

**Федеральное государственное автономное образовательное  
учреждение высшего образования  
«Московский физико-технический институт  
(национальный исследовательский университет)»**

**УТВЕРЖДЕНО**

**Директор физтех-школы  
прикладной математики и  
информатики**

**А.М. Райгородский**

	<b>Рабочая программа дисциплины (модуля)</b>
<b>по дисциплине:</b>	Обучение с подкреплением
<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра математических основ управления
<b>курс:</b>	1
<b>квалификация:</b>	магистр

Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен

Аудиторных часов: 30 всего, в том числе:

лекции: 30 час.

семинары: 0 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 30 час.

Подготовка к экзамену: 30 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составил: Н.Е. Юдин, ассистент

Программа обсуждена на заседании кафедры математических основ управления 05.03.2021

## Аннотация

В курсе рассматриваются ключевые понятия и методы обучения с подкреплением. В отличие от классического машинного обучения, в обучении с подкреплением (ОсП) на вход не поступает обучающая выборка прецедентов. Тем не менее, современные алгоритмы глубокого обучения с подкреплением способны решать задачи искусственного интеллекта методом проб и ошибок без использования каких-либо априорных знаний о решаемой задаче. В предлагаемом курсе изучаются принципы работы основных алгоритмов ОсП, позволивших достичь прорывных результатов во многих задачах: от игрового искусственного интеллекта до робототехники. Все необходимые теоретические результаты приводятся с доказательствами, использующими единый подход, унифицированные обозначения и определения.

Основная задача курса – не только предоставить актуальную информацию о задачах обучения с подкреплением и алгоритмах их решения, но и разъяснить разницу между алгоритмами различного вида и причины их представления в конкретных формах.

Курс содержит как обсуждение базовых вопросов обучения с подкреплением, так и разбор задач. Для успешного освоения курса слушателю необходимо владеть основами теории вероятностей, численных методов оптимизации.

## 1. Цели и задачи

### Цель дисциплины

Изучение основных понятий и методов обучения с подкреплением.

### Задачи дисциплины

- освоение студентами базовых знаний в области машинного обучения;
- приобретение теоретических знаний в области обучения с подкреплением;
- оказание консультаций и помощи студентам в решении теоретических и практических задач ОсП.

## 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области прикладной математики и информатики
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности
ПК-3 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами, устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований, проведения корректуры, редактирования, реферирования работ	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания
	ПК-3.2 Умеет вести корректную дискуссию в области информационно-коммуникационных технологий и информационных систем, задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий

## 3. Перечень планируемых результатов обучения по дисциплине (модулю)

В результате освоения дисциплины обучающиеся должны знать:

фундаментальные понятия, законы, и методы ОсП;  
 понятия, аксиомы, методы доказательств и доказательства основных теорем в разделах, входящих в базовую часть цикла;  
 основные свойства соответствующих математических объектов;  
 аналитические и численные подходы и методы для решения типовых прикладных задач ОсП.

уметь:

понять поставленную задачу;  
 использовать свои знания для решения фундаментальных и прикладных задач ОсП;  
 оценивать корректность постановок задач;  
 строго доказывать или опровергать утверждение;  
 самостоятельно находить алгоритмы решения задач, в том числе и нестандартных, и проводить их анализ;  
 самостоятельно видеть следствия полученных результатов.

владеть:

навыками освоения большого объема информации и решения задач ОсП;  
 навыками самостоятельной работы и освоения новых дисциплин;  
 культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов и методов ОсП;  
 предметным языком и навыками грамотного описания решения задач и представления полученных результатов.

#### 4. Содержание дисциплины (модуля), структурированное по темам (разделам) с указанием отведенного на них количества академических часов и видов учебных занятий

##### 4.1. Разделы дисциплины (модуля) и трудоемкости по видам учебных занятий

№	Тема (раздел) дисциплины	Трудоемкость по видам учебных занятий, включая самостоятельную работу, час.			
		Лекции	Семинары	Лаборат. работы	Самост. работа
1	Введение в курс. Постановка задачи обучения с подкреплением. Табличные методы обучения с подкреплением.	10			6
2	Элементы теории принятия решений и случайных процессов. Q-обучение и его вариации.	5			8
3	Policy gradient-подход с натуральным градиентом и схемы «актёр-критик».	5			8
4	Задачи непрерывного управления и имитационное обучение.	10			8
Итого часов		30			30
Подготовка к экзамену		30 час.			
Общая трудоёмкость		90 час., 2 зач.ед.			

