https://doi.org/10.31891/2307-5732-2025-351-49
УДК 004.8

ПЕТРОВСЬКИЙ ОЛЕКСАНДР
Національний університет «Львівська політехніка»
https://orcid.org/0000-0002-5729-544X
e-mail: oleksandr.s.petrovskyi@lpnu.ua

ПРЕТРЕНУВАННЯ НА ПРИБЛИЗНИХ СИМУЛЯЦІЯХ ЯК МЕТОД ЗМЕНШЕННЯ ПОТРЕБ У ДОСЛІДЖЕННІ SOFT ACTOR-CRITIC ПРИ КЕРУВАННІ БІОРЕАКТОРОМ

Через обмеження домінуючих методів реалізації систем автономного контролю біореакторів, заснованих на пропорціонально-інтегрально-диференціальних законах, нечіткій логіці та прогнозуючих моделях, стабільно зростає теоретичний і практичний інтерес до створення розумних контролерів на базі навчання з підкріпленням, що здатні самостійно вивчати оперовану систему без точної моделі біопроцесу та у реальному часі підлаштовуватися до різноманітних змін середовища. Проте на практиці впровадження таких контролерів часто супроводжується багатьма складнощами, зокрема високою ціною та тривалістю дослідження середовища. У статті запропоновано метод претренування, який полягає у навчанні агента протягом багатьох короткотривалих епох переводити приблизні симуляції з випадкового стану до бажаного і дозволяє таким чином отримати хорошу початкову стратегію управління, при використанні якої на реальному біореакторі суттєво зменшуються потреби у дорогому дослідженні середовища та ймовірність доведення системи до незворотних критичних станів. На прикладі симуляції біореактору для вирощування дріжджів продемонстровано, що за такого підходу алгоритм Soft Actor-Critic швидше збігається до оптимальної стратегії та при достатньому претренуванні уникає дослідження потенційно небезпечних станів системи навіть попри те, що жодна з використаних під час претренування симуляцій точно не відображала цільовий біопроцес. Впровадження претренованого агенту у порівнянні зі звичайним дозволило зменшити MSE майже у 50 разів, а ITSE – майже у 160. Використання цього методу спрощує впровадження контролерів на базі навчання з підкріпленням для управління біопроцесами, зменшуючи економічну вартість, складність та трудомісткість реалізації систем автономного керування індустріальними біореакторами, що, у свою чергу, підвищує обсяги виготовлення, якість та доступність багатьох корисних продуктів.

***Ключові слова*:** *адаптивний контроль, контроль біопроцесів, біореактор, навчання з підкріпленням, Soft Actor-Critic, offline-to-online навчання.*

PETROVSKYI OLEKSANDR
Lviv Polytechnic National University

APPROXIMATE SIMULATION PRETRAINING AS A WAY TO REDUCE SOFT ACTOR-CRITIC EXPLORATION NEEDS IN TERMS OF BIOREACTOR CONTROL

Bioreactors are at the heart of most biotechnological processes as they make it possible to grow stem cells and artificial organs, process waste and pollutants, produce biofuels, pharmaceuticals, vaccines, popular food and beverages, and many more. Due to the limitations of the dominant methods for implementing autonomous bioreactor control systems based on proportional-integral-differential laws, fuzzy logic, and predictive models, there is a steadily growing theoretical and practical interest in creating smart reinforcement-learning-based controllers that are capable of learning the operated system by themselves and adapting to various changes in real-time without the need in an accurate bioprocess model. However, implementing RL-based controllers in practice is often accompanied by many challenges, with the high cost and duration of environment exploration being among the most significant. In this article, an approximate simulation pretraining method is proposed that will allow obtaining a flexible initial control policy, the use of which in a real bioreactor will significantly reduce the need for expensive environment exploration and the probability of the controller accidentally bringing the system to an irreversible critical state. Using a baker yeast bioreactor simulation as an example, it is demonstrated that with this approach, the RL-agent converges to an optimal policy significantly faster and, with a sufficient amount of pretraining, avoids exploring potentially dangerous states of the system even though none of the approximate simulations used for pertaining accurately reflected the target system's dynamics. The agent adapted to the "real" environment 9 times faster, reducing the MSE by a factor of 50 and the ITSE by 160. This method simplifies the implementation of reinforcement learning-based controllers for many bioprocesses, reducing the economic cost, complexity, and laboriousness of the development of autonomous control systems for industrial bioreactors, which in turn will allow for increased production volume, quality, and availability of many valuable products.

