелинейной RL-политике π_θ в силу известной связи между линейно-квадратичной постановкой RL-задачи и её сходимостью к структуре линейной обратной связи с прямой компенсацией измеряемых возмущений [16].

Робастность оценивается методом Монте-Карло для каждого из n=60 независимых прогонов параметры «истинного» объекта φ сэмплируются равномерно в диапазоне ±25 % от номинала, а запаздывания – в диапазоне ±35 %, что шире, чем диапазон, условно принятый за «известный двойнику» при калибровке ±15 % для канала возмущения, таким образом проверяется устойчивость схемы не только к параметрам, использованным при синтезе, но и к частичной экстраполяции за пределы предполагаемого множества неопределённости. В момент t=20 мин моделируется ступенчатое возмущение состава питания. Горизонт моделирования – 80 мин, шаг интегрирования – 0,05 мин (метод Эйлера с явным буфером запаздывания). Управляющие воздействия ограничены физическим пределом $|u| \leq 6$ условные единицы хода клапана.

Использованы стандартные для сравнения регуляторов интегральные показатели:

$$ISE = \int_0^T (x_D^2(t) + x_B^2(t))dt, \quad (20)$$

$$IAE = \int_0^T (|x_D(t)| + |x_B(t)|)dt \quad (21)$$

время установления – момент последнего выхода отклика x_D за 5 %-ю полосу после нанесения возмущения; энергия управления –

$$E_u = \int_0^T (L^2(t) + V^2(t))dt \quad (22)$$

как косвенный индикатор эксплуатационных затрат/износа исполнительных механизмов; число насыщений клапанов за прогон, как индикатор нарушения физических ограничений.

Результаты вычислительного эксперимента

Результаты усреднения по 60 реализациям Монте-Карло приведены в таблице 1.

Таблица 1.

Сравнение показателей качества управления: базовый децентрализованный ПИД и гибридная схема «цифровой двойник + прямая связь»

Показатель	ПИД (базовый)	Гибрид DT+FF	Изменение, %
ISE	33,35 ± 19,87	16,44 ± 14,75	−50,7
IAE	38,05 ± 14,78	27,48 ± 12,26	−27,8
Время установления, мин	58,25 ± 5,29	55,80 ± 7,74	−4,2
Энергия управления (усл. ед.)	6,20 ± 6,34	9,91 ± 11,03	+59,8
Число насыщений клапанов	0	0	0

Для наглядной иллюстрации механизма улучшения на рисунке 4 показан отклик составов дистиллята и куба на ступенчатое возмущение состава питания для одной репрезентативной реализации параметрической неопределённости, здесь величина улучшения ISE для неё близка к среднему значению по всей выборке Монте-Карло 50,7 %.

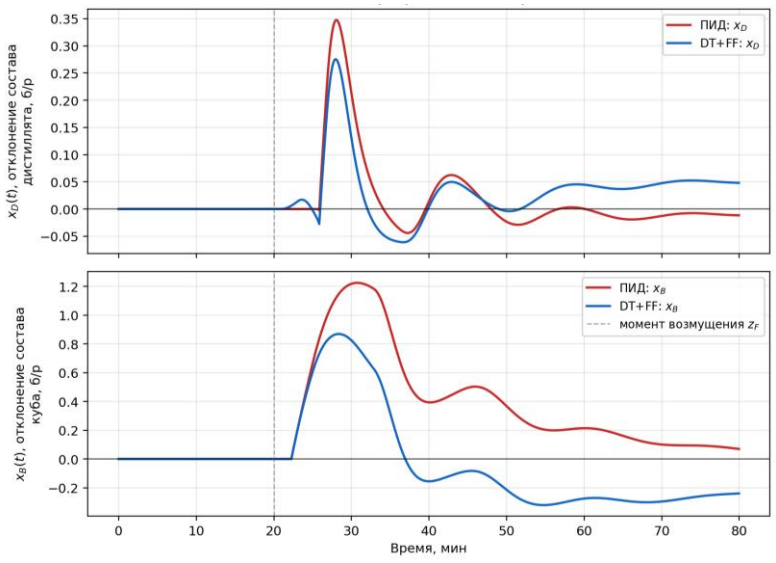


Рисунок 4. Временные отклики составов дистиллята $x_D(t)$ и куба $x_B(t)$.

