Отчет о проверке

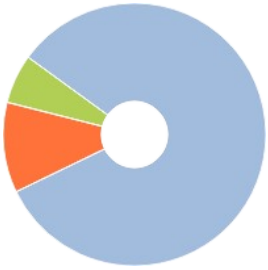
Автор: dinhngoctuan6789@gmail.com / ID: 11491219

Проверяющий:

Название документа: ВКР Динь Нгок Туан R34372

РЕЗУЛЬТАТЫ ПРОВЕРКИ

Тариф: FULL



Совпадения: 11,43%

Оригинальность: 82,1%

Цитирования: 6,47%

Самоцитирования: 0%

«Совпадения», «Цитирования», «Самоцитирования», «Оригинальность» являются отдельными показателями, отображаются в процентах и в сумме дают 100%, что соответствует проверенному тексту документа.

- **Совпадения** — фрагменты проверяемого текста, полностью или частично сходные с найденными источниками, за исключением фрагментов, которые система отнесла к цитированию или самоцитированию. Показатель «Совпадения» – это доля фрагментов проверяемого текста, отнесенных к совпадениям, в общем объеме текста.
- **Самоцитирования** — фрагменты проверяемого текста, совпадающие или почти совпадающие с фрагментом текста источника, автором или соавтором которого является автор проверяемого документа. Показатель «Самоцитирования» – это доля фрагментов текста, отнесенных к самоцитированию, в общем объеме текста.
- **Цитирования** — фрагменты проверяемого текста, которые не являются авторскими, но которые система отнесла к корректно оформленным. К цитированиям относятся также шаблонные фразы; библиография; фрагменты текста, найденные модулем поиска «СПС Гарант: нормативно-правовая документация». Показатель «Цитирования» – это доля фрагментов проверяемого текста, отнесенных к цитированию, в общем объеме текста.
- **Текстовое пересечение** — фрагмент текста проверяемого документа, совпадающий или почти совпадающий с фрагментом текста источника.
- **Источник** — документ, проиндексированный в системе и содержащийся в модуле поиска, по которому проводится проверка.
- **Оригинальный текст** — фрагменты проверяемого текста, не обнаруженные ни в одном источнике и не отмеченные ни одним из модулей поиска. Показатель «Оригинальность» – это доля фрагментов проверяемого текста, отнесенных к оригинальному тексту, в общем объеме текста.

Обращаем Ваше внимание, что система находит текстовые совпадения проверяемого документа с проиндексированными в системе источниками. При этом система является вспомогательным инструментом, определение корректности и правомерности совпадений или цитирований, а также авторства текстовых фрагментов проверяемого документа остается в компетенции проверяющего.

ИНФОРМАЦИЯ О ДОКУМЕНТЕ

Номер документа: 2

Тип документа: Не указано

Дата проверки: 12.05.2024 12:52:49

Дата корректировки: Нет

Количество страниц: 41

Символов в тексте: 53516

Слов в тексте: 6988

Число предложений: 3134

Комментарий: не указано

ПАРАМЕТРЫ ПРОВЕРКИ

Выполнена проверка с учетом редактирования: Да

Выполнено распознавание текста (OCR): Нет

Выполнена проверка с учетом структуры: Нет

Модули поиска: Переводные заимствования (KyEn), Цитирование, Коллекция НБУ, СПС ГАРАНТ: нормативно-правовая документация, СПС ГАРАНТ: аналитика, Переводные заимствования (KkEn), Патенты СССР, РФ, СНГ, Переводные заимствования*, Переводные заимствования по Интернету (KyRu), Переводные заимствования (RuEn), Переводные заимствования по коллекции Интернет в английском сегменте, ИПС Адилет, Шаблонные фразы, Публикации РГБ, Диссертации НББ, Библиография, Публикации eLIBRARY, Перефразирования по СПС ГАРАНТ: аналитика, Перефразирования по Интернету (EN), Переводные заимствования издательства Wiley, Перефразированные заимствования по коллекции Интернет в английском сегменте, Кольцо вузов, Медицина, Перефразирования по коллекции издательства Wiley, Переводные заимствования IEEE, Переводные заимствования по Интернету (EnRu), Переводные заимствования по Интернету (KkRu), Переводные заимствования по коллекции Гарант: аналитика, Перефразирования по коллекции IEEE, Публикации eLIBRARY (переводы и перефразирования), IEEE, Перефразированные заимствования по коллекции Интернет в русском сегменте, Издательство Wiley, Сводная коллекция ЭБС, СМИ России и СНГ, Перефразирования по Интернету, Публикации РГБ (переводы и перефразирования), Переводные заимствования по коллекции Интернет в русском сегменте, Интернет Плюс*

Модули, недоступные в рамках тарифа: Интернет Free

ИСТОЧНИКИ

№	Доля в тексте	Доля в отчете	Источник	Актуален на	Модуль поиска	Комментарий
[01]	5,58%	5,58%	не указано	13 Янв 2022	Библиография	
[02]	3,61%	3,56%	Reinforcement Learning and Feed... https://ieeexplore.ieee.org	22 Июн 2023	Переводные заимствования IEEE	
[03]	2,43%	0,46%	Optimal and Autonomous Control ... https://ieeexplore.ieee.org	07 Дек 2017	IEEE	
[04]	2%	0,98%	Р. С. Саттон, Э. Г. Барто ; пер. с а... http://dlib.rsl.ru	01 Дек 2014	Публикации РГБ (переводы и перефразирования)	
[05]	1,9%	1,9%	Предвзятость алгоритмов искусс... http://ivo.garant.ru	30 Окт 2021	Перефразирования по СПС ГАРАНТ: аналитика	
[06]	1,68%	0%	Books and Papers http://uta.edu	31 Янв 2017	Интернет Плюс*	
[07]	1,65%	0%	Reinforcement Learning and Feed... https://ieeexplore.ieee.org	22 Июн 2023	IEEE	
[08]	1,65%	0%	https://iris.poliba.it/retrieve/dd89f... https://iris.poliba.it	12 Мая 2024	Интернет Плюс*	
[09]	1,63%	0%	http://www.ijmlc.org/vol6/594-L00... http://ijmlc.org	12 Мая 2024	Интернет Плюс*	
[10]	1,62%	0,71%	courses:ml:%d0%be%d0%b1%d1... https://se.moevm.info	29 Мар 2023	Перефразированные заимствования по коллекции Интернет в русском сегменте	
[11]	1,58%	1,07%	120479 http://biblioclub.ru	15 Апр 2016	Сводная коллекция ЭБС	
[12]	1,58%	0%	Обучение с подкреплением http://biblioclub.ru	21 Янв 2020	Сводная коллекция ЭБС	
[13]	1,55%	0%	Neural Network Based Online Sim... https://ieeexplore.ieee.org	12 Окт 2012	IEEE	
[14]	1,48%	0%	Actor-critic neural network reinfor... https://ieeexplore.ieee.org	18 Дек 2014	IEEE	
[15]	1,43%	0%	Off-Policy Reinforcement Learning... https://ieeexplore.ieee.org	09 Мая 2014	IEEE	
[16]	1,38%	1,38%	Нейронные сети: полный курс - ... https://djvu.online	20 Апр 2024	Перефразированные заимствования по коллекции Интернет в русском сегменте	
[17]	1,37%	0,14%	Reinforcement learning of LQR co... https://ieeexplore.ieee.org	09 Июн 2023	IEEE	
[18]	1,32%	0%	Off-policy reinforcement learning ... https://arxiv.org	10 Июл 2017	Интернет Плюс*	
[19]	1,32%	0%	Neural Network Based Online Sim... https://doi.org	06 Сен 2019	Интернет Плюс*	
[20]	1,31%	0%	https://arxiv.org/pdf/1311.6107.pdf https://arxiv.org	12 Мая 2024	Интернет Плюс*	

[21]	1,3%	0%	基于数据的自学习优化控制:研究进展... http://aas.net.cn	12 Мая 2024	Интернет Плюс*	
[22]	1,29%	0%	https://www.princeton.edu/~nda... https://princeton.edu	06 Окт 2022	Интернет Плюс*	
[23]	1,29%	0%	66325 http://e.lanbook.com	09 Мар 2016	Сводная коллекция ЭБС	
[24]	1,29%	0%	Обучение с подкреплением https://e.lanbook.com	22 Янв 2020	Сводная коллекция ЭБС	
[25]	1,28%	0%	Optimal Adaptive Control of Const... https://ieeexplore.ieee.org	05 Апр 2024	Перефразирования по коллекции IEEE	
[26]	1,28%	0%	[IEEE 2012 American Control Conf... https://doi.org	18 Дек 2018	Интернет Плюс*	
[27]	1,27%	0%	Model-based vs data-driven adap... https://doi.org	31 Мая 2018	Издательство Wiley	
[28]	1,14%	0%	Model-Free H∞ Control Design for...	31 Мар 2016	Издательство Wiley	
[29]	1,13%	0%	Adaptive Optimal Control of Highl... https://doi.org	19 Окт 2019	Интернет Плюс*	
[30]	1,13%	0%	Online adaptive algorithm for opti... https://doi.org	18 Апр 2021	Перефразирования по коллекции издательства Wiley	
[31]	1%	0%	Model-based reinforcement learn... https://doi.org	01 Окт 2020	Перефразирования по коллекции издательства Wiley	
[32]	0,97%	0%	Data-Driven Adaptive Critic Appro... https://doi.org	31 Янв 2018	Издательство Wiley	Источник исключен. Причина: Маленький процент пересечения.
[33]	0,94%	0%	Reinforcement learning control fo... https://core.ac.uk	30 Окт 2020	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[34]	0,9%	0%	https://ieeexplore.ieee.org/ielaam... https://ieeexplore.ieee.org	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[35]	0,89%	0,89%	не указано	13 Янв 2022	Шаблонные фразы	
[36]	0,86%	0%	Adaptive optimal tracking control ... https://ieeexplore.ieee.org	29 Окт 2015	IEEE	Источник исключен. Причина: Маленький процент пересечения.
[37]	0,86%	0%	Adaptive neural network-based o... https://doi.org	04 Сен 2013	Издательство Wiley	Источник исключен. Причина: Маленький процент пересечения.
[38]	0,77%	0,77%	CITech_abstracts.pdf http://conf.nsc.ru	04 Мая 2023	Перефразированные заимствования по коллекции Интернет в английском сегменте	
[39]	0,72%	0%	Adaptive Optimal Control of Unkn... https://ieeexplore.ieee.org	03 Мая 2024	Перефразирования по коллекции IEEE	
[40]	0,67%	0%	Online solution of nonlinear two-... https://doi.org	21 Июл 2011	Издательство Wiley	Источник исключен. Причина: Маленький процент пересечения.
[41]	0,62%	0%	Off-policy reinforcement learning ... http://arxiv.org	08 Янв 2018	Перефразирования по Интернету (EN)	
[42]	0,6%	0%	melcer_s_p_issledovanie-sredy-s-...	09 Янв 2024	Кольцо вузов	
[43]	0,59%	0%	Autonomy and Machine Intelligen... http://ece.ucsb.edu	09 Янв 2018	Перефразирования по Интернету (EN)	
[44]	0,56%	0%	http://www.mit.edu/~dimitrib/RL_... http://mit.edu	28 Мая 2022	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[45]	0,52%	0%	zisman_i_a_proekt-kooperaciya-eg...	09 Янв 2024	Кольцо вузов	
[46]	0,47%	0%	Review on Reinforcement Learnin... https://ieeexplore.ieee.org	30 Дек 2022	IEEE	Источник исключен. Причина: Маленький процент пересечения.
[47]	0,46%	0,46%	Выборнов, Андрей Олегович дис... http://dlib.rsl.ru	20 Янв 2010	Публикации РГБ (переводы и перефразирования)	
[48]	0,45%	0%	https://hal.science/hal-00756747/f... https://hal.science	24 Фев 2023	Перефразированные заимствования по коллекции Интернет в английском сегменте	
[49]	0,45%	0%	https://hal.archives-ouvertes.fr/ha... https://hal.archives-ouvertes.fr	28 Ноя 2022	Перефразированные заимствования по коллекции Интернет в английском сегменте	
[50]	0,45%	0%	A Survey of Actor-Critic Reinforce... https://ieeexplore.ieee.org	22 Июн 2023	Перефразирования по коллекции IEEE	
[51]	0,44%	0%	Multi Objective Combinatorial Opt...	09 Янв 2024	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[52]	0,43%	0%	ISBN9785996325009.txt	26 Окт 2017	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[53]	0,41%	0%	https://spb.hse.ru/mirror/pubs/sh... https://spb.hse.ru	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.

[54]	0,41%	0%	https://spb.hse.ru/mirror/pubs/sh... https://spb.hse.ru	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[55]	0,4%	0%	Q-Learning Algorithms: A Compre... https://ieeexplore.ieee.org	13 Сен 2019	IEEE	Источник исключен. Причина: Маленький процент пересечения.
[56]	0,4%	0%	Adaptive Optimal Control of Unkn... https://ieeexplore.ieee.org	03 Мая 2024	Переводные заимствования IEEE	Источник исключен. Причина: Маленький процент пересечения.
[57]	0,4%	0%	Novel Discounted Adaptive Critic ... https://ieeexplore.ieee.org	23 Апр 2024	Переводные заимствования IEEE	Источник исключен. Причина: Маленький процент пересечения.
[58]	0,39%	0%	Robust Adaptive Dynamic Progra... http://arxiv.org	31 Янв 2017	Перефразирования по Интернету (EN)	Источник исключен. Причина: Маленький процент пересечения.
[59]	0,38%	0%	Model-based and Model-free Opti... https://ieeexplore.ieee.org	12 Июл 2023	Переводные заимствования IEEE	Источник исключен. Причина: Маленький процент пересечения.
[60]	0,37%	0%	Предотвращение столкновений ... http://elibrary.ru	10 Янв 2019	Публикации eLIBRARY (переводы и перефразирования)	Источник исключен. Причина: Маленький процент пересечения.
[61]	0,36%	0%	Ascertaining properties of weighti...	10 Ноя 2020	Издательство Wiley	Источник исключен. Причина: Маленький процент пересечения.
[62]	0,36%	0%	Data-Driven Actuator Allocation fo... https://ieeexplore.ieee.org	05 Апр 2024	Перефразирования по коллекции IEEE	Источник исключен. Причина: Маленький процент пересечения.
[63]	0,36%	0%	Adaptive Optimal Stabilization Re... https://ieeexplore.ieee.org	28 Мар 2024	Перефразирования по коллекции IEEE	Источник исключен. Причина: Маленький процент пересечения.
[64]	0,35%	0%	Robust Policy Iteration of Uncertai... https://ieeexplore.ieee.org	19 Апр 2024	Перефразирования по коллекции IEEE	Источник исключен. Причина: Маленький процент пересечения.
[65]	0,34%	0%	Design and Comparison Base Ana... https://doi.org	25 Мая 2021	Перефразирования по коллекции издательства Wiley	Источник исключен. Причина: Маленький процент пересечения.
[66]	0,33%	0%	Data-driven Control and Optimiza... http://uta.edu	06 Янв 2018	Перефразирования по Интернету (EN)	Источник исключен. Причина: Маленький процент пересечения.
[67]	0,33%	0%	Integral Reinforcement Learning f... http://uta.edu	07 Янв 2018	Перефразирования по Интернету (EN)	Источник исключен. Причина: Маленький процент пересечения.
[68]	0,32%	0%	ОБУЧЕНИЕ С ПОДКРЕПЛЕНИЕМ.	19 Окт 2023	Публикации eLIBRARY (переводы и перефразирования)	Источник исключен. Причина: Маленький процент пересечения.
[69]	0,32%	0%	Стратегии исследования окруже... http://machinelearning.ru	03 Июн 2020	Переводные заимствования по коллекции Интернет в русском сегменте	Источник исключен. Причина: Маленький процент пересечения.
[70]	0,31%	0%	https://books.ifmo.ru/file/pdf/258... https://books.ifmo.ru	07 Июн 2022	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[71]	0,31%	0%	Data-Driven Adaptive Critic Appro... https://doi.org	31 Янв 2018	Перефразирования по коллекции издательства Wiley	Источник исключен. Причина: Маленький процент пересечения.
[72]	0,31%	0%	Adaptive optimal tracking control ... https://sciencedirect.com	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[73]	0,29%	0%	https://arxiv.org/pdf/1908.02077.p... https://arxiv.org	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[74]	0,29%	0%	Using Social Cues to Recognize Ta... https://arxiv.org	25 Мар 2023	Перефразированные заимствования по коллекции Интернет в английском сегменте	Источник исключен. Причина: Маленький процент пересечения.
[75]	0,29%	0%	«Стратегии исследования окруж... http://machinelearning.ru	09 Июн 2020	Переводные заимствования по коллекции Интернет в русском сегменте	Источник исключен. Причина: Маленький процент пересечения.
[76]	0,28%	0%	6342-9949-1-SM.pdf	15 Янв 2024	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[77]	0,28%	0%	https://ctlab.itmo.ru/~vaksenov/di... https://ctlab.itmo.ru	20 Апр 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[78]	0,27%	0%	https://www.cs.colostate.edu/~an... https://cs.colostate.edu	20 Мая 2022	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[79]	0,27%	0%	Machine Learning-Based Researc... https://link.springer.com	11 Дек 2022	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[80]	0,26%	0%	Обучение с подкреплением http://studentlibrary.ru	20 Дек 2016	Медицина	Источник исключен. Причина: Маленький процент пересечения.
[81]	0,25%	0%	chepenko_d_d_offlayn-algoritm-ob...	24 Мая 2022	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[82]	0,24%	0%	tarelov_v_o_proekt-razrabotka-inte...	01 Мар 2024	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[83]	0,24%	0%	Распознавание нечётко определ... http://studentlibrary.ru	20 Дек 2016	Медицина	Источник исключен. Причина: Маленький процент пересечения.
[84]	0,23%	0%	Root Locus Analysis for Swinging ... https://ieeexplore.ieee.org	07 Авг 2023	IEEE	Источник исключен. Причина: Маленький процент пересечения.
[85]	0,21%	0%	Solutions for the Linear-Quadratic...	раньше 2011	Публикации eLIBRARY	Источник исключен. Причина: Маленький процент пересечения.
[86]	0,21%	0%	Data-Driven Optimal Assistance C... https://frontiersin.org	06 Июл 2020	СМИ России и СНГ	Источник исключен. Причина: Маленький процент пересечения.