Keywords: adaptive control, bioprocess control, bioreactor, reinforcement learning, Soft Actor-Critic, offline-to-online learning.

Стаття надійшла до редакції / Received 04.04.2025
Прийнята до друку / Accepted 18.04.2025

Постановка проблеми

Функціонуючий біореактор – це надскладна стохастична динамічна система з високим ступенем нелінійності. Існує безліч факторів, що можуть впливати на перебіг біопроцесу як на макроскопічному (поглинання субстрату, насичення киснем, накопичення інгібіторів росту тощо), так і на мікроскопічному рівні окремих клітин [1, 2]; наявні суттєві обмеження у можливості видобутку інформації про реальний стан системи [1]; кожен біопроцес вимагає урахування особливих для нього параметрів, а кожен біореактор відрізняється від інших, оскільки спеціально конструюється для вирішення конкретних задач [3]. Окрім цього, нам ніколи не відома точна модель біопроцесу, що

здійснюється, а знайти приблизну модель, яка узгоджується з обмеженими експериментальними даними, може бути доволі важко [4]. Саме тому такі сучасні підходи до побудови систем керування як контролери на базі пропорціонально-інтегрально-диференціальних законів, нечіткої логіки та управління з прогнозуючими моделями попри розповсюдженість через свої обмеження не здатні забезпечити оптимальний контроль біопроцесів або вимагають для цього точну математичну модель, побудувати яку у більшості випадків вкрай важко [5].

Цим у контексті керування біореакторів зумовлений науковий інтерес до навчання з підкріпленням (англ. reinforcement learning, RL) [6–8], завдяки якому проблема оптимального контролю може розглядатися як Марковський процес вирішування (англ. Markov decision process), у якому агент намагається самостійно навчитись максимізовувати загальну винагороду, що отримує під час взаємодії зі складним, невизначеним середовищем або вибудовуючи власну його апроксимацію, або спираючись виключно на спостережувані стани і отримуваний зворотній зв'язок [9]. Широкий досвід застосування навчання з підкріпленням у різних виробничих середовищах показав, що RL-контролери здатні перевершити традиційні методи у точності і швидкості контролю, мають видатну здатність до генералізації та навчання без попереднього знання та моделей, побудованих експертами [10].

Проте однією з вагомих завад на шляху до повсюдного застосування RL для керування біопроцесами є потреба у початковому дослідженні середовища, під час якого протягом тривалого часу контролер може приймати неоптимальні та часто небезпечні рішення [11], на відміну від, наприклад, управління з прогнозуючими моделями, яке одразу здійснює контроль, близький до оптимального, але не здатне адаптуватися до мінливих умов та цілком залежить від якості моделі, побудованої експертами. Тому розробка гібридних методів, що дозволяють суттєво скоротити тривалість дорогого дослідження середовища, при цьому зберігаючи можливість контролеру після впровадження й надалі самостійно покращувати стратегію управління попри впливи випадкових шумів, збурень, нестаціонарність біопроцесів та неточності ідентифікованих моделей, постають надзвичайно актуальними [6].

Аналіз останніх досліджень і публікацій

Попри наявність потужних алгоритмів, вартість онлайн-тренування переважно занадто висока. Ситуація ускладнюється ще й тим, що комплексність самих біопроцесів робить важким створення їх якісних механістичних моделей та симуляцій. Тому, гібридні підходи, що можуть забезпечити компроміс, увібравши найкраще від механістичних та data-driven рішень, викликають особливий інтерес, попри те, що наразі не є зрозумілим те, як найкраще інтегрувати ці два компоненти, враховуючи при цьому розбіжності між моделями і реальним середовищем, що завжди супроводжують біопроцеси [6].