Итоговое сопоставление интегральных критериев ошибки по всей выборке Монте-Карло (n=60) с указанием разброса представлено на рисунке 5.

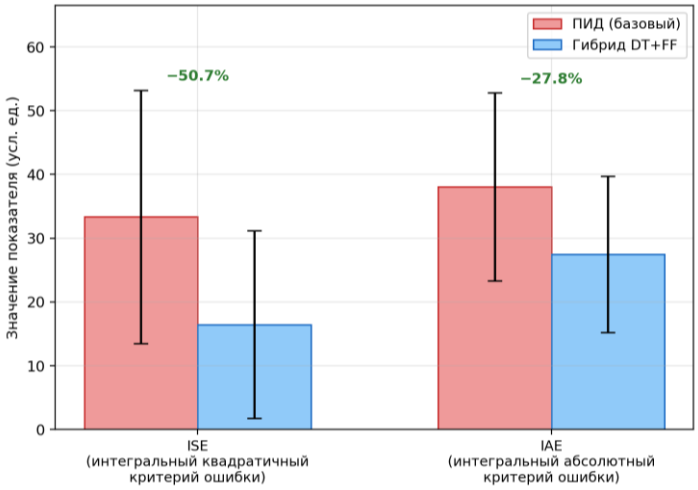


Рисунок 5. Сравнение интегральных показателей качества управления.

(cc) BY

Представленный вычислительный эксперимент сознательно использует линейную, интерпретируемую реализацию гибридной схемы вместо полного обучения нелинейной SAC-политики на нелинейной MESH-модели с рандомизацией домена. Это ограничение продиктовано требованиями воспроизводимости и прозрачности количественного анализа в рамках одной статьи; тем не менее, выбор именно такой редукции методологически обоснован известной эквивалентностью между RL-политикой, обучаемой на линейно-квадратичной формулировке задачи, и решением классической задачи линейно-квадратичного регулирования с интегральной/прямой связью [16] – то есть результаты можно интерпретировать как оценку выигрыша, достижимого уже на линейном участке сходимости полной нелинейной политики; ожидаемо, что для существенно нелинейных режимов выигрыш нелинейной RL-политики от использования цифрового двойника будет не меньшим, а вероятно, большим, поскольку линейные регуляторы в таких режимах теряют адекватность модели быстрее, чем нелинейная политика, обученная с учётом рандомизации в широком диапазоне режимов. Проверка этой гипотезы на полной нелинейной модели с обучением SAC – приоритетное направление последующей работы.

Ключевой практический риск гибридных DT+RL схем – разрыв между двойником и реальным объектом, способный привести к деградации или даже дестабилизации управления при переносе политики на физическую установку. В работе учтены три общепринятых механизма снижения этого риска рандомизация домена при офлайн-обучении, сохранение интегральной обратной связи как «страховочного» компонента, гарантирующего базовую устойчивость независимо от точности модели и консервативное онлайн дообучение вместо полного переобучения на реальном объекте. Тем не менее в работе не выполнена аппаратная или высокоточная нелинейная валидация – это ограничение, типичное для значительной части литературы по RL в управлении процессами и явно признаваемое, например, в работах по CIRL [4] и по промышленному пилоту FKDPP [13], где именно длительность и осторожность поэтапного внедрения стали определяющим фактором успеха.

Заключение

В работе предложен гибридный фреймворк, объединяющий цифровой двойник ректификационной колонны с политикой обучения с подкреплением, формализованной как решение робастного/распределительно-робастного марковского процесса принятия решений при рандомизации домена. Предложена control-informed структура политики, сочетающая классическую интегральную обратную связь с информированной двойником поправкой, что обеспечивает базовую устойчивость независимо от точности модели и одновременно позволяет использовать прогностическую информацию двойника для упреждающей компенсации возмущений и взаимосвязи каналов.