[87]	0,21%	0%	Survey of Image Processing Techn... https://frontiersin.org	22 Мая 2020	СМИ России и СНГ	Источник исключен. Причина: Маленький процент пересечения.
[88]	0,19%	0%	Шизофрения и вероятностное п... http://emll.ru	21 Дек 2016	Медицина	Источник исключен. Причина: Маленький процент пересечения.
[89]	0,18%	0%	MDP (марковский процесс приня... https://mlcentre.ru	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[90]	0,17%	0%	Качков_магистерская	11 Июн 2018	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[91]	0,17%	0%	Leveraging explainable machine l... https://doi.org	25 Сен 2023	Издательство Wiley	Источник исключен. Причина: Маленький процент пересечения.
[92]	0,17%	0%	https://nauchkor.ru/uploads/docu... https://nauchkor.ru	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[93]	0,17%	0%	Интеллектуальное управление и... https://nauchkor.ru	10 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[94]	0,16%	0%	Обучение с подкреплением: вве... https://russianblogs.com	19 Мая 2021	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[95]	0,16%	0%	Frontiers | From classical to quan... https://translated.turbopages.org	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[96]	0,16%	0%	smirnov_p_d_obuchenie-s-podkre...	09 Янв 2024	Кольцо вузов	Источник исключен. Причина: Маленький процент пересечения.
[97]	0,15%	0%	Применение глубокого обучения... http://elibrary.ru	01 Янв 2023	Публикации eLIBRARY	Источник исключен. Причина: Маленький процент пересечения.
[98]	0,15%	0%	Многомерные системы с неопре... http://elibrary.ru	28 Авг 2014	Публикации eLIBRARY	Источник исключен. Причина: Маленький процент пересечения.
[99]	0,14%	0%	Диссертация на тему «Метод и а... https://dissercat.com	12 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[100]	0,13%	0%	Что такое обучение усилению? https://machinelearningmastery.ru	09 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[101]	0,11%	0%	https://e-learning.bmstu.ru/iu6/pl... https://e-learning.bmstu.ru	11 Мая 2024	Интернет Плюс*	Источник исключен. Причина: Маленький процент пересечения.
[102]	0,11%	0%	Convergence of the approximate ...	01 Янв 2015	Публикации eLIBRARY	Источник исключен. Причина: Маленький процент пересечения.

ИССЛЕДОВАНИЕ АЛГОРИТМОВ ОПТИМАЛЬНОГО УПРАВЛЕНИЯ, ОСНОВАННЫХ НА ОБУЧЕНИИ С ПОДКРЕПЛЕНИЕМ

Титульный лист

Техническое задание (утвержденное из ИСУ)

Основные вопросы, подлежащие разработке / Key issues to be analyzed

Провести исследование метода обучения с подкреплением и способов его применения в задачах оптимизации, а также основных алгоритмов, основанных на этом методе.

Провести исследование оптимальных регуляторов на основе обучения с подкреплением.

Реализовать оптимальный адаптивный регулятор для динамических систем на основе алгоритмов обучения с подкреплением.

Провести моделирование оптимального алгоритма управления на основе обучения с подкреплением с объектами управления в MATLAB, выполнить анализ полученных результатов.

Аннотация

Цель исследования / Research goal

Исследование алгоритмов оптимального управления, основанных на обучении с подкреплением

Задачи, решаемые в ВКР / Research tasks

Для достижения цели требуется выполнение следующих подзадач: 1. Провести исследование метода обучения с подкреплением и способов его применения в задачах оптимизации, а также основных алгоритмов, основанных на этом методе. 2. Провести исследование оптимальных регуляторов на основе обучения с подкреплением 3. Реализовать оптимальный адаптивный регулятор для динамических систем на основе алгоритмов обучения с интегральным подкреплением 4. Провести моделирование оптимального алгоритма управления на основе обучения с подкреплением с объектами управления в симуляторе Matlab 5. Провести анализу работоспособности регулятора, основанного на обучении с подкреплением, на основе полученных результатов моделирования и сравните их с обычным оптимальным регулятором.

Краткая характеристика полученных результатов / Short summary of results/findings

Изучены типичные идеи обучения с подкреплением. Рассмотрены и исследованы математические модели построения оптимальных адаптивных регуляторов. Моделирование алгоритма выполнено в Matlab. В данной работе приближенное динамическое программирование в сочетании с алгоритмами обучения с подкреплением успешно нашло приближенное решение уравнения Гамильтона Якоби-Беллмана в реальном времени, тем самым было успешно реализовано построение адаптивного оптимального регулятора на основе алгоритмов обучения с интегральным подкреплением для непрерывных систем. Результаты сравнения и анализа показывают, что оптимальный адаптивный регулятор, основанный на обучении с подкреплением имеет примерно такую же производительность, как и оптимальный регулятор.

Оглавление

Введение

Оптимальные регуляторы обычно разрабатываются в автономном режиме путем решения уравнений Гамильтона-Якоби-Беллмана (HJB), например, уравнения Риккати, с использованием полных знаний о динамике системы. Такой регулятор имеет множество применений в науке, технике и исследованиях. Для расширения применимости оптимального управления необходима разработка алгоритмов оптимального управления, способных адаптироваться к изменению динамических свойств системы. Адаптивные регуляторы учатся в режиме реального времени управлять системами с неизвестными параметрами, используя данные, измеряемые в моменте вдоль траекторий системы. Однако адаптивные регуляторы обычно не сходятся к оптимальным решениям [1]. Метод обучения с подкреплением (RL) и адаптивное динамическое программирование стали прочной основой для разработки адаптивных оптимальных регуляторов. Такой метод предполагает причинно-следственную связь между действиями и вознаграждением или наказанием. Агент взаимодействует со своей средой посредством действия, и за этим действием следует вознаграждение (положительный сигнал RL), что дает снижение затрат на управление, а наказание (отрицательный сигнал RL) - увеличение затрат на управление. Действуя таким образом, алгоритм RL со временем обучается оптимальной стратегии. Алгоритм применяется для решения множества задач оптимального управления, включая обеспечение устойчивости, подавление шумов, оптимальное слежение за траекторией и т. д., без необходимости решения уравнений Гамильтона - Якоби - Беллмана, что позволяет решать задачу управления в условиях неполной информации о динамике объекта.

Список использованных источников:

1. Optimal control, Frank L. Lewis, Draguna L. Vrabie, Vassilis L. Syrmos.-3rd ed, 2012.

2. Kyriakos G. Vamvoudakis, Draguna Vrabie, Frank L. Lewis, Online adaptive learning of optimal control solutions using integral reinforcement learning, 2011.

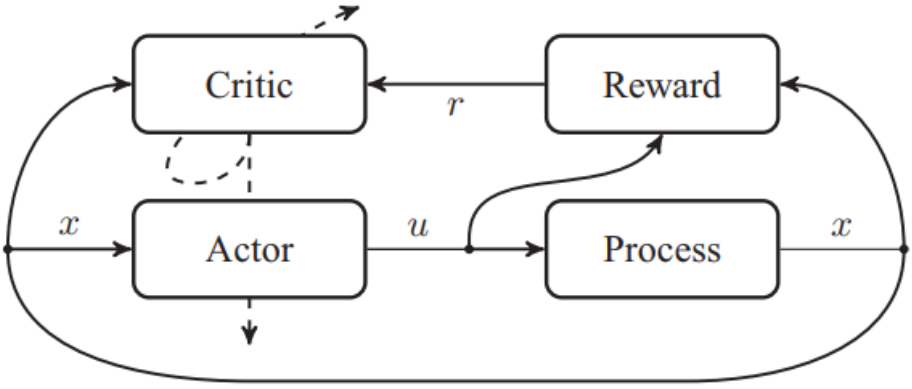
3. Bahare Kiumarsi, Member, IEEE, Kyriakos G. Vamvoudakis, Senior Member, IEEE, Hamidreza Modares, Member, IEEE, and Frank L. Lewis, Fellow, IEEE, Optimal and Autonomous Control Using Reinforcement Learning: A Survey, 2018.

1 Обучение с подкреплением

1.1 Обзор обучения с подкреплением (RL)

Обучение с подкреплением — это одно из машинного обучения. Речь идет о принятии правильных действий для получения максимальной выгоды в конкретной ситуации. Обучение с подкреплением (RL) — это наука о принятии решений. Речь идет об обучении оптимальному поведению в окружающей среде для достижения максимального вознаграждения. В области теории управления обучение с подкреплением относится к методу, который позволяет разрабатывать адаптивные регуляторы в реальном времени для решения определяемых пользователем задач оптимального управления.

Обучение с подкреплением подразумевает наличие причинно-следственной связи между действиями и вознаграждением или наказанием. Это структура, в которой агент (или регулятор) оптимизирует свое поведение, взаимодействуя с окружающей средой. Совершив действие в каком-то состоянии, агент получает скалярное вознаграждение или наказание от среды, которое дает агенту представление о качестве этого действия. Алгоритмы RL исходят из идеи, что успешное поведение (приносящее высокие награды) будет запоминаться в том смысле, что оно имеет тенденцию использоваться повторно в последующих случаях [1]. Идея RL возникла из экспериментов обучения в биологии. RL теоретически тесно связан прямо и косвенно с адаптивными методами оптимального управления.



Рисунка 1. Схематический обзор алгоритма актер-критик [2]

Одной из популярных структур RL является структура «Актер-критик» [Barto, Sutton, Anderson 1983], Актерско-критические методы сочетают в себе преимущества методов "только актер" и "только критик" [3], в которой компонент «Актер» выполняет действия (стратегия управления), влияющие на окружающую среду, а секция «Критик» оценивает это действие. На основе этой оценки используется множество методов для калибровки или улучшения действия, чтобы новое действие создавало большую ценность (награду), чем

предыдущее. Таким образом, структура «Актер-Критик» состоит из двух этапов: оценка поведения и улучшение поведения. Поведенческая оценка проводится путем наблюдения за результатами, полученными из окружающей среды после выполнения определенного поведения.

Метод «Актер-критик» обычно обладают хорошей сходимостью, в отличие от методов, основанных только на критике [4].

1.2 Элемент обучения с подкреплением

В этом разделе представлены некоторые элементы и термины обучения с подкреплением. Можно также рассмотреть некоторые элементы обучения с подкреплением, представленные в [5].

В обучении с подкреплением есть термин «агент», который означает субъект, который взаимодействует с окружающей средой посредством действий.

Среда - это окружающее агента пространство, в котором он существует и взаимодействует.

Действия - это методы агента, которые позволяют ему взаимодействовать с окружающей средой и изменять ее. Основываясь на состоянии S(t) текущей среды, агент будет выполнять действия a(t)

Стратегия - это то, что определяет, как агент действует в данный момент времени. Другими словами, стратегия - это отображение состояний среды на действия, которые будут выполняться в этих состояниях. Политика является ядром агента при определении поведения. В некоторых случаях политика может быть простой функцией или таблицей поиска. В других случаях политика может включать в себя обширные вычисления, например процесс поиска.

Вознаграждение - За каждое действие среда посылает агенту определенное вознаграждение. Цель агента - максимизировать общее вознаграждение, которое он получает в течение длительного периода времени. Сигнал о вознаграждении помогает определить, какие события являются хорошими, а какие плохими для агента, а также является основной базой для изменения стратегии. Если действие, выбранное в соответствии с стратегией, приносит низкое вознаграждение, стратегия может быть изменена. В будущем агент будет выбирать другие действия в аналогичных ситуациях.

Исследование и использование - Одна из проблем, возникающих в обучении с подкреплением - это компромисс между использованием и исследованием. Чтобы получить большее вознаграждение, агент должен отдавать предпочтение тем действиям, которые уже пробовал в прошлом и которые помогли ему достичь вознаграждения. Агент рассматривает все возможные действия для данного состояния, а затем выбирает действия, основываясь на максимальном значении этих действий. Это называется использованием, поскольку для принятия решения мы используем доступную информацию.

Кроме того, вместо выбора действий, основанных на максимальном будущем вознаграждении, агент может выбирать действия случайным образом. Случайные действия важны, поскольку они позволяют агенту исследовать и обнаруживать новые состояния, которые не были выбраны во время операции.

Агент должен использовать то, что он испытал, для получения вознаграждения, но он также должен исследовать, чтобы сделать лучший выбор в отношении будущих действий.

1.3 Марковский процесс принятия решений (МППР) - Markov Decision Processe

RL можно выразить с помощью марковского процесса принятия решений (МППР), как показано на рисунке 2. Каждое окружение представлено состоянием, которое отражает происходящее в окружении. Агент RL совершает действия в среде, которые вызывают изменение текущего состояния среды, порождая новое состояние, и получает вознаграждение в зависимости от результатов. Агент получает положительное вознаграждение за хорошие действия и отрицательное - за плохие, что помогает ему оценить выполненное действие в данном состоянии и учиться на опыте. Идея МППР представлена в статье [7].

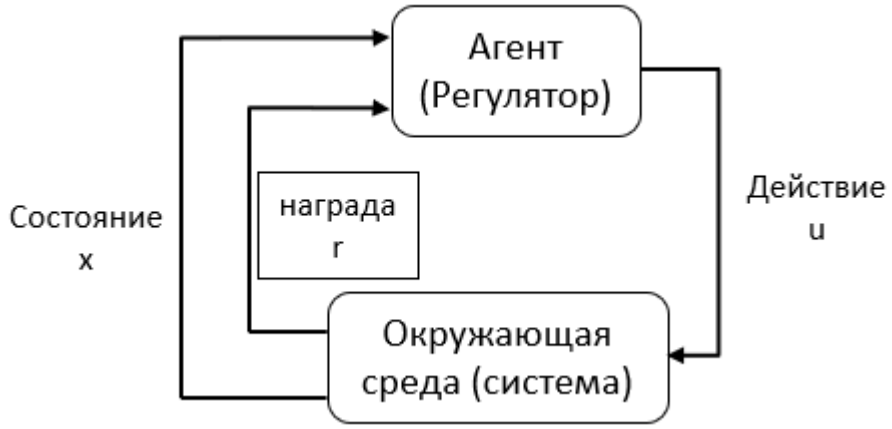


Рисунок 2.1 - Взаимодействие агента и среды в марковском процессе принятия решений.

Рассмотрим МППР *(X, U, P, R)*, где X — набор состояний, а U — набор действий или элементов управления. Вероятности перехода $P:X\times U\times X \to [0,1]$ описывают что для каждого состояния $x\in X$, и действие $u\in U$, условная вероятность $P_{x,x'u} = Pr_{x'}\ x,u\}$ перехода в состояние $x' \in X$ учитывая, что МППР находится в состоянии x и выполняет действие u. Функция ценности $R:X\times U\times X\to R$ это ожидаемая непосредственная стоимость $R_{x,x'u}$ уплаченная после перехода в состояние $x' \in X$ учитывая, что МППР запускается в состоянии $x\in X$ и выполняет действие u . Свойство Маркова относится к тому факту, что вероятности перехода $P_{x,x'u}$ зависят только от текущего состояния x, а не от истории того, как МППР достигла этого состояния.

Основная проблема МППР состоит в том, чтобы найти отображение $\pi:X\times U\to[0,1]$, которое дает для каждого состояния x и действия u условную вероятность $\pi_{x,u} = P_{rux}$ выполнения действия u при условии, что МППР находится в состоянии x. Такое отображение называется управлением с обратной связью или стратегией действий или стратегией. Стратегия $\pi_{x,u} =$

Prux называется стохастической или смешанной, если существует ненулевая вероятность выбора более одного элемента управления в состоянии x. Смешанные стратегии можно рассматривать как векторы распределения вероятностей, имеющие в качестве компонента i вероятность выбора i-го управляющего воздействия в состоянии x∈X . Если отображение π x,u = Prux допускает только одно управление с вероятностью единица, то в каждом состоянии x отображение называется детерминированной стратегией. Тогда π x,u = Prux соответствует функции, отображающей состояния в управления μ x :X→U .

МППР, которые имеют конечные пространства состояний и действий, называются конечными МППР.

1.3.1 Задачи оптимального последовательного решения

Обозначим значения состояния и действия в момент времени k через xk , uk . Зачастую желательно, чтобы системы, спроектированные человеком, были оптимальными с точки зрения экономии ресурсов, таких как стоимость, время, топливо и энергия.

Определение ценности этапа во время k по формуле:

rk = rk xk, uk, xk+1 (1)

Тогда функция ценности:

Rx,x'u =E rk xk =x, uk =u, xk+1 =x '} (2)

с E . - оператором ожидаемого значения.

Определение индекса производительности как сумму будущих затрат за интервал времени [k, k+T] :

Jk,T = i=0Tγirk+1 = i=kk+Tγi-kri (3)

где γ - коэффициент дисконтирования, уменьшающий вес затрат, понесенных в дальнейшем.