Одним з багатообічяючих способів вирішення цієї задачі є концепція offline-to-online reinforcement learning [12, 13], яка поєднює стадію офлайн-претренування на наявних даних чи симуляції з онлайн-адаптацією у реальній системі. І хоча претренування повсюдно використовується у інших галузях ML, offline-to-online в рамках навчання з підкріпленням – це доволі молодий напрямок з багатьма унікальними для нього проблемами [14]. Попри це, переднавчання RL вже зазнало успіху в таких сферах як робототехніка, рекомендаційні системи, та охорона здоров'я [15].

У контексті цієї парадигми цікавий підхід пропонується у статті [16]. Автори поєднують два найпоширеніших методи контролю біореакторів на основі нейронних мереж: обернені мережі, коли на доступних даних про динаміку системи застосовується навчання з учителем для створення моделі, яка здатна відображати поточний стан у дію, що призведе до бажаного результату, і навчання з підкріпленням, у якому агент самостійно визначає оптимальну поведінку у реальному часі. Проте, на відміну від попередніх робіт схожої тематики, обернена мережа використовується не як policy-функція напряму, а як засіб ініціалізації Q-таблиці. Такий гібридний підхід дозволяє одночасно знизити вимоги до обсягу наявних даних (недолік обернених мереж) та пришвидшити сходження алгоритму навчання з підкріпленням. Автори продемонстрували ефективність описаного підходу на реальній лабораторній установці. Проте, недоліками запропонованого алгоритму є вимога у використанні табличного Q-Learning та дискретизації неперервних змінних, що зумовлює його високу просторову складність і робить його важкозастосовним для вирішення складніших задач керування неперервних біопроцесів.

У статті [17] автори також пропонують застосування поетапного впровадження RL-контролера, у якому на першій стадії проста механістична апроксимація моделі цільового процесу використовується для offline-переднавчання Policy Gradient агента, на другій – деякі параметри мережі заморожуються і мережа дотреновується на даних, отриманих з реальної системи, а на третій – переднавчений RL-контролер застосовується для керування реальною системою в онлайн-режимі.

Отже, зважаючи на ефективність offline-to-online навчання у контексті задач оптимального керування, проблема розробки методу претренування на приблизних симуляціях для керування біореакторами набуває особливої актуальності.

Формулювання цілей статті

Метою роботи є розробка та апробація методу претренування на приблизних симуляціях для зменшення потреб у дослідженні середовища та пришвидшення сходження RL-агента до оптимальної

стратегії у контексті реалізації систем автономного контролю біореактора на базі навчання з підкріпленням.

Виклад основного матеріалу

Процедура претренування. Головною метою претренування є пришвидшення сходження агента до оптимальної policy та мінімізація ймовірності прийняття шкідливих рішень за рахунок надання агенту можливості попередньо вивчити в загальних рисах, як працюють системи, схожі з цільовою. У цьому контексті найвищий пріоритет надається генералізації, що також зумовлює доцільність використання декількох приблизних симуляцій зі спрощеною механікою або дещо зміненими параметрами, особливо якщо існує декілька альтернативних версій математичної моделі біопроцесу, отриманих у результаті застосування методів ідентифікації систем.

Оскільки навчання тому, як ідеально керувати однією приблизною симуляцією тривалий час за наявності розбіжностей між моделлю і реальним біопроцесом лише зашкодить агенту, претренування організується таким чином, щоб протягом кожного короткотривалого епізоду моделювалось випадкове збурення системи, а задачею агента було якнайшвидше перевести її у бажаний стан. Таким чином агент одночасно отримує більше інформації про простір станів біопроцесів та широкий досвід ефективного реагування на різкі зміни у середовищі.

Після описаного вище претренування на приблизних симуляціях, агент перед впровадженням у реальну систему може бути додатково дотренований на математичній моделі, що має найвищу схожість з цільовою системою, та/або реальних даних, щоб ще сильніше скоротити час адаптації.