Вычислительный эксперимент на признанном в литературе бенчмарке Wood–Berry с методом Монте-Карло по параметрической неопределённости продемонстрировал статистически значимое в среднем на 50,7 % по ISE и 27,8 % по IAE улучшение качества регулирования гибридной схемы относительно децентрализованного ПИД при сопоставимой настройке базового контура обратной связи, ценой ожидаемого роста энергозатрат на управление. Обсуждены ограничения представленного эксперимента относительно полной нелинейной архитектуры фреймворка, вопросы sim-to-real переноса, безопасности онлайн дообучения и интерпретируемости решений для операторского персонала, а также сформулированы конкретные направления дальнейших исследований, включая полное обучение нелинейной политики на физически информированной модели двойника и переход к распределительно-робастной формулировке с адаптивным сужением множества неопределённости.

(cc) BY

Литература

1. Nian R., Liu J., Huang B. A review on reinforcement learning: Introduction and applications in industrial process control // Computers & Chemical Engineering. – 2020. – Vol. 139. – Art. 106886.
2. Yoo H., Zavala V. M., Lee J. H. A dynamic penalty approach to state constraint handling in deep reinforcement learning // Journal of Process Control. – 2022. – Vol. 115. – P. 157–166.
3. Petsagkourakis P., Sandoval I. O., Bradford E., Galvanin F., Zhang D., del Rio-Chanona E. A. Chance constrained policy optimization for process control and optimization // Journal of Process Control. – 2022. – Vol. 111. – P. 35–45.
4. Bloor M., Ahmed A., Kotecha N., Mercangöz M., Tsay C., del Rio-Chanona E. A. Control-Informed Reinforcement Learning for Chemical Processes // Industrial & Engineering Chemistry Research. – 2025 (препринт: arXiv:2408.13566, 2024).
5. Soft Actor-Critic Reinforcement Learning Improves Distillation Column Internals Design Optimization // Processes / MDPI. – 2025. – Vol. 9(2). – Art. 34.
6. Kreuzer T., Papapetrou P., Zdravkovic J. Artificial Intelligence in Digital Twins – A Systematic Literature Review // Data & Knowledge Engineering. – 2024. – Vol. 151. – Art. 102304.
7. Mane et al. Digital twin in the chemical industry: A review // Digital Twins and Applications (IET/Wiley). – 2024.
8. Chen Y.-P., Tsai Y.-K., Karkaria V., Chen W. Uncertainty-Aware Digital Twins: Robust Model Predictive Control using Time-Series Deep Quantile Learning // arXiv:2501.10337. — 2025.
9. Physics-Informed Neural Network Digital Twin for Dynamic Tray-Wise Modeling of Distillation Columns under Transient Operating Conditions // arXiv:2603.24644. – 2026.
10. Cronrath C., Aderiani A. R., Lennartson B. Enhancing digital twins through reinforcement learning // Proc. IEEE 15th Int. Conf. on Automation Science and Engineering (CASE). – 2019. – P. 293–298.
11. Mohammed S. K., Singh S., Mizouni R., Otrok H., Damiani E. Digital twins paradigm: A systematic review from the reinforcement learning perspective // ACM Computing Surveys. – 2026. – Vol. 58(7). – P. 1–33.
12. Kuo H.-A., Hong T.-Y., Chien C.-F. A deep reinforcement learning based digital twin framework for resilient production planning under demand uncertainty and an empirical study in semiconductor wafer fabrication // Computers & Industrial Engineering. – 2025. – Art. 111389.
13. How to support decision making with reinforcement learning in hierarchical chemical process control? // ScienceDirect (открытый доступ). – 2025.
14. Nilim A., El Ghaoui L. Robust control of Markov decision processes with uncertain transition matrices // Operations Research. – 2005. – Vol. 53, No. 5. – P. 780–798.
15. Pinto L., Davidson J., Sukthankar R., Gupta A. Robust adversarial reinforcement learning // Proceedings of the 34th International Conference on Machine Learning (ICML). – 2017. – Vol. 70. – P. 2817–2826.
16. Recht B. A Tour of Reinforcement Learning: The View from Continuous Control // Annual Review of Control, Robotics, and Autonomous Systems. – 2019. – Vol. 2. – P. 253–279.
17. Wood R. K., Berry M. W. Terminal composition control of a binary distillation column // Chemical Engineering Science. – 1973. – Vol. 28, No. 9. – P. 1707–1717.
18. Skogestad S., Postlethwaite I. Multivariable Feedback Control: Analysis and Design. – 2nd ed. – Chichester: John Wiley & Sons, 2005.

(cc) BY