Предположим, что агент выбирает стратегию управления πk (xk , uk) , которая используется на каждом этапе k МППР. Нас в первую очередь интересуют стационарные стратегия, где условные вероятности πk (xk , uk) не зависят от k. Тогда πk x, u =π x,u = Pru x} для всех k . Нестационарная детерминированная стратегии имеет вид π={ μ0 , μ2, ...} , где каждая запись представляет собой функцию μk x :X→ U; k=0, 1, ... Стационарные детерминированные стратегия не зависят от времени, т. е. имеют вид π= μ, μ, ...

Выберите фиксированную стационарную стратегию π x,u = Pru x} . Тогда "замкнутая" МППР сводится к цепи Маркова с пространством состояний X. То есть вероятности переходов между состояниями фиксированы, и дальнейшая свобода выбора действий отсутствует. Вероятности переходов этой цепи Маркова задаются следующим образом:

px, x' ≡ Px, x'π = uPrx'x, u Prux = uπx,uPx, x'u (4)

где используется тождество Чепмена-Колмогорова [3].

Ценность стратегии определяется как условное ожидаемое значение будущих затрат при запуске в состоянии x в момент k и после этого следовании стратегии π(x,u) :

$$V_k^{\pi}(x) = E^{\pi}\{J_k, T(x_k) = x\} = E^{\pi}\{\sum_{i=k}^{k+T-1} \gamma^{i-k} r_i \mid x_k = x\} \quad (5)$$

где $E^{\pi}\{.\}$ - ожидаемое значение при условии, что агент придерживается стратегии $\pi(x,u)$, и $V^{\pi}(x)$ называется функцией ценности для стратегии $\pi(x,u)$, что является ценностью пребывания в состоянии x с учетом того, что является ценностью пребывания в состоянии x, учитывая, что стратегия $\pi(x,u)$.

Основная цель МППР — определить стратегию $\pi(x,u)$, позволяющую минимизировать ожидаемые будущие затраты

$$\pi^*(x,u) = \arg\min_{\pi} V_k^{\pi}(s) = \arg\min_{\pi} E^{\pi}\{\sum_{i=k}^{k+T-1} \gamma^{i-k} r_i \mid x_k = x\} \quad (6)$$

Такая стратегия называется оптимальной, а соответствующее оптимальное значение задается как

$$V_k^*(x) = \min_{\pi} V_k^{\pi}(x) = \min_{\pi} E^{\pi}\{\sum_{i=k}^{k+T-1} \gamma^{i-k} r_i \mid x_k = x\} \quad (7)$$

1.3.2 Обратная рекурсия для значения

Используя тождество Чепмена-Колмогорова и свойство Маркова, значение стратегии $\pi(x,u)$ можно записать как

$$V_k^{\pi}(x) = E^{\pi}\{J_k, T(x_k) = x\} = E^{\pi}\{\sum_{i=k}^{k+T-1} \gamma^{i-k} r_i \mid x_k = x\}$$

$$V_k^{\pi}(x) = E^{\pi}\{\sum_{i=k+1}^{k+T-1} \gamma^{i-(k+1)} r_i \mid x_k = x\},$$

$$V_k^{\pi}(x) = u^{\pi}(x,u)\{x'^{T}(x,u) R(x,u) + \gamma E^{\pi}\{\sum_{i=k+1}^{k+T-1} \gamma^{i-(k+1)} r_i \mid x_{k+1} = x'\} \quad (8)$$

Поэтому функция ценности для стратегии $\pi(x,u)$ удовлетворяет:

$$V_k^{\pi}(x) = u^{\pi}(x,u)\{x'^{T}(x,u) R(x,u) + \gamma V_{k+1}^{\pi}(x')\} \quad (9)$$

Это уравнение обеспечивает обратную рекурсию для значения в момент времени k в терминах значения в момент времени $k+1$.

1.3.3 Обзор адаптивного динамического программирования

Оптимальную стоимость можно записать как

$$V_k^*(x) = \min_{\pi} V_k^{\pi}(x) = \min_{\pi} u^{\pi}(x,u)\{x'^{T}(x,u) R(x,u) + \gamma V_{k+1}^{\pi}(x')\} \quad (10)$$

Принцип оптимальности Беллмана гласит: "Оптимальная стратегия обладает тем свойством, что независимо от того, какими были предыдущие управляющие воздействия, оставшиеся управления представляют собой оптимальную стратегию по отношению к состоянию, возникшему в результате этих предыдущих управлений". Поэтому мы можем записать:

$$V_k^*(x) = \min_{\pi} u^{\pi}(x,u)\{x'^{T}(x,u) R(x,u) + \gamma V_{k+1}^{\pi}(x')\} \quad (11)$$

Предположим, что в момент времени k применяется произвольное управление u, а начиная с момента времени $k+1$ применяется оптимальная . Тогда принцип оптимальности Беллмана гласит, что оптимальное управление в момент времени k задается следующим образом:

$$\pi^*(x_k = x, u) = \arg\min_{\pi} u^{\pi}(x,u)\{x'^{T}(x,u) R(x,u) + \gamma V_{k+1}^{\pi}(x')\} \quad (12)$$

В предположении, что цепь Маркова, соответствующая каждой стратегии, с вероятностями перехода, является эргодической, каждая МППР имеет стационарную детерминированную оптимальную стратегию. Тогда мы можем

эквивалентн минимизировать условное ожидание по всем действиям u в состоянии x. Поэтому:

$$Vk^*\ x = \min u' Pxx' u Rxx' u + \gamma Vk+1^* x' \quad (13)$$

$$uk^* = \arg\min u' Pxx' u Rxx' u + \gamma Vk+1^* x' \quad (14)$$

Обратная рекурсия (11), (13) составляет основу динамического программирования (ДП), которое дает автономные методы для работы в обратном направлении во времени для определения оптимальной стратегии. ДП - это автономная процедура поиска оптимального значения и оптимальных стратегий, которая требует знания полной динамики системы в виде вероятностей перехода $Px,x'u = Prx',u$ и ожидаемых затрат $Rxx'u = E\{ rk \mid xk = x, uk = u, xk+1 = x' \}$.

1.3.4 Уравнение Беллмана и уравнение оптимальности Беллмана

Обучение с подкреплением включает в себя определение решений для управления в реальном времени и с упреждением во времени. Ключом к этой проблеме является уравнение Беллмана, которое будет представлено ниже.

Динамическое программирование — это метод поиска оптимального значения и стратегия в обратном времени. В отличие от этого, обучение с подкреплением направлено на поиск оптимальной стратегии на основе каузального опыта путем выполнения последовательных решений, которые улучшают управляющие действия на основе наблюдаемых результатов использования текущей стратегии.

Эта процедура требует выведения методов поиска оптимальных значений и оптимальных стратегий, которые могут быть выполнены в прямом времени.

Чтобы получить методы, направленные в прямом времени, для поиска оптимальных значений и оптимальных стратегий, устанавливается временной горизонт T равным бесконечности и определяется стоимость бесконечного горизонта

$$Jk = \sum_{i=0}^{\infty} \gamma^i rk+1 = \sum_{i=k}^{\infty} \gamma^{i-k} ri \quad (15)$$

Бесконечная функция ценности связанного горизонта для стратегии $\pi(x,u)$ равна

$$V\pi\ x = E\pi\ Jk \mid xk=x = E\pi \sum_{i=k}^{\infty} \gamma^{i-k} ri \mid xk=x \quad (16)$$

Используя (4) с $T= \infty$, можно увидеть, что функция ценности стратегии $\pi(x, u)$ удовлетворяет уравнению Беллмана

$$V\pi\ x = u\pi x,u\ x' Pxx' u\ Rxx' u + \gamma Vk+1 \pi x' \quad (17)$$

Важность этого уравнения заключается в том, что в нем с обеих сторон фигурирует одна и та же функция стоимости, что связано с тем, что используется бесконечный горизонт стоимости. Поэтому уравнение Беллмана (17) можно интерпретировать как уравнение согласованности, которому должна удовлетворять функция стоимости на каждом временном этапе. Оно выражает связь между текущей стоимостью нахождения в состоянии x и стоимостью нахождения в следующем состоянии x' при использовании стратегии $\pi(x,u)$.

Если МППР конечен и имеет N состояний, то уравнение Беллмана (17) представляет собой систему N одновременных линейных уравнений для

значения $V\pi(x)$ пребывания в каждом состоянии x при текущей стратегии π x, u .

Оптимальное значение бесконечного горизонта удовлетворяет

$$V^* x = \min\pi V\pi(x) = \min\pi u\pi x,ux'Pxx'uRxx'u+\gamma V\pi x' \quad (18)$$

Тогда принцип оптимальности Беллмана приводит к уравнению оптимальности Беллмана

$$V^* x = \min\pi V\pi(x) = \min\pi u\pi x,ux'Pxx'uRxx'u+\gamma V^*x' \quad (19)$$

Эквивалентно, при условии эргодичности цепей Маркова, соответствующих каждой стратегии, уравнение оптимальности Беллмана можно записать как

$$V^* x = \min ux'Pxx'uRxx'u+\gamma V^*x' \quad (20)$$

Это уравнение известно как уравнение Гамильтона – Якоби – Беллмана (HJB) в системах управления. Если МППР конечен и имеет N состояний, то уравнение оптимальности Беллмана представляет собой систему N нелинейных уравнений для оптимального значения $V^*(x)$ пребывания в каждом состоянии. Оптимальное управление определяется выражением

$$u^* = \arg\min ux'Pxx'uRxx'u+\gamma V^*x' \quad (21)$$

1.4 Итерация по стратегии и итерация по критерию

Чтобы найти оптимальную стратегию для данной задачи, в обучении с подкреплением обычно используются два метода: итерация по стратегии и итерация по критерию.

Эти два метода являются методами автономной оптимизации в обучении с подкреплением. Чтобы иметь возможность реализовать методы онлайн-оптимизации, необходимо объединить их с методами приближенного динамического программирования. Это то, что нам нужно для построения алгоритма оптимального адаптивного регулятор в реальном времени.

1.4.1 Итерация по стратегии алгоритм

Чтобы понять алгоритм итерации по стратегии, нам сначала нужно обобщить функцию значения.

$$V\pi x = u\pi x,u\ x'Pxx'uRxx'u+\gamma V\pi x' \quad (22)$$

Это равенство называется равенством Беллмана и упрощает вычисление функции ценности за счет использования динамического программирования.

Алгоритм итерации по стратегии представлен следующим образом:

Выберите начальную стратегию $\pi 0(x,u)$. Начиная с j = 0, повторите по j до сходимости

Оценка стратегии (обновление ценности)

$$Vj x = u\pi jx,u\ x'Pxx'uRxx'u+\gamma Vj\pi x' , \forall x \in X \quad (23)$$

Улучшение стратегии (обновление стратегии)

$$\pi_{j+1}(x,u) = \arg\min_{\pi} x'Pxx'uRxx'u+\gamma V_j \pi x', \forall x \in X \quad (24)$$

На каждой итерации j, алгоритм итерации по стратегии определяет решение уравнения Беллмана для определения значения функции ценности $V_j(x)$, соответствующей текущей стратегии $\pi_j(x,u)$. Тогда улучшение стратегии осуществляется благодаря уравнению (24). Эти шаги повторяются до тех пор, пока и функция ценности, и стратегия придут к своим оптимальным значениям. Чтобы алгоритм работал корректно, нам нужно обратить внимание на начальную инициализацию стратегию $\pi_0(x,u)$ и начальное начальное значение $V_1 \leq V_0$.

1.4.2 Итерации по критерию

Алгоритм итерация по критерию основан на чрезвычайно простой, но очень эффективной идее. Рассматривая алгоритм, итерация по стратегии, мы видим, что он одновременно поддерживает и стратегию, и функции ценности, что делает вычисления громоздкими и трудоемкими. Алгоритм итерация по критерию начинает с того, что сначала пытается оптимизировать функцию ценности, затем стратегия, соответствующая оптимальной функции ценности, конечно же, также будет оптимальной стратегией.

Алгоритм итерации по стратегии представлен следующим образом:

Выберите начальную стратегию $\pi_0(x,u)$. Начиная с $j = 0$, итерации по j до сходимости

Обновление ценности

$$V_{j+1}(x) = u\pi_{jx,u} x'Pxx'uRxx'u+\gamma V_j \pi x', \forall x \in S_j \in X \quad (25)$$

Улучшение стратегии

$$\pi_{j+1}(x,u) = \arg\min_{\pi} x'Pxx'uRxx'u+\gamma V_{j+1} \pi x', \forall x \in S_j \in X \quad (26)$$

В каждом цикле j алгоритм итерация по критерию оценивает значение функции $V_{j+1}(x)$, соответствующее стратегии управления $\pi_j(x,u)$, на основе значения функции ценности на предыдущем этапе . Обновление стратегии производится по формуле (26). Повторяйте шаги алгоритма до сходимости.

1.4.3 Сравнение итерации по стратегии и итерации по критерию

В то время как значение $V_j(x')$ в алгоритме итерации по стратегии представляет собой фактическое значение текущей стратегии π_j, в алгоритме

итерации по критерию его можно рассматривать как оценку стоимости перехода из состояния x в будущее состояние x'.

В алгоритме итерации по критерию в каждом цикле мы обновляем значение функции ценности $V(x)$, и когда $V(x)$ достигает оптимального значения, стратегия автоматически становится оптимальной. В отличие от итерации по критерию, итерация по стратегии пытается определить эффективность текущей стратегии с помощью оценки стратегии, а затем обновляет новую стратегию, после чего стратегия и значение функции ценности постоянно обновляются до тех пор, пока не будет достигнуто оптимальное значение.

Итерации по стратегии обычно сходятся быстрее, чем итерации по критерию. Поэтому во многих случаях он предпочтительнее. При определенных условиях гарантируется сходимость алгоритма итерации по стратегии за ограниченное число шагов. Однако количество шагов, необходимых для сходимости алгоритм итерации по критерию, не обязательно ограничено.

2. Регулятора для линейных систем

Рассмотрим непрерывную линейную и инвариантную систему во времени систему с моделью в пространстве состояний.

$$x = Ax(t) + Bu(t), \quad (2.1)$$

Где $B \in R^{n\times m}$, $A \in R^{n\times n}$, $x(t) \in R^{n}$ - состояния системы, $u(t) \in R^{n}$ – входной сигнал управления, пара матрицы (A, B) – стабилизируемость.

Задачи включается себе в том, что нужно разрабатывать стратегию управления с обратной связью $u = -Kx$ так, чтобы

$$\lim_{t\to\infty} x(t) = 0 \quad (2.2)$$

2.1 Оптимальный регулятор

Одним из наиболее известных типов оптимальных регуляторов является линейный квадратичный регулятор (LQR). Для системы (2.1), задача LQR с обратной связью по состоянию $u = -Kx$ состоит в поиске оптимального регулятора u^*, который минимизирует функцию стоимости $J(x,u)$ на бесконечном горизонте, связанную с системой:

$$J(x,u) = \int_0^\infty x^T(t)Qx(t) + u^T(t)Ru(t)dt \quad (2.3)$$

Здесь можно показать, что оптимальное управление является линейной функцией от x

$$u = -Kx = R^{-1} B^T P x \qquad (2.4)$$

где P — определенное положительное решение алгебраического уравнения Риккати

$$A^T P + PA - PB R^{-1} B^T P + Q = 0 \qquad (2.5)$$

Если ≥ 0, > 0, (A, B) стабилизируемость и (Q, A) обнаруживаемость, то алгебраическое уравнение Риккати имеет единственное симметричное положительное решение и соответствующий ему регулятор делает систему (3.1) асимптотически устойчивой.

2.2 Адаптивный регулятор

В случае адаптивного регулятора для системы (2.1) матрица A считается неизвестной. Полагая $A = A_0 + b\theta^T$, уравнение системы (2.1) принимает вид

$$\dot{x} = A_0 x + b\theta^T x + u \qquad (2.6)$$

Где θ - неизвестный параметр и A_0 – эталонная матрица Гурвица.

Сигнал управления имеет следующую формулу

$$u = -\theta^T x \qquad (2.7)$$

Где θ - оценка вектора , динамическая модель системы будет иметь вид

$$\dot{x} = A_0 x + b\theta^T x - b\theta^T x = A_0 x + b\theta^T x \qquad (2.8)$$

Где $\theta = \theta - \theta$ - вектор параметрических ошибок.

Алгоритм адаптивной стабилизации объекта (2.1) имеет вид

$$\dot{\theta} = \gamma x\, x^T Pb = \gamma x\, b^T Px \qquad (2.9)$$

Где симметрическая положительно определенная матрица P является решением уравнения Ляпунова

$$A_0^T P + P A_0 + Q = 0 \,,\; Q = Q^T > 0 \qquad (2.10)$$

2.3 Оптимальный адаптивный регулятор

Для системы (2.1) задача включается в том, что, нужно найти оптимальный регулятор u^* , который минимизирует квадратичную функцию стоимости бесконечного горизонта, связанную с системой.

$$V\ xt0, t0 = t0\infty xT(\tau)Qx\tau+uT(\tau)Ru\tau d\tau \qquad (2.11)$$

Где для системы (2.1) матрица A неизвестна, $Q \geq 0$, $R \geq 0$ и Q, A – обнаруживаемость.

Тогда оптимальный регулятор записывается следующим образом:

$$u^* t = \text{argminu}(t)Vt0,xt0, ut\ ,\ t0 \leq t \leq \infty \qquad (2.12)$$

Решение этой задачи оптимального управления, определенное в соответствии с принципом оптимизации Беллмана, дается с помощью обратной связи по состоянию $u\ t = - Kx = - R\text{-}1\ BT\ Px$.