Математична модель біореактору. Апробація ефективності запропонованого методу претренування у контексті задачі утримання концентрації біомаси на визначеному наперед рівні здійснювалась з використанням симуляції біореактора для ферментації дріжджів на базі математичної моделі, яку описали і використовували у своїй роботі Pandian і Noel [16]. Ця модель була обрана через те, що попри невелику кількість параметрів і керованих змінних у порівнянні з, наприклад, поширенною симуляцією IndPenSim, вона зберігає усі характерні риси складної нелінійної системи, легко піддається модифікації та базується на рівнянні Моно, що описує динаміку росту мікроорганізмів і широко застосовується у моделюванні біопроцесів.

Динаміка моделі задається наведеною нижче системою з двох пов'язаних дифиренційних рівнянь (1), перше з яких описує швидкість зміни концентрації біомаси x_1 (г/л), а друге – концентрації субстрату x_2 (г/л).

$$\begin{aligned} \frac{dx_1}{dt} &= (r - u_1 - \theta_4) \cdot x_1 \\ \frac{dx_2}{dt} &= -\frac{r x_1}{\theta_3} + u_1 \cdot (u_2 - x_2) \\ r &= \frac{\theta_1 x_2}{\theta_2 + x_2} \end{aligned} \quad (1)$$

У Табл.1 наведений опис параметрів моделі, що використовувались при апробації алгоритму. Розглядаючи задачу, як Марковський процес вирішування, стан системи складається з концентрацій біомаси та субстрату у біореакторі $s = (x_1, x_2)$, а дією виступає концентрація субстрату $a_t = u_2 \in [5,35]$, що подається у біореактор на кожному часовому кроці. Функція винагород, як і у оригінальній роботі, була визначена як від'ємна абсолютна похибка між реальним і бажаним значеннями концентрації біомаси у реакторі $R(s_t) = -|x_1^* - x_1|$.

Таблиця 1

Опис параметрів математичної моделі біореактору

Символ	Опис	Значення
x_1	Концентрація біомаси (г/л)	$x_1 \in R^+$, початкове - 7.0
x_2	Концентрація субстрату (г/л)	$x_2 \in R^+$, початкове - 0.1
u_1	Фактор розведення ($г^{-1}$)	0.1
u_2	Субстрат, що подається (г/л)	Керована змінна, $u_2 \in [5,35]$
$\theta_{1..4}$	Інші параметри моделі	$0.31, 0.18, 0.55, 0.05$
x_1^*	Цільова концентрація біомасси (г/л)	10.0

Алгоритм навчання з підкріпленням. Для реалізації контролеру був обраний популярний алгоритм Soft Actor-Critic [18], оскільки він демонструє передову продуктивність на задачі керування біореакторами [19], спорідненій задачі керування системою опалення [20] та ідеалізованому середовищі Inverted Pendulum [21].

Приблизні симуляції. За браком реальної лабораторної установки, були використані 3 симуляції: TRUE з доданим до стану гаусівським шумом $\mathcal{N}(0, \sigma_s)$, яка відображає реальний біопроцес,

та детерміністичні AUX1 і AUX2, параметри моделей яких були дещо змінені вдносно TRUE (Табл.2). Симуляції були реалізовані за допомогою *odeint* з бібліотеки *scipy*.

Таблиця 2

Параметри моделей для симуляцій TRUE, AUX1 та AUX2

Параметр	**TRUE**	**AUX1**	**AUX2**
θ_1	0.31	0.34	0.28
θ_2	0.18	0.16	0.20
θ_3	0.55	0.60	0.50
θ_4	0.05	0.03	0.07
σ_s	0.05	0	0

Експерименти та аналіз результатів. У ході претренування використовувались детерміністичні симуляції AUX1 та AUX2 з випадковою реініціалізацією $x_1, x_2 \sim \mathcal{U}(0, 35)$ на початку кожного епізоду, де $\mathcal{U}(a, b)$ – рівномірний розподіл на проміжку (a, b). Тривалість кожного епізоду дорівнювала 32 часовим крокам (розміру одного батчу). Використовувався реалізований за допомогою *PyTorch* агент SAC з автоматичним підлаштуванням ентропії, темпом навчання $\alpha = 0.001$, коефіцієнтом м'якого оновлення у $\tau = 0.1$. Було здійснене претренування протягом 100, 125, 150, 175 та 200 епізодів.