## 4.2. Содержание дисциплины (модуля), структурированное по темам (разделам)

### Семестр: 2 (Весенний)

1. Введение в курс. Постановка задачи обучения с подкреплением. Табличные методы обучения с подкреплением.

Кросс-энтропийный метод (CEM).

Динамическое программирование. Value Iteration, Policy Iteration.

Библиотека OpenAI gym. Реализация табличного кросс-энтропийного метода.

Метод зеркального спуска в обучении с подкреплением.

2. Элементы теории принятия решений и случайных процессов. Q-обучение и его вариации.

Марковский процесс принятия решений. Оптимизационная формализация. Двойственность Фенхеля-Рокафеллара.

Deep Q-Network (DQN) и его модификации.

Distributional RL. Categorical DQN (c51), Quantile Regression DQN (QR-DQN).

Анализ сложности алгоритма Q-обучения.

3. Policy gradient-подход с натуральным градиентом и схемы «актёр-критик».

Внутренняя мотивация для исследования среды.

Подход Advantage Actor-Critic (A2C).

Оценивание по методу REINFORCE.

Trust-Region Policy Optimization (TRPO).

Generalized Advantage Estimation (GAE). Proximal Policy Optimization (PPO).

Методы вида Natural Policy Gradient с энтропийной регуляризацией, их глобальная сходимость.

4. Задачи непрерывного управления и имитационное обучение.

Непрерывное управление.

Имитационное обучение. Обратное обучение с подкреплением.

Monte Carlo Tree Search. AlphaZero, MuZero.

Linear Quadratic Regulator (LQR). Model-based RL.

## 5. Описание материально-технической базы, необходимой для осуществления образовательного процесса по дисциплине (модулю)

Учебная аудитория, оснащенная компьютером и мультимедийным оборудованием (проектор, звуковая система).

## 6. Перечень рекомендуемой литературы

### Основная литература

1. Python для сложных задач: наука о данных и машинное обучение, Python data science handbook. Essential tools for working with data, Электронная версия печатной публикации / . — Санкт-Петербург, Питер, 2018

### Дополнительная литература

1. Глубокое обучение / Я. Гудфеллоу, И. Бенджио, А. Курвилль . — Москва, ДМК Пресс, 2018.— URL: <https://e.lanbook.com/book/107901> (дата обращения: 29.01.2021). - Полный текст (Режим доступа : из сети МФТИ / Удаленный доступ)

## **7. Перечень ресурсов информационно-телекоммуникационной сети "Интернет", необходимых для освоения дисциплины (модуля)**

<http://www.machinelearning.ru> – профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

<http://shad.yandex.ru> – сайт школы анализа данных Яндекса.

<https://openai.com/blog/>

<https://lilianweng.github.io/posts/>

<https://distill.pub/>

<https://hackernoon.com/>

## **8. Перечень информационных технологий, используемых при осуществлении образовательного процесса по дисциплине (модулю), включая перечень необходимого программного обеспечения и информационных справочных систем (при необходимости)**

<http://www.machinelearning.ru> – профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных.

<http://shad.yandex.ru> – сайт школы анализа данных Яндекса.

<https://openai.com/blog/>

<https://lilianweng.github.io/posts/>

<https://distill.pub/>

<https://hackernoon.com/>

## **9. Методические указания для обучающихся по освоению дисциплины (модуля)**

Студент, изучающий курс "Обучение с подкреплением", должен с одной стороны, овладеть общим понятийным аппаратом, а с другой стороны, должен научиться применять теоретические знания на практике.

В результате изучения дисциплины студент должен знать основные определения, понятия, аксиомы, методы доказательств.