Пусть K - стабилизирующий коэффициент обратной связи по состоянию для (2.1) такой, что $x = A\text{-}BK$ представляет собой устойчивую замкнутую систему. Тогда квадратичная бесконечная стоимость или ценность на бесконечном горизонте определяется выражением

$$V\ xt = 0\infty(xT\tau Qx\tau+uT\tau Ru)\ d(\tau)\ ,\ u= -Kx$$

$$V(x\ t\ )= 0\infty xT\tau Q+KTRKxT\tau d(\tau) = xT\ \tau\ Px(t) \qquad (2.13)$$

Где P - симметричная положительно определенная матрица, является решением матричного уравнения Ляпунова

$$A\text{-}BKT\ P+P\ A\text{-}BK = -(\ KT\ RK+Q) \qquad (2.14)$$

Тогда $V(x(t))$ служит функцией Ляпунова для (2.1) с коэффициентом усиления регулятора K. Функция ценности (2.8) можно записать в следующем виде

$$V\ xt = tt+TxT\tau Q+kTRKx\tau d\tau + V\ (\ x\ t+T\ ) \qquad (2.15)$$

Это уравнение Беллмана для задачи LQR. Используя уравнение Беллмана IRL, можно обойти отмеченные в разделе (1.3.2) проблемы применения обучения с подкреплением к системам с непрерывным временем.

Интегральное армирование

$$\rho\ xt,t,T = tt+TxT\tau Q+KTRKx\tau d(\tau) \qquad (2.16)$$

Формула (2.15) называется уравнением Беллмана обучения с интегральным подкреплением (IRL).

Далее показано, как найти оптимальный регулятор, применяя интегральное обучение с подкреплением к линейным системам (итерации по стратегии) (2.1).

$$W_k^T[\phi(x(t)) - \phi(x(t+T))] = \int_t^{t+T} x^T(\tau)(Q + K_k^T R K_k)\,\mathrm{d}\tau.$$

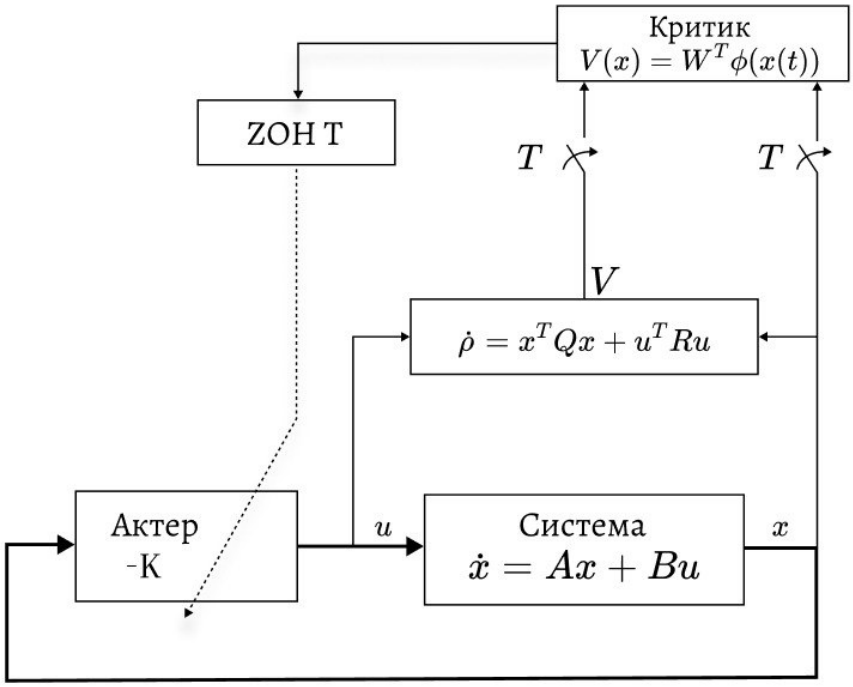


Рисунок 2.1 Структура системы с оптимальным адаптивным регулятором для линейных систем.

На основе интегрального обучения с подкреплением (2.15) можно написать следующий алгоритм RL:

$$V\ x(t) = \int_{t}^{t+T} x^T Q + K_i^T R K_i x\,d(\tau) + x_{t+T}^T\ P\ x_{t+T} \qquad (2.16)$$

Параметры функции ценности $V_i\ x_t = x^T\ t\ P x\ (\ t\ )$ на итерации i IRL являются элементами симметричной ядерной матрицы P_i . Они должны быть найдены на каждой итерации i путем измерения данных (x(t), x(t+ T) , u(t)) в

моменты времени t вдоль траекторий системы. Для вычисления этих параметров термин xT t Pi x(t) записывается как

$$xT\ t\ Px\ t = WT\ \varphi\ (\ x\ t\ ) \qquad (2.17)$$

Где φ xt – вектор функции активации. φ xt обозначает базисный вектор квадратичного полинома произведения Кронекера

$$\varphi\ xt = x12x1x2...x1xnx22x2x3...xn2T \qquad (2.18)$$

Применяя формулу аппроксимации значения функции (2.17), можно записать (2.15) следующим образом

$$Wj+1\ (\varphi\ xt - \varphi\ x(t+T\ )) = tt+TxtTQ+KiTRKixt\ dt = d(\varphi\ xt, Ki\ , \qquad (2.17)$$

Применяя к этой проблеме структуру «Актер-Критик», Критик использует уравнение (2.17) для вычисления матрицы Pi . Когда Критик сходится к решению Pi , Актер рассчитает новое значение управления для управления системой на основе Pi по формуле:

$$ui+1 = -\ R\text{-}1\ BT\ Pi\ x = Ki+1\ x \qquad (2.16)$$

С помощью использования алгоритма итерации по стратегии мы можем найти оптимальный регулятор без необходимости знания матрицы для линейной системы.

3. Задача оптимального управления для нелинейных систем

К непрерывным системам обучение с подкреплением применить сложнее, чем к дискретным, и оно дает меньше результатов. Один из методов, описанных в [13], [14], называется обучением с интегральным подкреплением. Этот метод может быть использован для разработки методов оптимального управления непрерывными системами, не требуя полного понимания динамики системы. Этот метод называется методом оптимального адаптивного управления. Ниже изучается и рассматривается оптимальный адаптивный регулятор для нелинейных систем.

3.1 Оптимальное управление и уравнение Гамильтона Якоби – Беллмана

Рассмотрим нелинейную динамическую систему с непрерывным временем

$$x = f\ x + g\ x\ u \qquad (3.1)$$

Где $x\ t \in Rn$ – состояние системы, входное управление $u\ t \in Rm$, точка равновесии $x=0$, $f\ 0 =0$ и $f\ x + g\ x\ u$ Липшиц (Lipschitz) на съемочной площадке $\Omega \in Rn$. Предположим, что система стабилизируема на Ω, т. е. существует непрерывная функция управления $u(t)$ такая, что замкнутая система асимптотически устойчива на Ω.

Определите меру эффективности или функцию ценности, которая имеет значение, связанное со стратегией управления с обратной связью $u= \mu(x)$, заданное как

$$V\mu\ xt = t\infty rx\tau, u\tau d\tau \qquad (3.2)$$

Где $r\ x,u = Q\ x + u\tau\ Ru$, положительно определенный $Q(x)$, т. е. $Q\ x >0$ для всех x и $x=0 => Q\ x =0$ и положительно определенная матрица $R= RT \in Rm\times m$.

Для LQR с непрерывным временем приведенные выше выражения имеют вид как показан в пункте 2.

$$x = Ax + Bu \qquad (3.3)$$

$$V\mu\ xt = 12\ t\infty(xTQx+uTRu)d\tau\ . \qquad (3.4)$$

Определение 1. Стратегия управления $\mu(x)$ определяется как допустимая относительно (3.2) на Ω, обозначаемая $u \in \Psi(\Omega)$, если $\mu(x)$ непрерывна на Ω, $u\ 0 =0$, $\mu(x)$ стабилизирует (3.1) на Ω и $V(x0)$ конечно $\forall\ x0 \in \Omega$.

Для любой допустимой стратегии управления $\mu\in\Psi(\Omega)$, если соответствующая функция стоимости $V(x0)$ равна $c1$, то бесконечно малым эквивалентом (3.2) является уравнение Беллмана

$$0 = r (x , \mu (x)) + \nabla V\mu T (f (x)+ g (x) \mu (x)),\ V\mu (0) = 0 \qquad (3.5)$$

Где $\nabla\ V\mu$, взятый здесь как вектор-столбец, обозначает вектор градиента функции ценности $V\mu$ по x. Учитывая допустимый регулятор $\mu(x) \in \Psi(\Omega)$, уравнение Беллмана можно решить через соответствующую функцию ценности $V\mu\ (x)$. В линейном случае оно становится уравнением Ляпунова. Учитывая, что $\mu(x)$ является допустимой стратегии управления, если $V\mu\ (x)$ удовлетворяет

(3.5) с $r(x, \mu(x)) \geq 0$, то можно доказать, что $V_\mu(x)$ является функцией Ляпунова для системы (1) со стратегией управления $\mu(x)$.

Теперь можно сформулировать задачу оптимального управления: для непрерывной системы (3.1), множества $u \in \Psi(\Omega)$ допустимых стратегий управления и функции ценности на бесконечном горизонте (3.2) найти допустимую стратегию управления такую, что индекс ценности (3.2), связанная с системой (3.1), минимизируется.

Определите гамильтониан

$$H(x, \mu x, \nabla V_\mu) = r(x, \mu x) + \nabla V_\mu^T (f(x) + g(x)u) \qquad (3.6)$$

Тогда оптимальная функция ценности $V^*(x)$ удовлетворяет уравнению HJB

$$0 = \min_\mu H(x, \mu x, \nabla V^*) \qquad (3.7)$$

и удовлетворяющие оптимальному управлению

$$\mu^* = \arg\min_\mu H(x, \mu x, \nabla V^*) \qquad (3.8)$$

$$\mu^* = -\frac{1}{2} R^{-1} g^T(x) \nabla V^*(x) \qquad (3.9)$$

Подставляя эту стратегию оптимального управления в гамильтониан, мы получаем формулировку уравнения HJB в терминах $V^*(x)$:

$$0 = Q(x) + \nabla V^{*T}(x) f(x) - \frac{1}{4} \nabla V^{*T}(x) g(x) R^{-1} g^T(x) \nabla V^*(x), \qquad (3.10)$$

$$V^*(0) = 0$$

Это достаточное условие функции оптимального значения [10]. В случае линейной системы с квадратичной функцией стоимости это уравнение HJB становится уравнением Риккати.

Чтобы найти оптимальное управляющее решение задачи, нужно всего лишь решить HJB (3.10) для функции ценности, а затем подставить решение в (3.9) для получения оптимального управления. Однако решение уравнения HJB, как правило, затруднено, поскольку оно представляет собой нелинейное дифференциальное уравнение второго порядка, следующее за градиентом функции стоимости, и его решение также требует полного знания системы системной динамики (т. е. динамики системы, описываемой уравнением

функции f(x) и g(x) должны быть известны). Более того, глобально гладких решений может не существовать.

3.2 Адаптивный алгоритм оптимального управления на основе итераций по стратегии (PI).

Итерация по стратегии (PI) — это итерационный метод обучения с подкреплением [4, 12] для решения (3.10) на основе простых уравнений. PI состоит из двух этапов: улучшение стратегии на основе (3.9) и оценка стратегии на основе Беллмана (3.5). Говорят, что PI имеет структуру обучения с подкреплением «Актер-Критик», где улучшение стратегии выполняется актером, а оценка стратегии выполняется критиком, решая уравнение Беллмана. Алгоритм итерация по стратегии задается следующим образом.

Пусть u(x) — допустимая стратегия для (3.1), такая что замкнутая система асимптотически устойчива на Ω. Тогда стоимость бесконечного горизонта для любого x t ∈ Ω определяется формулой (3.2), а V(x(t)) служит функцией Ляпунова для (3.1). Функцию стоимости (3.2) можно записать в виде

$$V\mu\ xt = tt+Tr(x\tau, \mu x\tau d\tau+V\mu xt+T \qquad (3.11)$$

На основании (3.11) и (3.7), учитывая начальную допустимую стратегию управления u0 (x) , можно вывести следующую схему итерации по стратегии

1. Решить для Vμi (x) с использованием

$$V\mu i\ xt = tt+Trx\tau, uix\tau d\tau+V\mu(xt+T \qquad (3.12)$$

2. Обновить стратегию управления с помощью

$$\mu i+1\ x = argmin\mu\{H(x, \mu, Vxui\} \qquad (3.13)$$

Где

$$\mu i+1\ x = -\ 12\ R\text{-}1\ gT\ (\ x\ )\ \nabla Vxui \qquad (3.14)$$

Уравнения (3.13) и (3.14) формулируют новый алгоритм итерации по стратегии для поиска оптимального управления без использования каких-либо знаний о внутренней динамике системы f(x) .

3.3 Онлайн реализация алгоритма итерации по стратегии

Применяя аппроксимационные возможности параллельной нейронной сети для уменьшения объема вычислений по сравнению со структурой «Актор-критик», этот алгоритм использует нейронную сеть для аппроксимации

оптимальной функции стоимости V (x) с помощью x ∈ Ω следующим образом:

$$V\,x = W^T \varphi\,(\,x\,) \qquad (3.15)$$

Где $W^T \in R^N$ – неизвестная веса, N - количество нейронов, φ(x) – вектор функции активации.

Используя нейронную сеть для аппроксимации оптимальной функции ценности, подставьте формулу (3.15) в формулу (3.12), чтобы получить:

$$W^T \varphi\, x(t) = \int_{t}^{t+T} r(x_\tau, u_i x_\tau d\tau + W^T \varphi(x_{t+T}) \qquad (3.16)$$

Появляется ошибка (), которая является ошибкой аппроксимации функции Беллмана.

$$e\,x_{t,T} = W^T \varphi x_{t+T} - \varphi x_t = -\int_{t}^{t+T} r(x_\tau, u_i x_\tau d\tau \qquad (3.17)$$

Обозначим что

$$h\,t = \varphi\,x_{t+T} - \varphi\,x_t \qquad (3.18)$$

$$y\,t = \int_{t}^{t+T} r(x_\tau, u_i x_\tau d\tau \qquad (3.19)$$

Уравнение (3.16) можно написать в следующий вид

$$e\,t = W^T h\,t + y\,(\,t\,) \qquad (3.20)$$

Где $H=[\,h_{t1},\ldots,h_{tN}\,]$, $Y = y_{t1},\ldots,y_{tNT}$ (3.21)

Уравнение (3.20) является линейной функцией по параметру W . Следовательно, мы можем применить алгоритм наименьших квадратов ошибки, чтобы найти оптимальное значение для W .

Данные системы собирается из N различных выборок за период времени T, поэтому мы вычисляем (3.19) в n точках от $t1 \rightarrow tN$, чтобы получить следующие функции:

$$H= h_{t1},\ldots,h_{tN} \qquad (3.22)$$

$$Y= y_{t1},\ldots,y_{tN} \qquad (3.23)$$

Чтобы определить веса W нейронной сети, аппроксимирующей функцию V, что приводит к минимизации следующей целевой функции

$$S= \Omega ex,Te(x,T)\partial x \quad (3.24)$$

Произведение $< f , g > = \Omega fg\partial x$ интеграла Лебега можно записать:

$$< dex,TdW , e\ x,T > \Omega =0 \quad (3.25)$$

Использовать уравнение (3.20) для уравнения (3.25), получили

$$H\ HW+Y =0 \quad (3.26)$$

$$W = -\ HHT-1\ HY \quad (3.27)$$

Алгоритм онлайн обучения с интегральным подкреплением использует нейронные сети представлен следующим образом

Шаг 1. $\forall\ x \in \Omega x$, инициализировать допустимый закон управления $u (x) \in \psi (\Omega)$. Поместите в систему управляющий сигнал $u0$ и соберите необходимую данные системы о состоянии и управляющих сигналах на N различных выборках за период времени T. Назначьте $i \leftarrow 0$, инициализируйте εw

Шаг 2. Используйте данные, собранную о системе, для расчета H и Y. Определить W из уравнения (3.27)

Шаг 3: обновите закон управления для следующего цикла

$$ui+1\ x = -\ 12\ R-1\ g\ xT\ \nabla\ Vx\ Wi \quad (3.28)$$

Если критерий сходимости удовлетворяется так, что $||\ Wi+1 - Wi\ ||< \varepsilon w$, алгоритм завершается. Если не удовлетворено, присвойте $i \leftarrow i+1$, подайте сигнал ui в систему и соберите необходимую информацию системы о состоянии и управляющих сигналах в N различных выборках за период T, затем вернитесь к шагу 2.

4. Результаты моделирование

В этом разделе для применения и исследования алгоритмов, представленных ранее в разделах 2 и 3, была выбрана система перевернутого маятника и тележки [17][18].