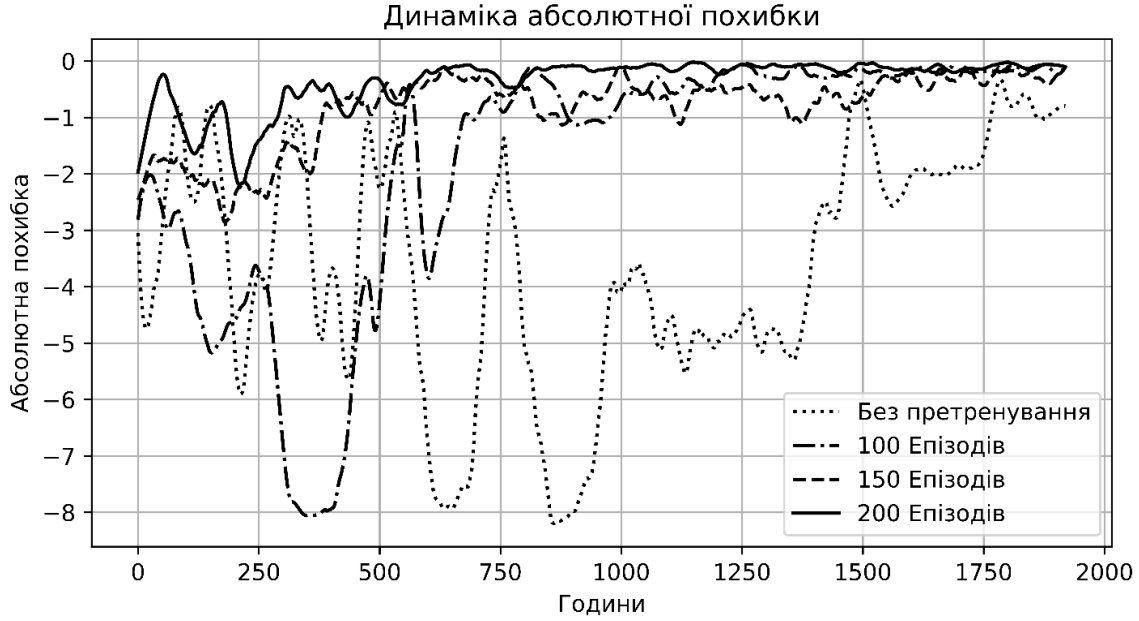

Рис.1. Динаміка абсолютної похибки на симуляції TRUE

Після цього отримані агенти, а також SAC без претренування, були застосовані для керування симуляції TRUE протягом 1920 кроків (32×60). Динаміка абсолютної похибки після згладжування фільтром Савицького–Голея з розміром вікна 64 та поліномами 3 степеню для 0, 100, 150 та 200 епізодів претренування наведена на Рис.1.

Таблиця 3 наводить залежність характеристик отриманих послідовностей абсолютних похибок від тривалості претренування на приблизних симуляціях. Використовувались такі метрики як середня абсолютна похибка (MAE), середня квадратична похибка (MSE), зважена за часом абсолютна похибка (ITAE) та зважена за часом квадратична похибка (ITSE).

Таблиця 3

Вплив претренування на якість контролю

Тривалість претрен.	0	100	125	150	175	200
MAE	3.599	1.76	1.989	0.823	0.594	0.363
MSE	17.833	8.337	9.298	1.183	0.424	0.359
ITAE	3157.215	767.207	1044.375	531.003	539.945	172.612
ITSE	14991.04	2904.138	3920.154	488.928	341.523	93.914

Обговорення результатів досліджень. Результати експериментів доводять, що за достатнього претренування на приблизних симуляціях RL-агент набагато швидше збігається до оптимальної стратегії управління цільовим середовищем та уникає дій, які можуть сильно відхилити концентрацію дріжджів від бажаного значення, що у реальних умовах може призвести до загибелі усієї мікробної культури. Претренування протягом 200 епізодів зменшило показники MSE та ITSE майже у 50 та 160 разів відповідно, що демонструє ефективність запропонованого підходу.