Успешное освоение курса требует напряжённой самостоятельной работы студента. В программе курса приведено минимально необходимое время для работы студента над темой. Самостоятельная работа включает в себя:

- чтение и конспектирование рекомендованной литературы,
- проработку учебного материала (по конспектам лекций, учебной и научной литературе), подготовку ответов на вопросы, предназначенных для самостоятельного изучения, доказательство отдельных утверждений, свойств;
- подготовку к экзамену.

Руководство и контроль за самостоятельной работой студента осуществляется в форме индивидуальных консультаций.

Показателем владения материалом служит умение решать задачи. Для формирования умения применять теоретические знания на практике студенту необходимо решать как можно больше задач. При решении задач каждое действие необходимо аргументировать, ссылаясь на известные теоретические сведения.

Важно добиться понимания изучаемого материала, а не механического его запоминания. При затруднении изучения отдельных тем, вопросов, следует обращаться за консультациями к лектору или преподавателю, ведущему практические занятия.

Литература для самостоятельной работы:

1. Ivanov S. Reinforcement Learning Textbook //arXiv preprint arXiv:2201.09746. – 2022.
2. Gasnikov A. et al. Lecture Notes on Stochastic Processes //arXiv preprint arXiv:1907.01060. – 2019.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Прикладная математика и информатика
<b>профиль подготовки:</b>	Прикладная математика и информатика Физтех-школа Прикладной Математики и Информатики кафедра математических основ управления
<b>курс:</b>	<u>1</u>
<b>квалификация:</b>	магистр
Семестр, формы промежуточной аттестации: 2 (весенний) - Экзамен	
<b>Разработчик:</b>	Н.Е. Юдин, ассистент

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен решать актуальные задачи фундаментальной и прикладной математики	ОПК-1.1 Знает и способен использовать в профессиональной деятельности фундаментальные научные знания и новые научные принципы и методы исследований в области прикладной математики и информатики
	ОПК-1.2 Способен обобщать и критически оценивать опыт и результаты научных исследований в области профессиональной деятельности
	ОПК-1.3 Понимает междисциплинарные связи в области прикладной математики и информатики и способен их применять при решении задач профессиональной деятельности
ПК-3 Владеет навыками участия в научных дискуссиях, выступления с сообщениями и докладами, устного, письменного и виртуального (размещение в информационных сетях) характера, представления материалов собственных исследований, проведения корректуры, редактирования, реферирования работ	ПК-3.1 Знает основы ведения научной дискуссии и формы устного научного высказывания
	ПК-3.2 Умеет вести корректную дискуссию в области информационно-коммуникационных технологий и информационных систем, задавать вопросы и отвечать на поставленные вопросы по теме научной работы
	ПК-3.3 Имеет практический опыт участия в научных студенческих конференциях, очных, виртуальных, заочных обсуждениях научных проблем в области информационных технологий

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Обучение с подкреплением» обучающийся должен:

### знать:

фундаментальные понятия, законы, и методы ОсП;  
понятия, аксиомы, методы доказательств и доказательства основных теорем в разделах, входящих в базовую часть цикла;  
основные свойства соответствующих математических объектов;  
аналитические и численные подходы и методы для решения типовых прикладных задач ОсП.

### уметь:

понять поставленную задачу;  
использовать свои знания для решения фундаментальных и прикладных задач ОсП;  
оценивать корректность постановок задач;  
строго доказывать или опровергать утверждение;  
самостоятельно находить алгоритмы решения задач, в том числе и нестандартных, и проводить их анализ;  
самостоятельно видеть следствия полученных результатов.

### владеть:

навыками освоения большого объема информации и решения задач ОсП;  
навыками самостоятельной работы и освоения новых дисциплин;  
культурой постановки, анализа и решения математических и прикладных задач, требующих для своего решения использования математических подходов и методов ОсП;  
предметным языком и навыками грамотного описания решения задач и представления полученных результатов.

### **3. Перечень типовых (примерных) вопросов, заданий, тем для подготовки к текущему контролю**

С целью контроля освоения обучающимися учебного материала проводится устный опрос в начале занятия по теме прошлого занятия.