4.1 Регулятор для линейных систем

Рассмотрим непрерывную инвариантную во времени линейная система

$$x =Ax+Bu, \quad (4.1)$$

$A = \begin{matrix} 0 & 1 \\ -10 & 0 \end{matrix}$, $B = \begin{matrix} 1 \\ 1 \end{matrix}$

Где $x \in R^2$ – состояние системы, $u \in R^1$ – сигнал управления системы

Задача оптимального управления состоит в оптимизации следующей функции

$$J(x,u) = \int_0^\infty x^T(\tau) Q x(\tau) + u^T(\tau) R u(\tau) d\tau \qquad (4.2)$$

$Q = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix}$, $R = 1$

В случае оптимального регулятора (LQR), решая уравнение Риккати

$$A^T P + PA - PB R^{-1} B^T P + Q = 0$$

Мы получали матрицу $P = \begin{matrix} 2.0739 & 0.2110 \\ 0.2110 & 0.0672 \end{matrix}$, следовательно обратная матрица K вычислена

$K = R^{-1} B^T P = \begin{matrix} 2.2840 & 0.2783 \end{matrix}$

Параметры для адаптивного регулятора

$\gamma = 1$, $A_0 = A - BK = \begin{matrix} -2.2840 & 0.7217 \\ -7.7151 & -0.2783 \end{matrix}$, $P_a = P = \begin{matrix} 2.0739 & 0.2110 \\ 0.2110 & 0.0672 \end{matrix}$

Далее представлены результаты моделирования

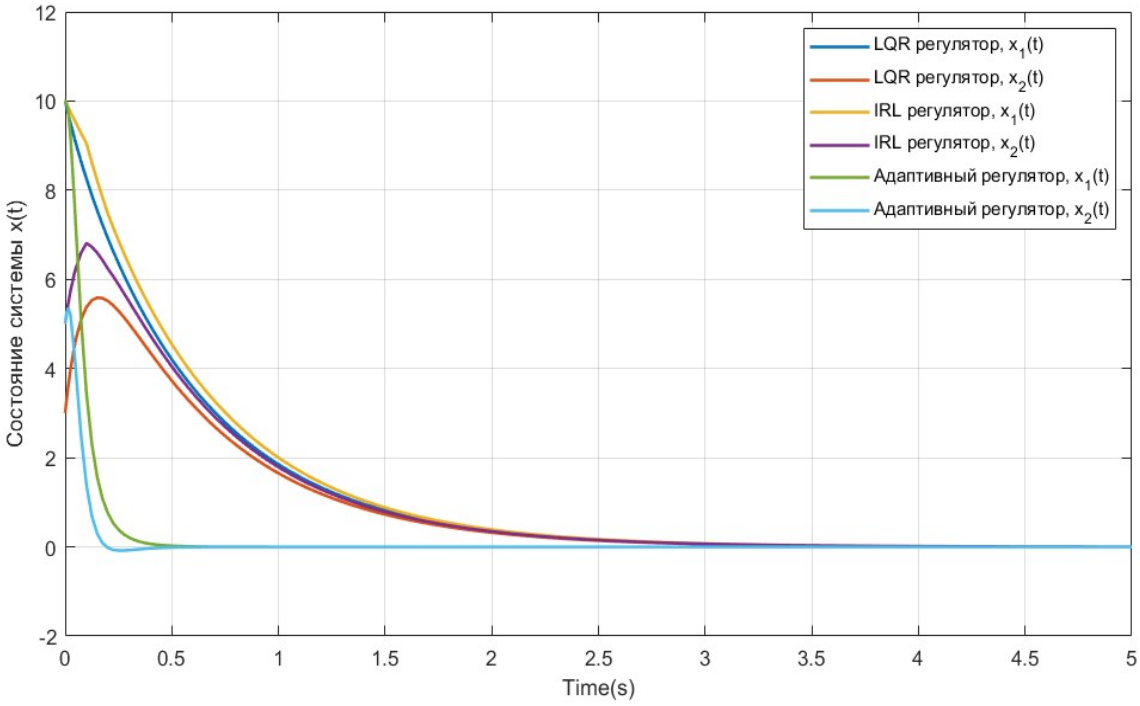


Рисунок 4.1 - Графика вектор состояния x(t) для оптимального регулятора, адаптивный регулятор и оптимального адаптивного регулятора.

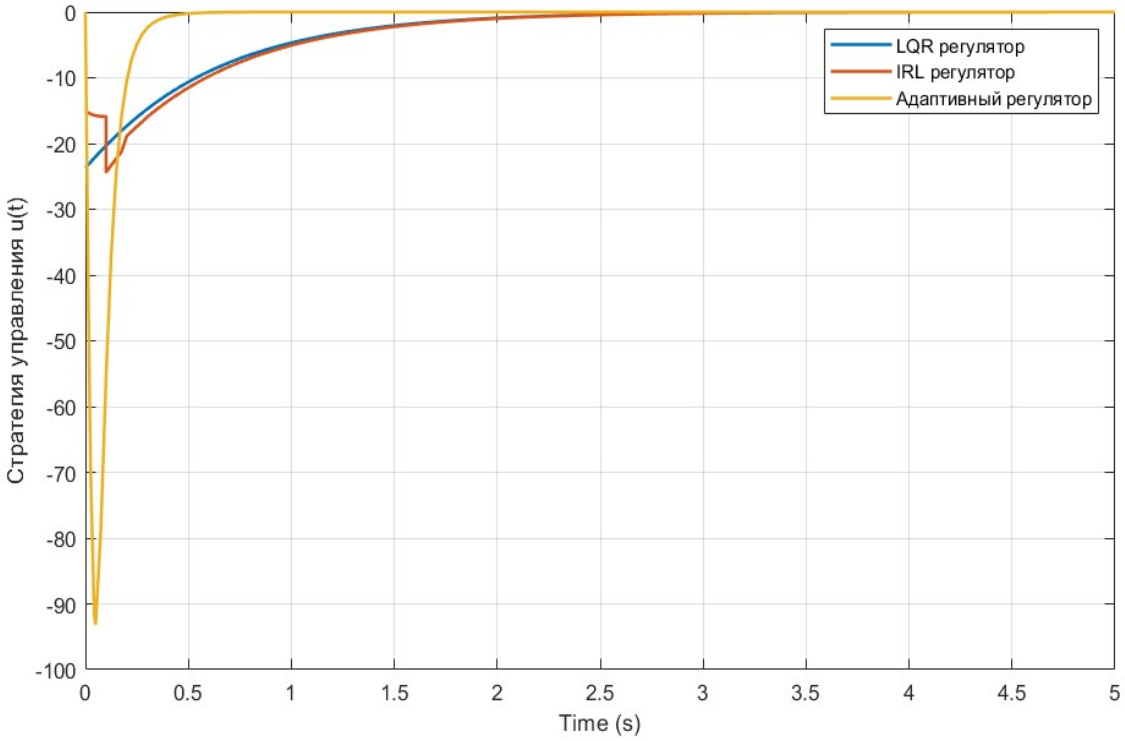


Рисунок 4.2 - Графика сигнал управления для оптимального регулятора, адаптивный регулятор и оптимального адаптивного регулятора (IRL)

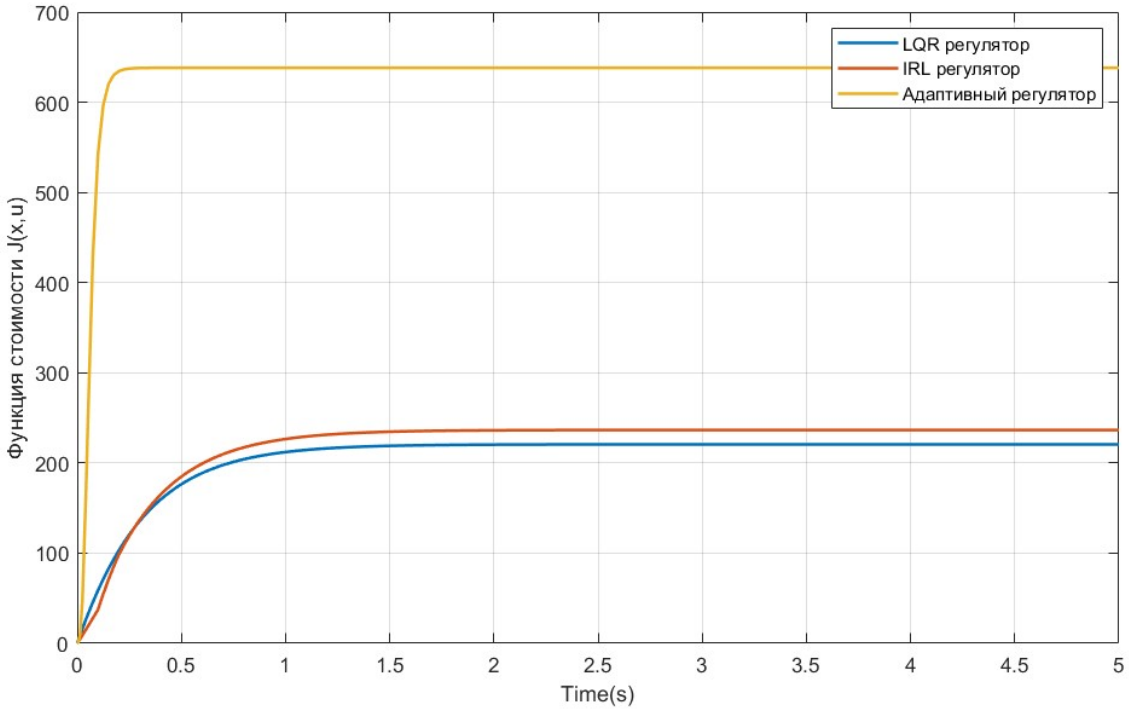


Рисунок 4.3 - Значение функции стоимости для оптимального регулятора, адаптивный регулятор и оптимального адаптивного регулятора (IRL)

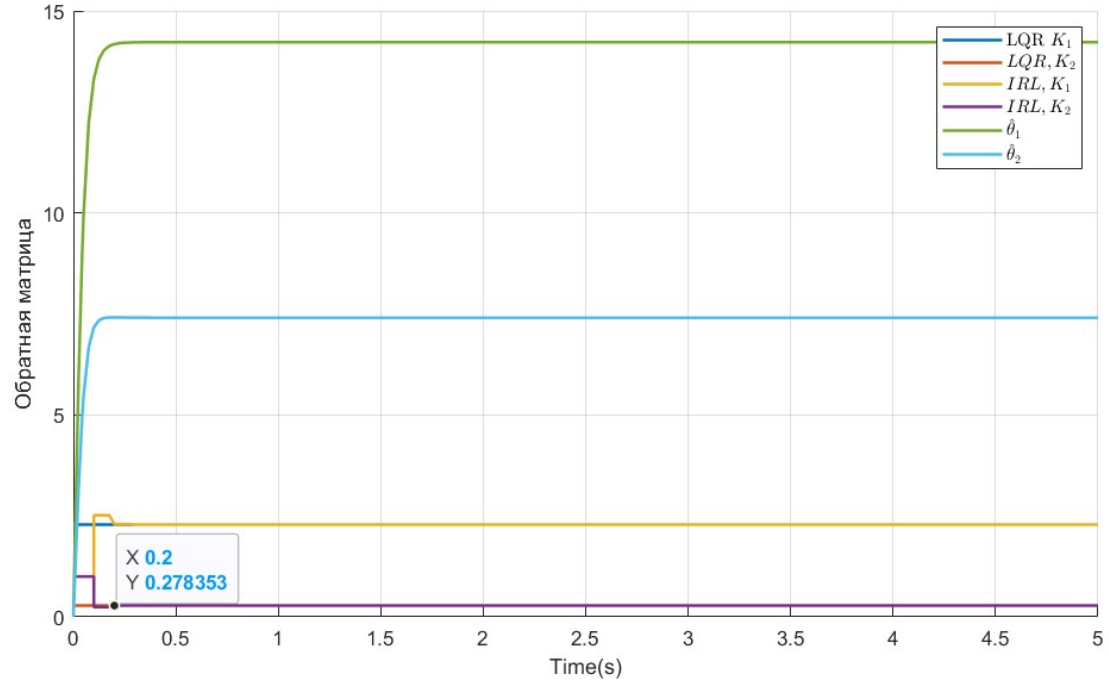


Рисунок 4.4 – Графика матрицы регулятора `K` (LQR), параметры `θ` для адаптивного регулятора и сходимости значения обратной матрицы оптимального регулятора (IRL)

Приведенные выше результаты моделирования показывают, что с помощью алгоритма итерации по стратегии обучения с подкреплением оптимальный адаптивный регулятор сошелся к асимптоте оптимального управления через 0.2 секунды. При этом адаптивный регулятор не сходится к оптимальному управлению.

Кроме того, производительность адаптивного регулятора зависит от выбора параметров γ и Pa для системы. Поэтому при изменении значений матрицы Q и R нам приходится заново выбирать параметры оптимального регулятора. Для оптимального адаптивного регулятора нам не нужно настраивать параметры при изменении Q и R.

4.2 Исследование влияния времени выборки T на оптимальный адаптивный регулятор.

При применении алгоритма обучения с интегральным подкреплением к оптимальному адаптивному регулятору нам нужно точно аппроксимировать оптимальную матрицу P, используя данные по траектории системы на интервалах времени T. Графики влияния времени выборки T к производимости оптимального адаптивного регулятора представлены ниже.

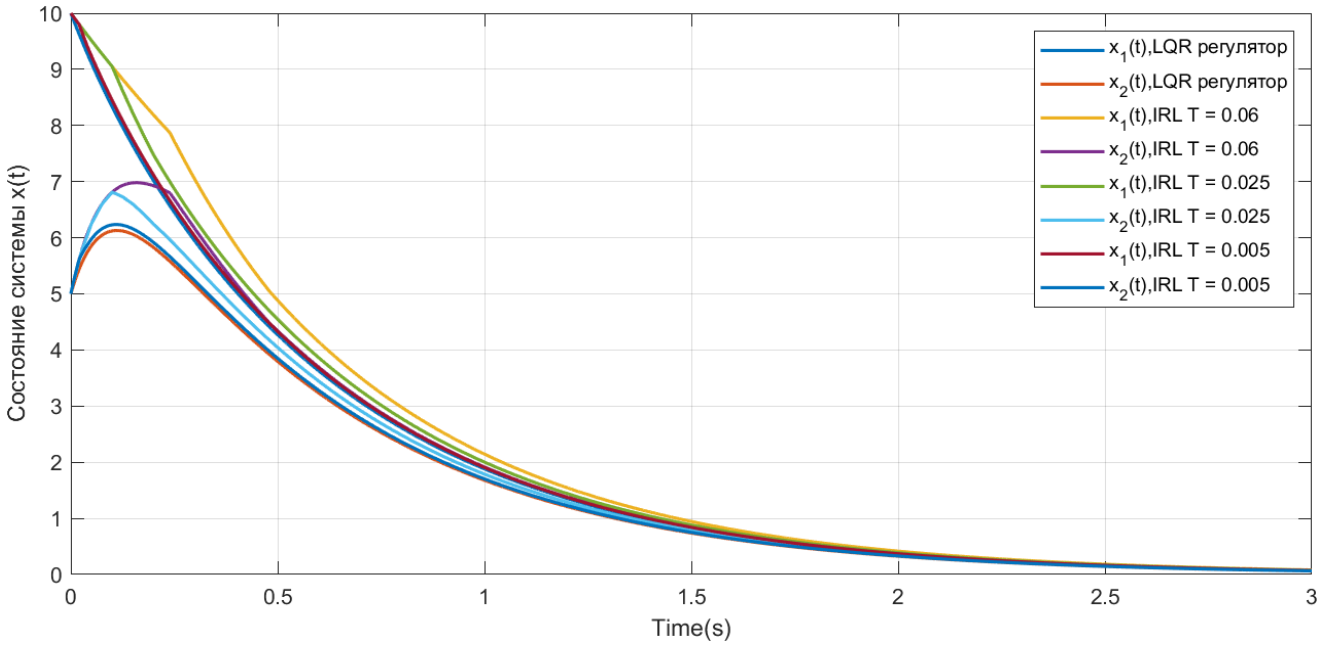


Рисунок 4.5 – Графика состояния системы $x\ t$ оптимального регулятора и оптимального адаптивного регулятора при некоторых значениях T

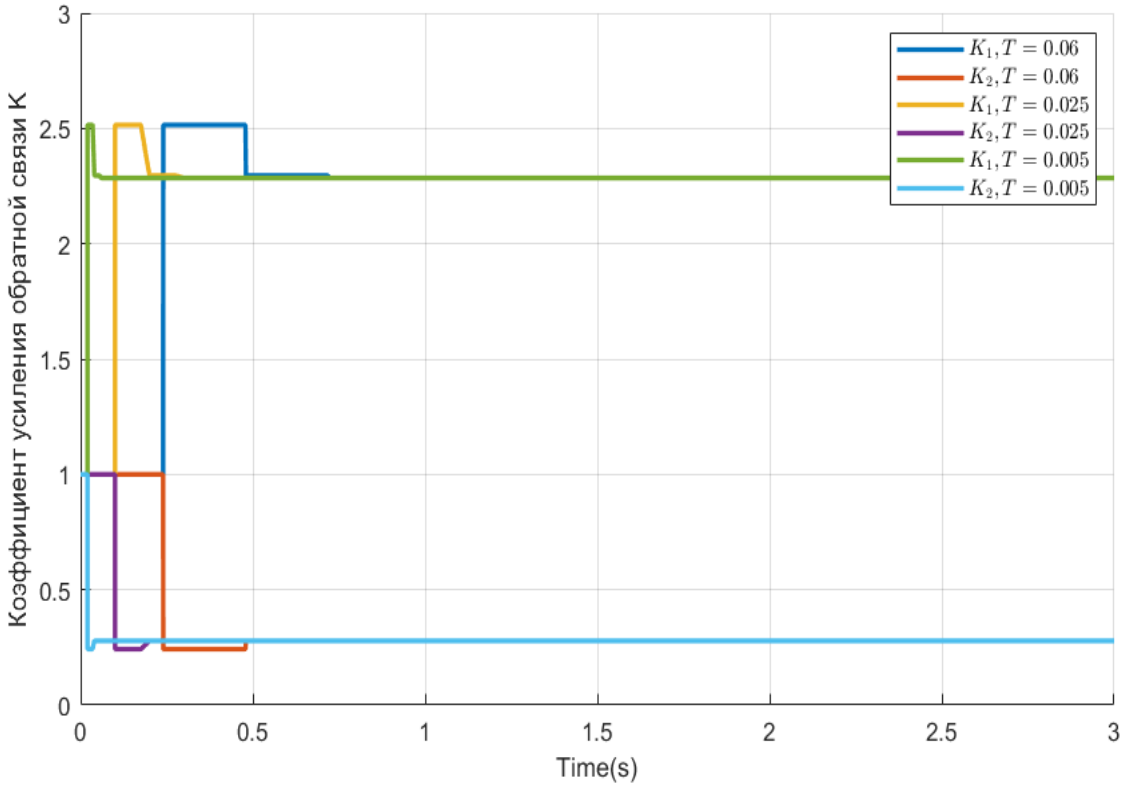


Рисунок 4.6 – Графики коэффициента усиления обратной связи K оптимального адаптивного регулятора для разных значений T