Наукова новизна отриманих результатів дослідження полягає у розробці і апробації методу offline-to-online навчання, що здатен суттєво зменшити потреби у дослідженні реального середовища RL-агентом завдяки претренуваню на декількох приблизних симуляціях.

Практична значущість результатів дослідження – можливість використання описаного методу для пришвидшення адаптації контроллеру на базі навчання з підкріпленням до реальних умов та мінімізації пов'язаних з цим процесом економічних ризиків.

Висновки

У статті розглянуто одну з найкритичніших проблем, що заважає широкому практичному використанню контролерів на базі навчання з підкріпленням для керування біопроцесами (висока ціна і тривалість дослідження середовища), та потенційний спосіб її вирішення в рамках парадигми offline-to-online навчання – претренування.

Запропонований підхід полягає у тому, щоб перед впровадженням RL-контролера у реальну систему на декількох приблизних симуляціях навчити агента протягом багатьох короткотривалих епізодів наближати схожі системи з випадкових станів до бажаного. Таким чином агент здатен отримати попереднє уявлення про загальні закони функціонування цільового біопроцесу та широкий досвід подолання випадкових збурень та раптових змін динаміки у схожих середовищах, у результаті чого при застосуванні контролера на реальній системі суттєво пришвидшується сходження алгоритму та зменшується ймовірність прийняття агентом потенційно небезпечних рішень. Ефективність переднавчання на приблизних симуляціях була продемонстрована на прикладі симуляції біореактору для вирощування дріжджів.

Відкритою проблемою залишається те, що попри суттєве зменшення ймовірності виконання небажаних дій, навчання SAC є стохастичним процесом, тому гарантувати, що погані рішення не будуть прийняті, неможливо, особливо у випадку складних нестаціонарних біопроцесів. Тому існує потреба у подальших дослідженнях та розробці гібридних методів, що поєднують попередні знання про систему з адаптивним контролем, завдяки яким процес впровадження розумного контролеру міг би супроводжуватися меншими економічними ризиками.

References

1. Pérez, P. A. L., López, R. A., & Femat, R. (2020). *Control in Bioprocessing: Modeling, Estimation and the Use of Soft Sensors*. John Wiley & Sons.

2. Soni, A., & Parker, R. (2004). Closed-Loop Control of Fed-Batch Bioreactors: A Shrinking-Horizon Approach. *Industrial & Engineering Chemistry Research, 43*. https://doi.org/10.1021/ie030535b

3. Palladino, F., Marcelino, P. R. F., Schlogl, A. E., José, Á. H. M., Rodrigues, R. de C. L. B., Fabrino, D. L., Santos, I. J. B., & Rosa, C. A. (2024). Bioreactors: Applications and Innovations for a Sustainable and Healthy Future – A Critical Review. *Applied Sciences*, *14*(20), Article 20. https://doi.org/10.3390/app14209346

4. Ma, Y., Zhu, W., Benton, M. G., & Romagnoli, J. (2019). Continuous control of a polymerization system with deep reinforcement learning. *Journal of Process Control*, *75*, 40–47. https://doi.org/10.1016/j.jprocont.2018.11.004

5. Bolmanis, E., Dubencovs, K., Suleiko, A., & Vanags, J. (2023). Model Predictive Control – A Stand Out among Competitors for Fed-Batch Fermentation Improvement. *Fermentation*, *9*(3), Article 3. https://doi.org/10.3390/fermentation9030206

6. Monteiro, M., & Kontoravdi, C. (2024). Bioprocess Control: A Shift in Methodology Towards Reinforcement Learning. In F. Manenti & G. V. Reklaitis (Eds.), *Computer Aided Chemical Engineering* (Vol. 53, pp. 2851–2856). Elsevier. https://doi.org/10.1016/B978-0-443-28824-1.50476-2