### **4. Перечень типовых (примерных) вопросов и тем для проведения промежуточной аттестации обучающихся**

1. Кросс-энтропийный метод в общем виде. Его применение для решения задач оптимизации и задач обучения с подкреплением.
2. Уравнения Беллмана для функций ценности. Алгоритмы Policy/Value Iteration.
3. Табличные методы: Монте-Карло, Q-learning, Sarsa.
4. Алгоритм DQN и его модификации: Double DQN, приоритизированный буфер, дуэльная архитектура, шумные сети, многошаговый DQN, память.
5. Distributional-подход в RL. Алгоритмы c51 и QR-DQN.
6. Подход Policy gradient. Алгоритмы Reinforce и A2C.
7. Метод Trust-Region Policy Optimization (TRPO), его теоретическое обоснование.
8. Bias-variance trade-off в обучении с подкреплением. Оценка GAE. Алгоритм Proximal Policy Optimization (PPO).
9. Детерминированный градиент по политике. Off-policy алгоритмы для задач непрерывного управления: DDPG, Twin Delayed DDPG (TD3).
10. Обучение с подкреплением с добавлением энтропии. Алгоритм Soft Actor-Critic.
11. Имитационное обучение и обратное обучение с подкреплением. Схема Guided Cost Learning. Генеративно-сопоставительное имитационное обучение (GAIL).
12. Задача многоруких бандитов, UCB-бандиты. Внутренняя мотивация: дистилляция случайной сети (RND) и внутренний модуль любопытства (ICM).
13. Monte Carlo Tree Search в общем виде. Методы AlphaZero и MuZero.
14. Линейно-квадратичный регулятор и его итеративная версия. Общая схема Model-based RL.

Примеры экзаменационных билетов

Билет №1

1. Кросс-энтропийный метод в общем виде. Его применение для решения задач оптимизации и задач обучения с подкреплением.
2. Линейно-квадратичный регулятор и его итеративная версия. Общая схема Model-based RL.

Билет №2

1. Алгоритм DQN и его модификации: Double DQN, приоритизированный буфер, дуэльная архитектура, шумные сети, многошаговый DQN, память.
2. Обучение с подкреплением с добавлением энтропии. Алгоритм Soft Actor-Critic.

### **Критерии оценивания**

- оценка «отлично (10)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение уверенно применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;



- оценка «отлично (9)» выставляется студенту, показавшему всесторонние, систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, свободное и правильное обоснование принятых решений;
- оценка «отлично (8)» выставляется студенту, показавшему всесторонние систематизированные, глубокие знания учебной программы дисциплины и умение применять их на практике при решении конкретных задач, и правильное обоснование принятых решений;
- оценка «хорошо (7)» выставляется студенту, если он твердо знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (6)» выставляется студенту, если он знает материал, грамотно и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «хорошо (5)» выставляется студенту, если он знает материал, и по существу излагает его, умеет применять полученные знания на практике, но допускает в ответе или в решении задач некоторые неточности;
- оценка «удовлетворительно (4)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «удовлетворительно (3)» выставляется студенту, показавшему фрагментарный, разрозненный характер знаний, недостаточно правильные формулировки базовых понятий, нарушения логической последовательности в изложении программного материала, но при этом он владеет фрагментарно основными разделами учебной программы, необходимыми для дальнейшего обучения и может применять полученные знания по образцу в стандартной ситуации;
- оценка «неудовлетворительно (2)» выставляется студенту, который не знает большей части основного содержания учебной программы дисциплины, допускает грубые ошибки в формулировках основных понятий дисциплины и не умеет использовать полученные знания при решении типовых практических задач;
- оценка «неудовлетворительно (1)» выставляется студенту, который не знает формулировок основных понятий дисциплины.

## **5. Методические материалы, определяющие процедуры оценивания знаний, умений, навыков и (или) опыта деятельности**

Во время проведения экзамена обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций.

Экзамен проводится путем организации специального опроса, проводимого в устной форме.