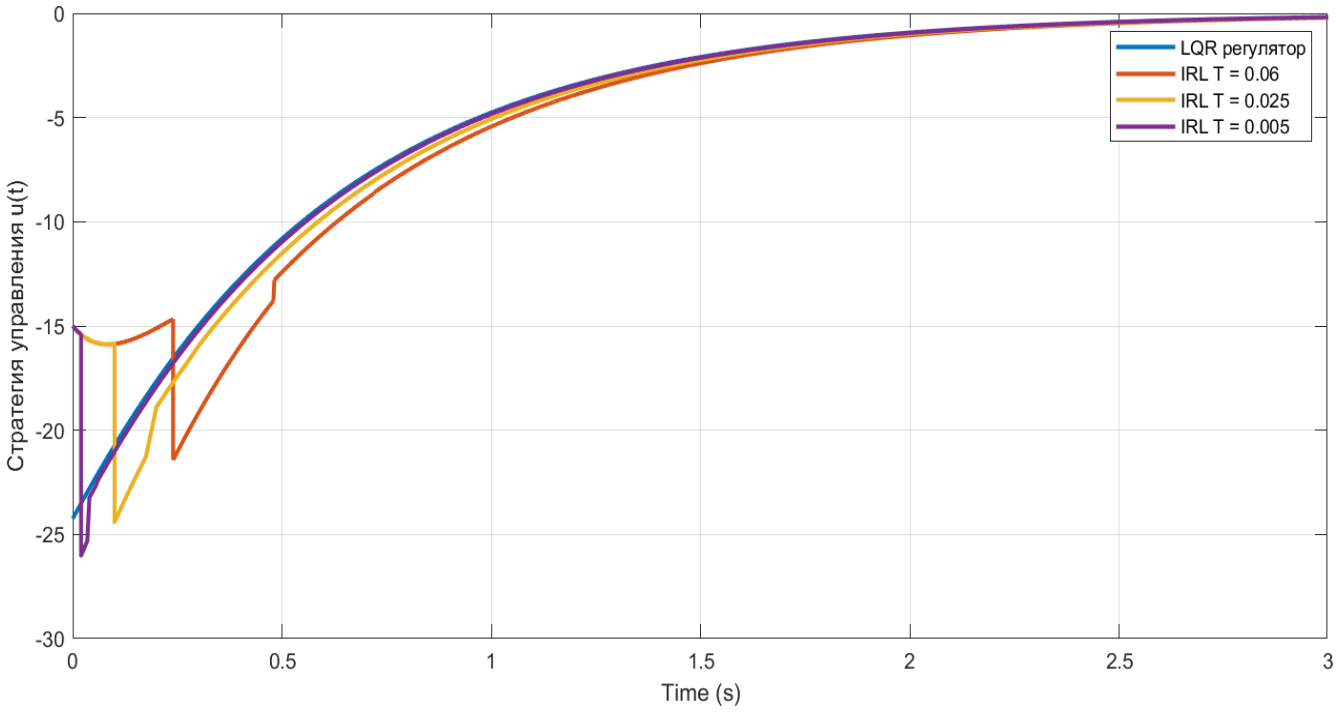


Рисунок 4.7 – График, представляющий управляющий сигнал оптимального регулятора LQR и оптимального адаптивного регулятора с различными значениями T .

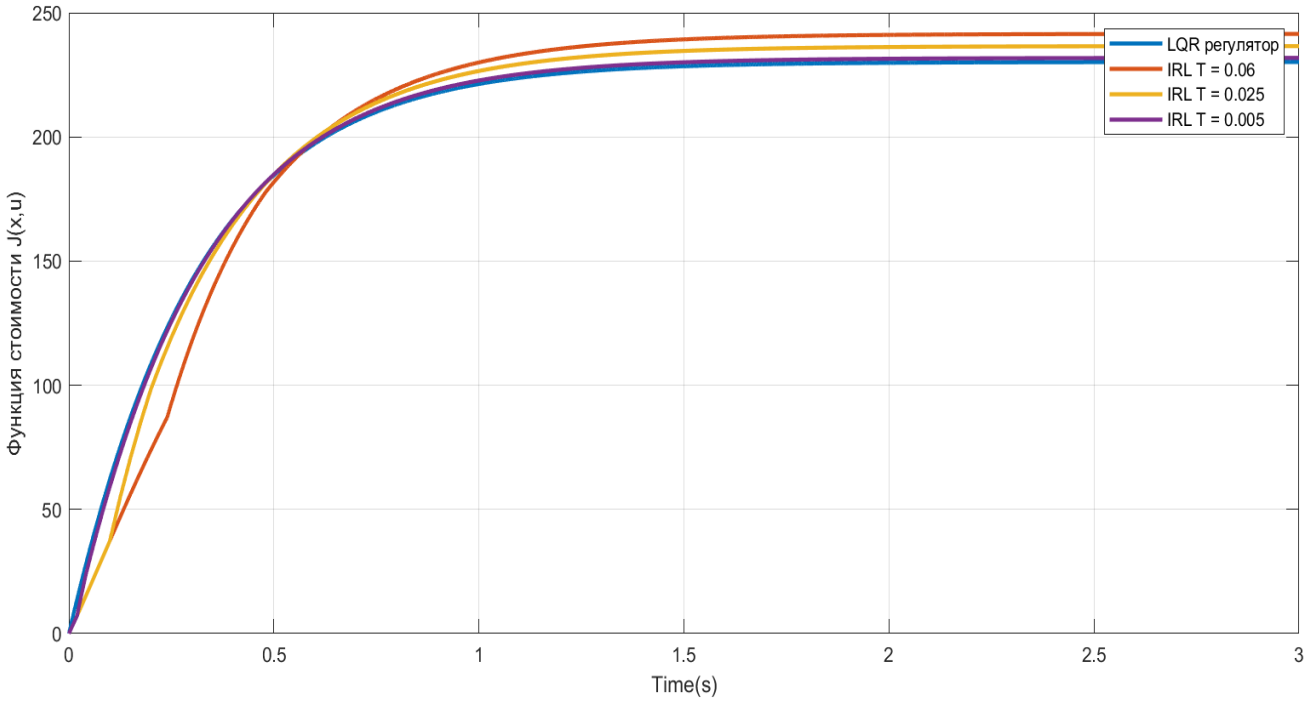


Рисунок 4.8 - График, отражающий значение функции стоимости для оптимального регулятора и оптимального адаптивного регулятора с различными значениями T .

Поскольку алгоритм использует данные, полученные из траектории системы, система не будет устойчивой, когда состояние системы x(t) равно примерно 0, и в этот момент алгоритм будет работать неправильно, что приведет к тому, что управление перестанет быть оптимальным. Поэтому мы должны выключить алгоритм, когда стратегия управления приблизится к оптимальному значению. В этом случае нам необходимо отключить алгоритм, когда коэффициент матрицы обратной связи K приблизится к оптимальному значению. Еще один случай, на который следует обратить внимание, заключается в том, что мы применяем управляющий сигнал u0 , чтобы сделать систему устойчивой. Поэтому может случиться так, что система сойдётся до того, как будет найдено оптимальное значение управления. Поэтому нам нужно выбрать подходящее время выборки T.

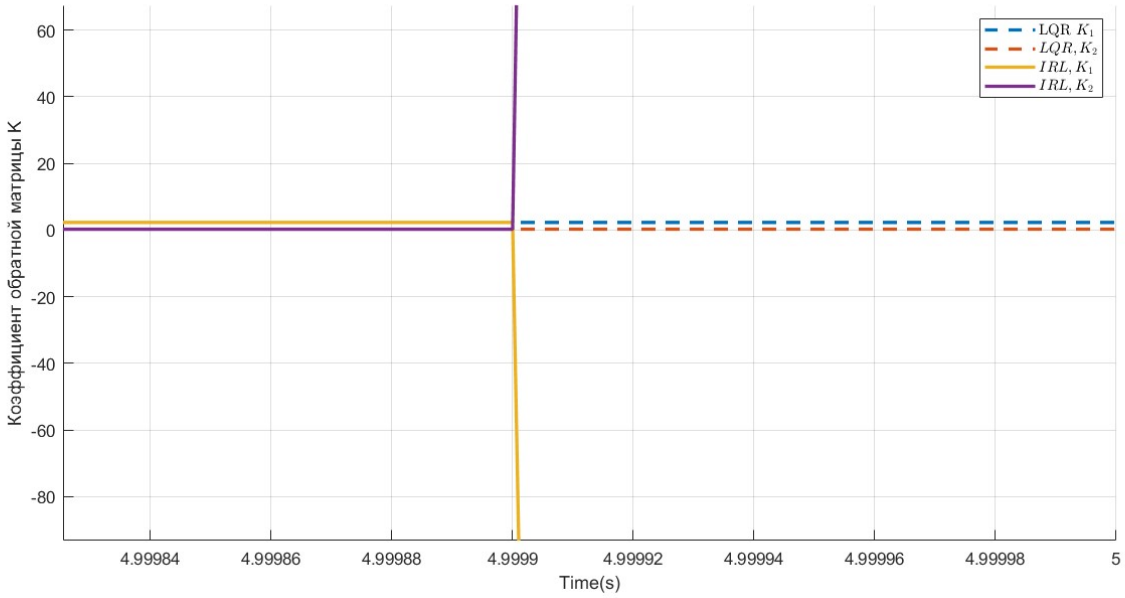


Рисунок 4.9 - Графики коэффициента усиления матрицы обратной связи K оптимального регулятора и оптимального адаптивного регулятора при T=0.25

На графике выше мы видим, что значение коэффициента матрицы обратной связи K приблизилось к оптимальному значению, как и в случае использования LQR. Но когда состояние системы сходится к нулю, алгоритм не отключается, в результате чего значение коэффициента матрицы обратной связи K быстро расходится.

При применении обучении с интегральным подкреплением и приближенного динамического программирования для линейных систем мы часто используем квадратичную функцию стоимости. Пусть n - размер состояния системы, x ∈ Rn . Тогда объем данных, который необходимо собрать для вычисления новой стратегии управления, равен nn+12 . Таким образом, для квадратичной системы (n = 2) необходимо собрать не менее 3 точек данных, а для квадратичной системы (n = 4) необходимо собрать не менее 10 точек данных, чтобы аппроксимировать параметры матрицы P . То есть, собрав необходимое количество точек данных, алгоритм будет аппроксимировать параметры матрицы P. Алгоритм продолжается до тех пор, пока значение P не сойдется с оптимальным значением.

4.3 Регулятор для нелинейных систем

Рассмотрим нелинейных систем

$$x = -x1+x2-0.5x1-0.5x2(1-\cos 2x1+22) + 0\cos 2x1+2\ u \qquad (4.3)$$

Функция ценности:

$$J = \int_0^\infty x^T Q x + u^T R u dt \quad (4.4)$$

$$Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, R = 1$$

Решить уравнение HJB мы получили функцию Беллмана и сигнал оптимального управления

$$V^*(x) = w_1^* x_1^2 + 2 w_{12}^* x_1 x_2 + w_2^* x_2^2 = 0.5 x_1^2 + x_2^2 \quad (4.5)$$

$$u^*(x) = -\cos(2x_1 + 2) x_2 \quad (4.6)$$

Функция вектора активации выбрана

$$\varphi(x) = [x_1^2 \; x_1 x_2 \; x_2^2] \quad (4.7)$$

Инициализировать начальный вес $W = [-2 \; 0 \; 5]^T$ и $x_0 = [1 \; 2]^T$

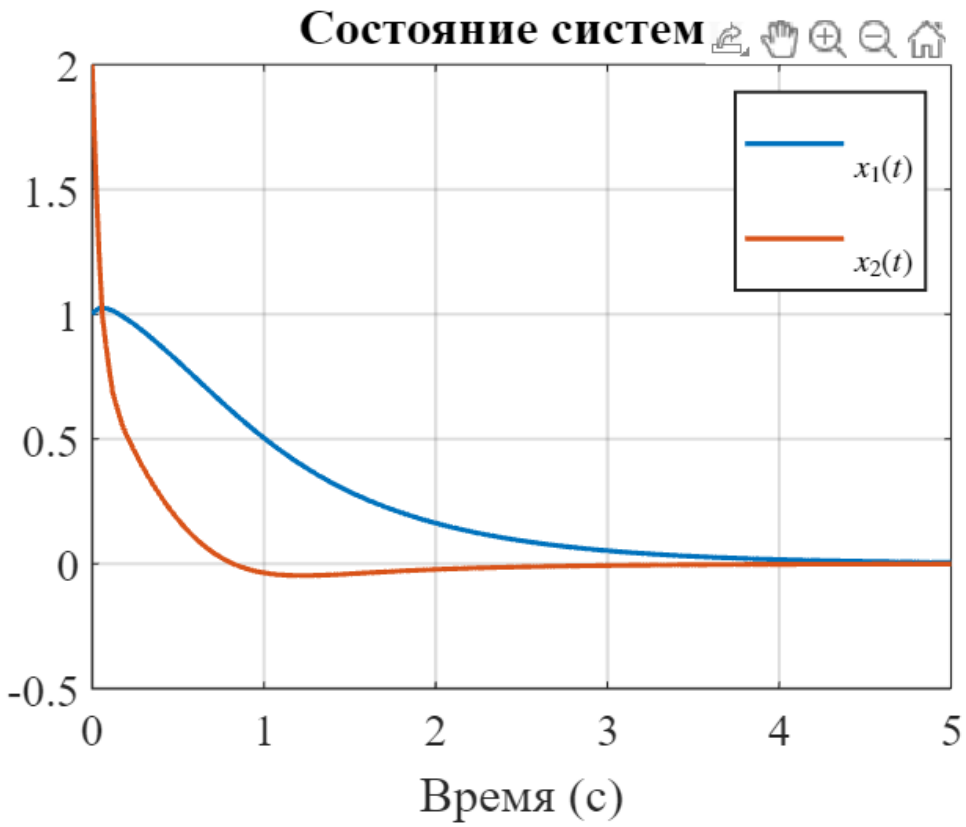


Рисунок 4.10 – График состояния системы с использованием онлайна обучения с интегральным подкреплением