7. Oh, T. H. (2024). Quantitative comparison of reinforcement learning and data-driven model predictive control for chemical and biological processes. *Computers & Chemical Engineering*, *181*, 108558. https://doi.org/10.1016/j.compchemeng.2023.108558

8. Yoo, H., Byun, H. E., Han, D., & Lee, J. H. (2021). Reinforcement learning for batch process control: Review and perspectives. *Annual Reviews in Control*, *52*, 108–119. https://doi.org/10.1016/j.arcontrol.2021.10.006

9. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement Learning, second edition: An Introduction*. MIT Press.

10. Nievas, N., Pagès-Bernaus, A., Bonada, F., Echeverria, L., & Domingo, X. (2024). Reinforcement Learning for Autonomous Process Control in Industry 4.0: Advantages and Challenges. *Applied Artificial Intelligence*, *38*(1), 2383101. https://doi.org/10.1080/08839514.2024.2383101

11. Dulac-Arnold, G., Mankowitz, D., & Hester, T. (2019). *Challenges of Real-World Reinforcement Learning*, No.arXiv:1904.12901. https://doi.org/10.48550/arXiv.1904.12901

12. Ball, P. J., Smith, L., Kostrikov, I., & Levine, S. (2023). Efficient Online Reinforcement Learning with Offline Data. *Proceedings of the 40th International Conference on Machine Learning*, 1577–1594. https://proceedings.mlr.press/v202/ball23a.html

13. Zhang, H., Xu, W., & Yu, H. (2023). *Policy Expansion for Bridging Offline-to-Online Reinforcement Learning*, No.arXiv:2302.00935 https://doi.org/10.48550/arXiv.2302.00935

14. Xie, Z., Lin, Z., Li, J., Li, S., & Ye, D. (2022). *Pretraining in Deep Reinforcement Learning: A Survey*, No.arXiv:2211.03959. https://doi.org/10.48550/arXiv.2211.03959

15. Guo, S., Zou, L., Chen, H., Qu, B., Chi, H., Yu, P. S., & Chang, Y. (2024). Sample Efficient Offline-to-Online Reinforcement Learning. *IEEE Transactions on Knowledge and Data Engineering*, *36*(3), 1299–1310. https://doi.org/10.1109/TKDE.2023.3302804

16. Pandian, B. J., & Noel, M. M. (2018). Control of a bioreactor using a new partially supervised reinforcement learning algorithm. *Journal of Process Control*, *69*, 16–29. https://doi.org/10.1016/j.jprocont.2018.07.013

17. Petsagkourakis, P., Sandoval, I. O., Bradford, E., Zhang, D., & del Rio-Chanona, E. A. (2020). Reinforcement learning for batch bioprocess optimization. *Computers & Chemical Engineering*, *133*, 106649. https://doi.org/10.1016/j.compchemeng.2019.106649

18. Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., & Levine, S. (2019). *Soft Actor-Critic Algorithms and Applications*, No.arXiv:1812.05905. https://doi.org/10.48550/arXiv.1812.05905

19. Zhu, W., Castillo, I., Wang, Z., Rendall, R., Chiang, L. H., Hayot, P., & Romagnoli, J. A. (2022). Benchmark study of reinforcement learning in controlling and optimizing batch processes. *Journal of Advanced Manufacturing and Processing*, *4*(2), e10113. https://doi.org/10.1002/amp2.10113

20. Biemann, M., Scheller, F., Liu, X., & Huang, L. (2021). Experimental evaluation of model-free reinforcement learning algorithms for continuous HVAC control. *Applied Energy*, *298*, 117164. https://doi.org/10.1016/j.apenergy.2021.117164

21. Shelke, A., Vyas, D., & Srivastava, A. (2023). Comparative Study for Deep Deterministic Policy Gradient and Soft Actor Critic Using an Inverted Pendulum System. *International Conference on Electrical, Electronics, Communication and Computers (ELEXCOM)*, 1–6. https://doi.org/10.1109/ELEXCOM58812.2023.10370289