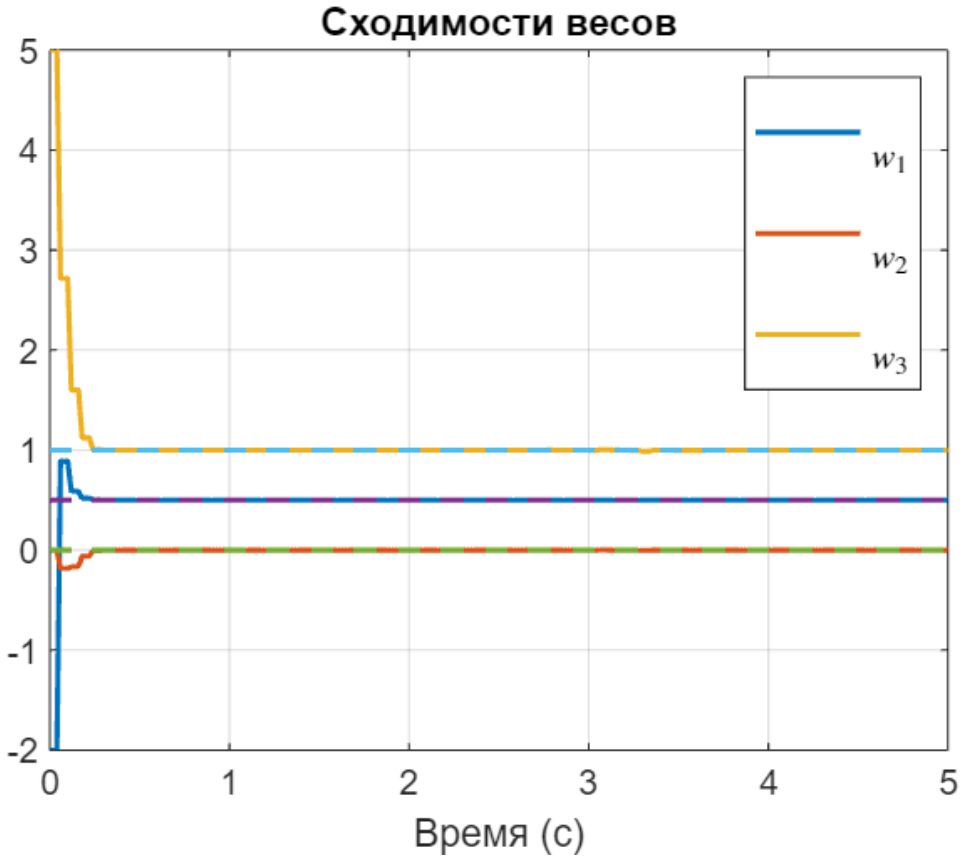


Рисунок 4.11 Сходимость весов W

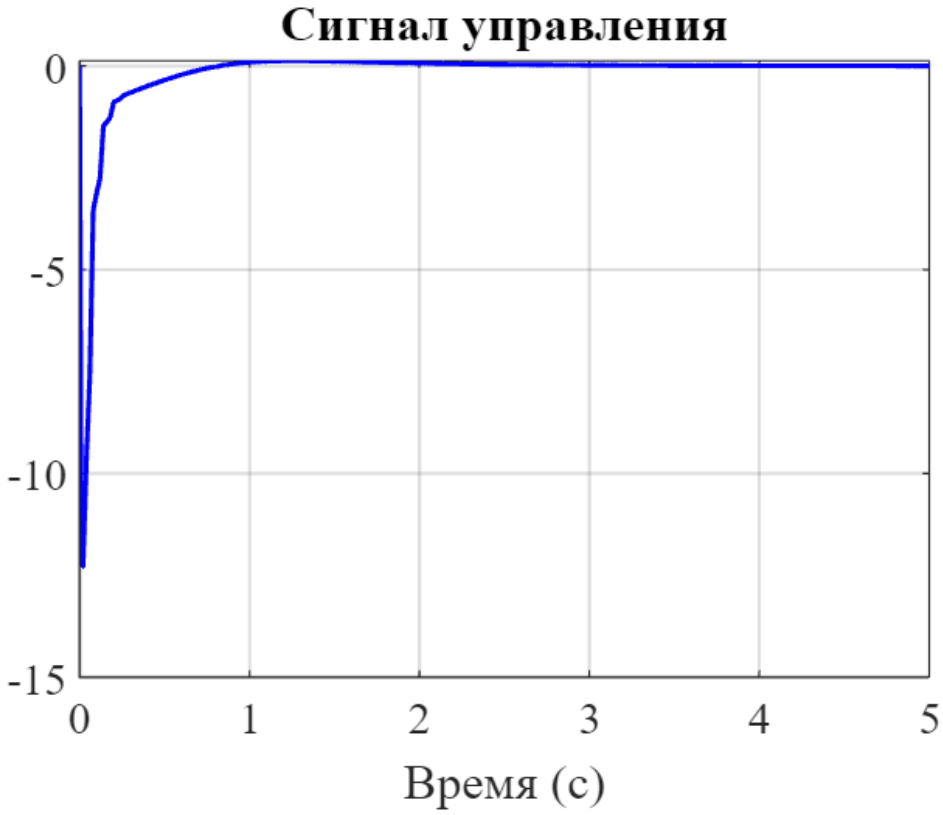


Рисунок 4.12 – Сигнал управления системой с алгоритмом обучения с подкреплением

Как видно из графика, вес W сходится точно к своему оптимальному значению через 0.24 секунды, в то время как сигнал управления по-прежнему помогает системе стабилизироваться на достаточно хорошей скорости. Тем самым показывая корректность алгоритма.

4.4 Регулятор линейной системы для перевернутого маятника и тележки

Данная система

$$x = Ax + Bu ,$$

$$A= 0100M+mgML0000001-mgM000 , B= 0-1ML01M$$

Где $x= \theta\theta xx$ – состояние системы, $u \in R1$ – сигнал управления системы

Параметры для моделирования из [18]

$M =2.4$ кг, $m=0.23$ кг, $g=9.81$ мс2 , $L=0.46$ м

Задача оптимального управления состоит в оптимизации следующей функции

$$J\ x,u = 0\infty xTtQxt+uTtRutdt$$

$$Q = 10.1000.11000010.1000.11 , R=1$$

Матрица оптимального регулятора (LQR)

$K = R\text{-}1\ BT\ P = -62.7\text{-}13.1\text{-}1.0\text{-}2.92$

Результаты моделирования

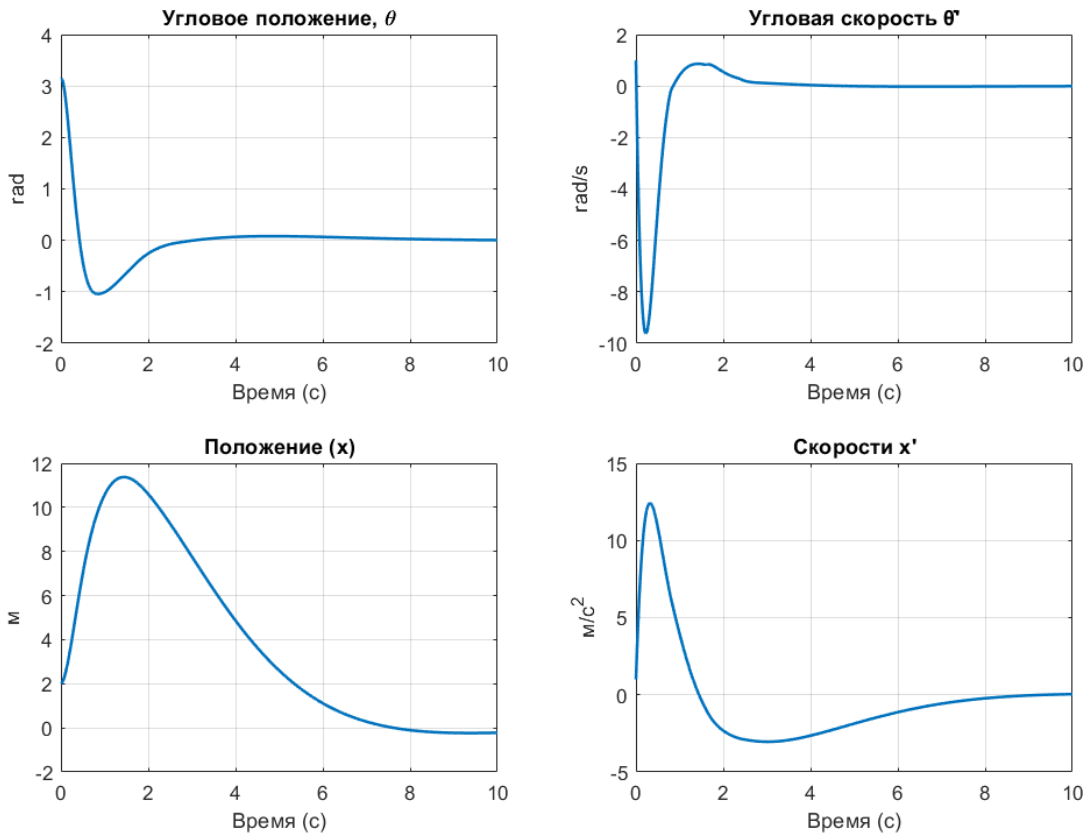


Рисунок 4.13 – Графика состояния системы x(t)

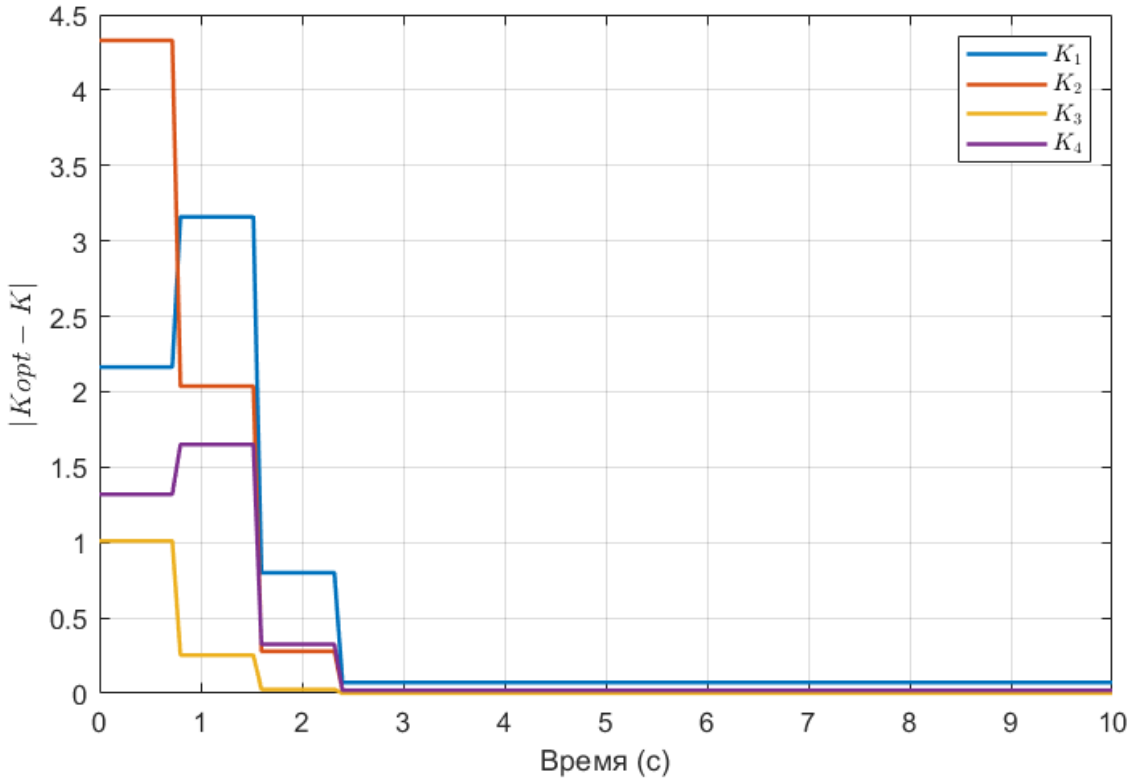


Рисунок 4.14 – График отклонения коэффициентов усиления управления K от их оптимальных значений Kopt

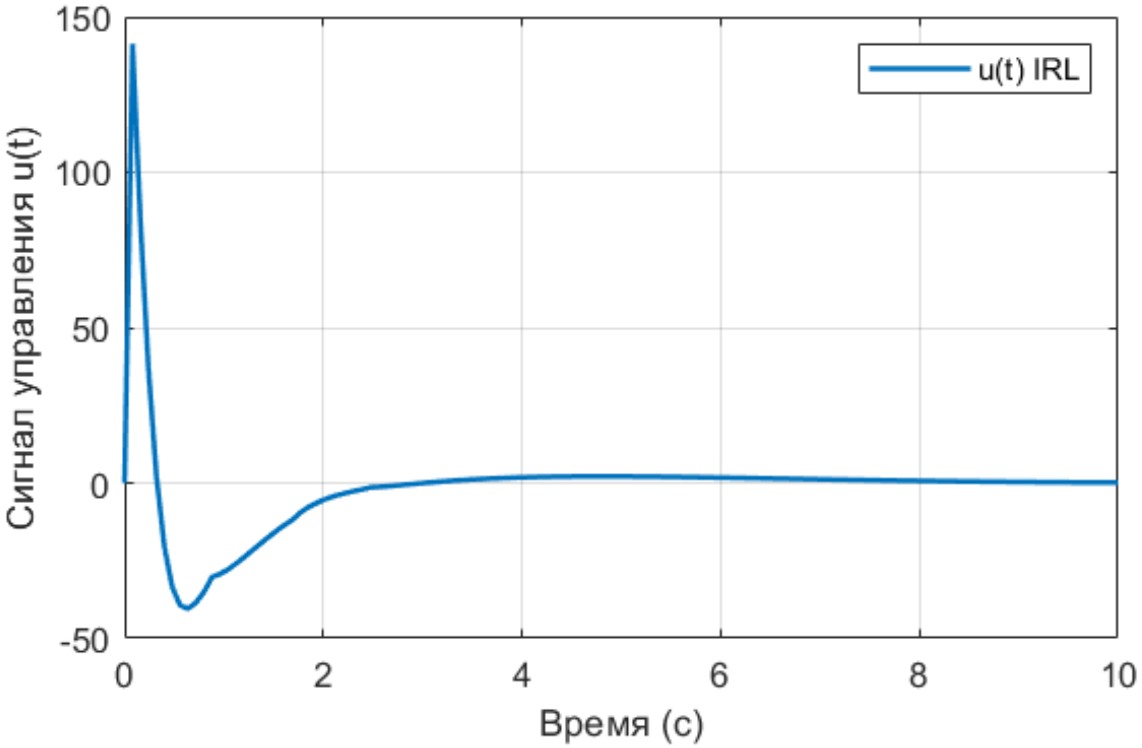


Рисунок 4.15 – График сигнал управления системы u(t)

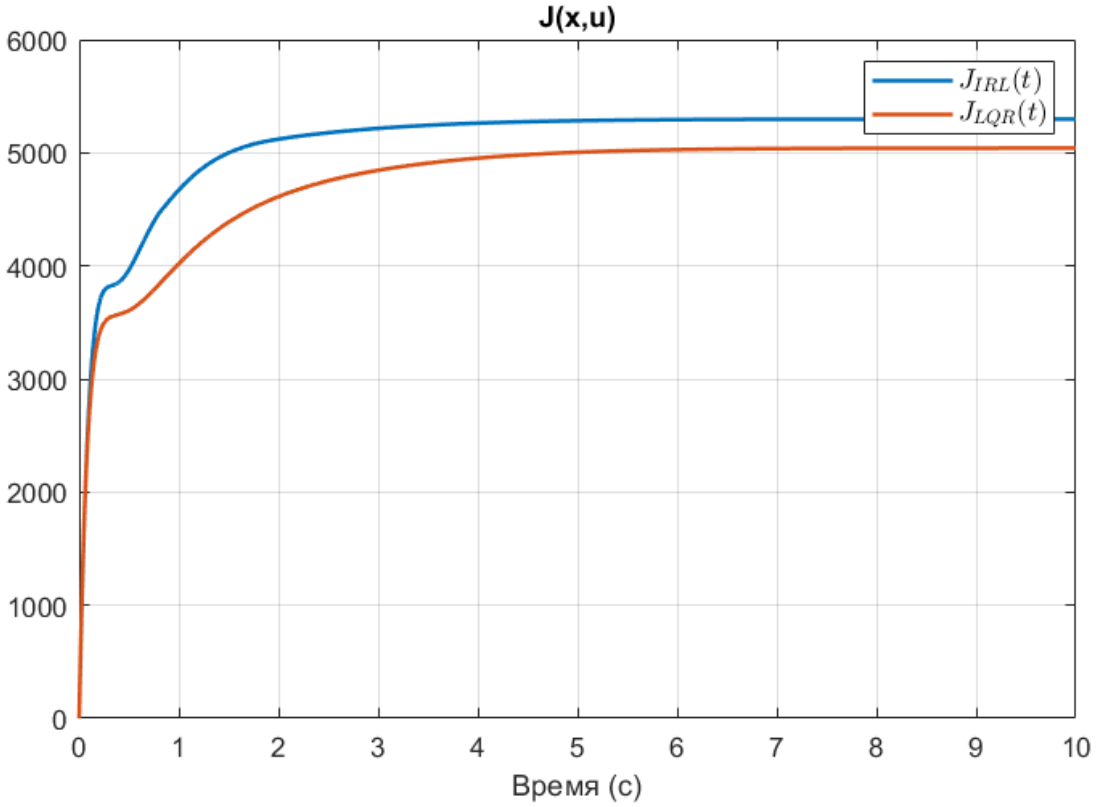


Рисунок 4.16 – График, отражающий значение функции стоимости оптимального регулятора и оптимального адаптивного регулятора

4.5 Моделирование для маятника

Рассмотрим систему маятника

$$x = \theta\text{-mglsintheta-B}\theta\text{ml2} + 01\text{ml2}\ u \qquad (4.8)$$

Где $x = \theta\theta$ - состояние системы, u – сигнал управление

Функция ценности:

$$J = 0\infty xTQx+uTRudt \qquad (4.4)$$

$$Q = 1001\ ,\ R=1$$

Для моделирования, функции вектора активации выбирается следующим образом

$$\varphi\ x = [\ x12x1x2x22\ ]$$

Инициализировать начальный вес $W = 020T$ и $x0 = \pi60T$

Результаты моделирования

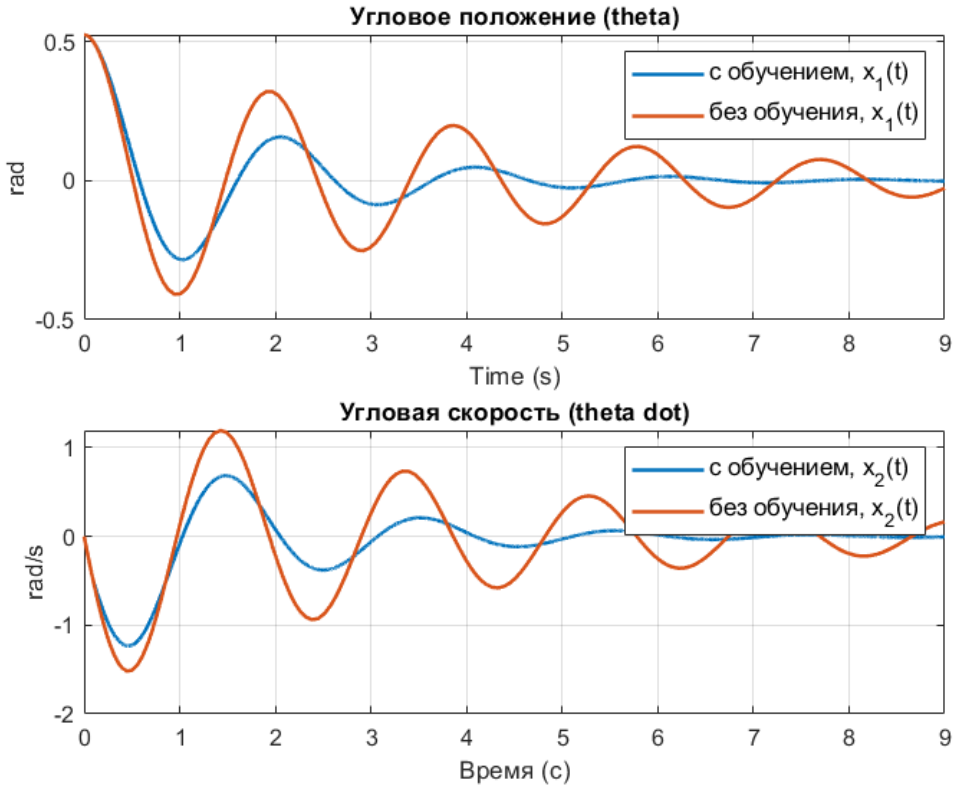


Рисунок 4.17 – Состояние с обучением и без обучения

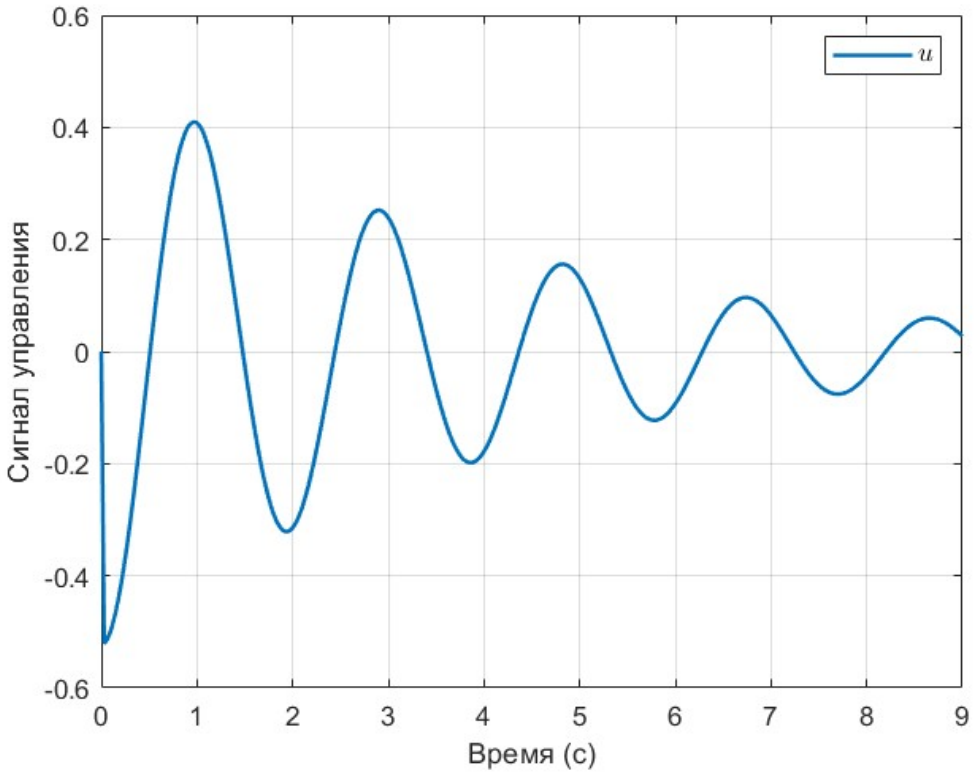


Рисунок 4.18 – Сигнал управления при без обучения

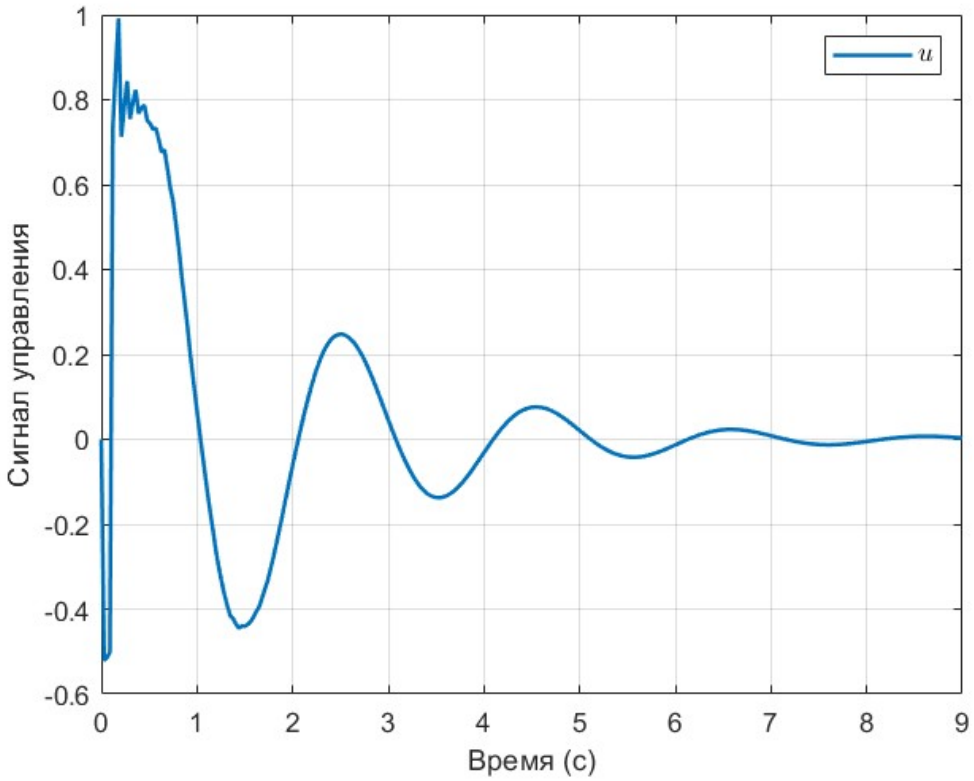


Рисунок 4.19 – Сигнал управления u с обучением

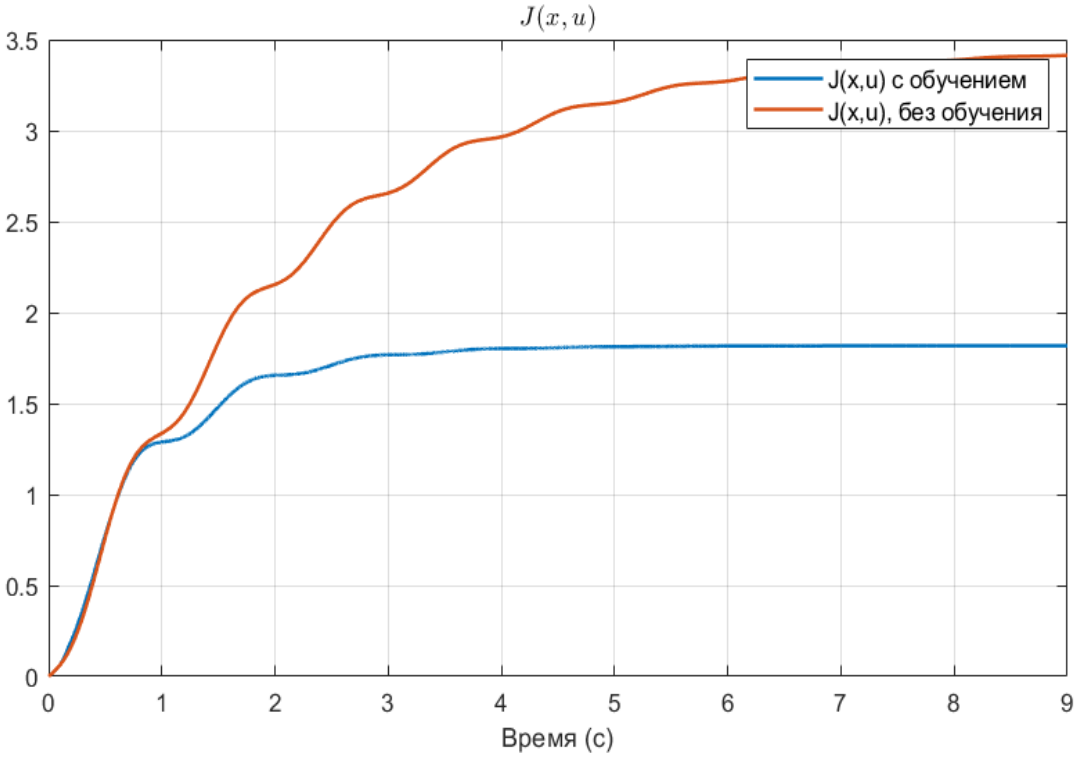


Рисунок 4.20 – Значение функции стоимости J(x,u) с обучением и без обучения

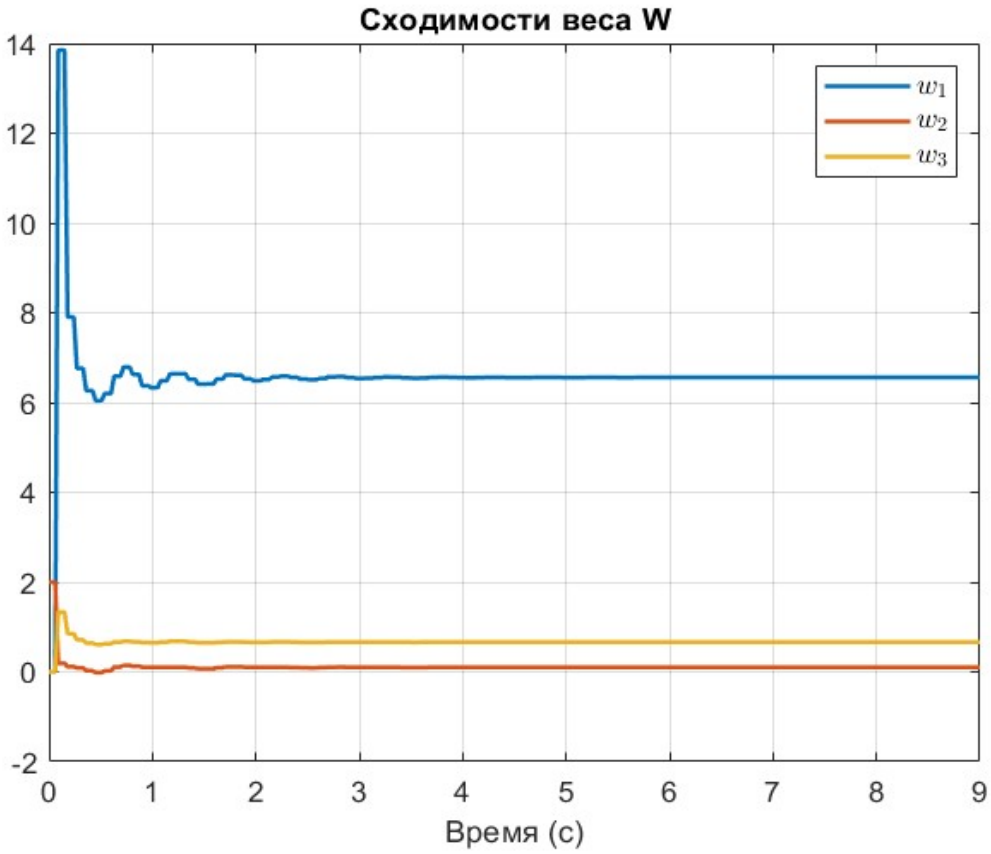


Рисунок 4.21 – Сходимости весов при применении обучения с интегральным подкреплением

Мы видим, что состояние системы сходилось лучше в случае с обучением. Значение функции стоимости также минимизировано, видно, что оно меньше в случае без обучения. Таким образом, при использовании обучение с интегральным подкреплением, регулятор улучшает исходный управляющий сигнал u0 и минимизирует функцию стоимости J(x,u). Для этой системы сложно решить задачу HJB для сравнения с результатами системы. Но в целом алгоритм IRL сошелся к стратегии управления, которая лучше, чем изначально инициализированный сигнал управления. Пример в разделе 4.3 более наглядно показывает сходимость алгоритма к оптимальному значению управления для нелинейной системы.

5. Заключение

В данной работе были изучены основные концепции и идеи, лежащие в основе алгоритма обучения с подкреплением. Были изучены два популярных алгоритма в обучении с подкреплением — это алгоритмы итерации по стратегии и по критерию, предназначенные для решения проблемы поиска оптимальных стратегии.

Представлены и исследованы математические регуляторы, таким как адаптивный регулятор, оптимальное регулятор (LQR) и оптимальный адаптивный регулятор, основанные на обучении с интегральным подкреплением.

Моделирование выполняется в MATLAB. Для линейных систем был реализован анализ и исследование эффективности оптимального адаптивного регулятора на основе обучения с интегральным подкреплением. Проведено сравнение эффективности оптимального адаптивного регулятора с эффективностью оптимального регулятора и адаптивного регулятора. В результате оптимальный адаптивный регулятор имеет эффективность, примерно равную эффективность оптимального регулятора. При этом адаптивный регулятор не сходится к оптимальному управлению. Для нелинейных систем корректность оптимального адаптивного регулятора доказана, когда веса

функции аппроксимируют сходящееся значение к решению уравнения Гамильтона Якоби Беллмана.

Оптимальный адаптивный регулятор, основанный на обучении с интегральным подкреплением, строится на основе данных, собранных по траектории движения системы. Поэтому проанализировано и изучено влияние времени собрки данных T на сходимость оптимального управления оптимального адаптивного регулятора. В результате, чем меньше время T, тем лучше система приближается к оптимальному управлению. Выбор подходящего времени сборки данных был представлен. Если регулятор сходится к оптимальному значению, мы должны прекратить обновление стратегии управления системой, подробности представлены в разделе 4.2.

Таким образом, был успешно исследован оптимальный адаптивный регулятор на основе интегрального обучения с подкреплением. Преимущество этого регулятора в том, что мы получаем оптимальный управляющий сигнал без необходимости полного понимания динамики системы (без знания матрицы A для линейных систем и без знания f(x) для нелинейных систем).

Список использованных источников

1. *Frank Lewis and Draguna Vrabie // Reinforcement learning and adaptive dynamic programming for feedback control. Circuits and Systems Magazine, IEEE, 9:32 – 50, 01 2009.*
2. *Ivo Grondman, Lucian Bus¸oniu, Gabriel A. D. Lopes, and Robert Babuska // A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients*
3. *Ivo Grondman, Lucian Bus¸oniu, Gabriel A.D. Lopes and Robert Babuska // A Survey of Actor-Critic Reinforcement Learning: Standard and Natural Policy Gradients*
4. *V. R. Konda and J. N. Tsitsiklis // "On Actor-Critic Algorithm", SIAM Journal on Control and Optimization, vol. 42, No. 4, pp. 1143–1166, 2003.*
5. *Richard S. Sutton and Andrew G. Barto // Reinforcement Learning: An Introduction, Second Edition, MIT Press, Cambridge, MA, 2018*
6. *Fadi AlMahamid, and Katarina Grolinger, Department of Electrical and Computer Engineering Western University // Reinforcement Learning Algorithms: An Overview and Classification*

7. *A. Geramifard, T. J. Walsh, S. Tellex, G. Chowdhary, N. Roy, and J. P. How// A Tutorial on Linear Function Approximators for Dynamic Programming and Reinforcement Learning, Vol. 6, No. 4 (2013) 375–454, 2013*

8. *Beakcheol Jang, Myeonghwi Kim, Gaspard Harerimana, and Jong Wook Kim // Q-Learning Algorithms: A Comprehensive Classification and Applications, Department of Computer Science, Sangmyung University, Seoul 03016, South Korea*

9. *Draguna Vrabie, Kyriakos G. Vamvoudakis and Frank L. Lewis // Optimal Adaptive Control and Differential Games by Reinforcement Learning Principles*

10. *Frank L. Lewis, Draguna L. Vrabie, Vassilis L. Syrmos // Optimal control-3rd ed, 2012.*

11. *Bahare Kiumarsi, Member, IEEE, Kyriakos G. Vamvoudakis, Senior Member, IEEE, Hamidreza Modares, Member, IEEE, and Frank L. Lewis, Fellow, IEEE, Optimal and Autonomous Control Using Reinforcement Learning: A Survey, 2018.*

12. *Kyriakos G. Vamvoudakis, Draguna Vrabie, Frank L. Lewis, Online adaptive learning of optimal control solutions using integral reinforcement learning, 2011.*

13. *D. Vrabie, O. Pastravanu, M. Abu-Khalaf, and F. L. Lewis, "Adaptive optimal control for continuous-time linear systems based on policy iteration," Automatica, vol. 45, no. 2, pp. 477–484, 2009.*

14. *D. Vrabie and F. L. Lewis, "Neural network approach to continuoustime direct adaptive optimal control for partially-unknown nonlinear systems," Neural Netw., vol. 22, no. 3, pp. 237–246, Apr. 2009.*

15. *Vamvoudakis KG, Lewis FL. Online actor-critic algorithm to solve the continuous-time infinite horizon optimal control problem. Automatica 2010.*

16. *Kleinman D. On an iterative technique for Riccati equation computations. IEEE Transactions on Automatic Control 1968;*

17. *Uğur Yıldıran, Yildiz Technical University, Adaptive Control of an Inverted Pendulum by a Reinforcement Learning based LQR Method.*

18. *Vinayak Kumar, Ruchi Agarwal, Modeling and Control of Inverted Pendulum cart system using PID-LQR based Modern Controller, July 01-03, 2022, Prayagraj, India.*