

**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>	Математика больших данных
<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Математическое моделирование и компьютерные технологии Физтех-школа Прикладной Математики и Информатики кафедра математических основ управления
<b>курс:</b>	4
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

Аудиторных часов: 45 всего, в том числе:

лекции: 30 час.

семинары: 15 час.

лабораторные занятия: 0 час.

Самостоятельная работа: 45 час.

Всего часов: 90, всего зач. ед.: 2

Количество контрольных работ, заданий: 2

Программу составили:

Э.А. Горбунов, ассистент

Г.А. Кабатянский, д-р физ.-мат. наук

Программа обсуждена на заседании кафедры математических основ управления 05.03.2021

## Аннотация

В последние годы стали появляться университетские курсы, посвященные важным математическим принципам, часто используемым при решении задач больших размерностей, возникающих в анализе данных. Приведем лишь один, наиболее яркий, на наш взгляд, пример такого курса и книги, написанной на его основе: Blum A., Hopcroft J., Kannan R. Foundations of data science. – Cambridge University Press, 2020.

Отличие таких курсов от многих других математических курсов, в которых рассматриваются продвинутое способы решения задач анализа данных больших размеров заключается в акцентировании внимания на общих математических принципах, повсеместно используемых в анализе больших массивов данных. К таким принципам можно отнести: принцип концентрации меры, малоранговые приближения матриц, стохастический градиентный спуск.

Настоящий курс ставит целью познакомить студентов с соответствующей математикой, что впоследствии должно помочь им в изучении специализированных разделов анализа данных (машинного обучения, статистики, обучения с подкреплением, численных методов оптимизации, моделирования больших сетей и т.д.).

В первой (и главной) части курса (6 лекций) акцент будет сделан на явлении концентрации меры. Начиная с классических результатов Гаусса, Максвелла, Пуанкаре, Леви, Мильмана, планируется постепенно перейти к современным результатам и приложениям, в том числе, возникающим в разнообразных задачах анализа данных.

Во второй части курса (4 лекции) классическая теорема о SVD-разложении и его различные обобщения будут продемонстрированы в приложениях к данным, хранящимся в многомерных массивах.

В заключительной третьей части курса будут рассмотрены вопросы, связанные с численными методами решения задач (выпуклой) стохастической оптимизации в пространствах больших размеров. Такие задачи часто возникают в разнообразных приложениях, в том числе, в анализе данных (принцип максимального правдоподобия в статистике, минимизация риска в машинном обучении).

## 1. Цели и задачи

### Цель дисциплины

- ознакомление студентов с некоторыми типами оптимизационных задач, возникающих в современном анализе данных, вопросами теории адаптивных численных методов первого порядка для задач минимизации, вариационных неравенств, седловых задач, основами теории методов для задач невыпуклой оптимизации.

### Задачи дисциплины

- приобретение слушателями теоретических знаний и практических умений и навыков в области концентрации меры и ее приложений,
- приобретение слушателями навыков владения аппаратом матричных разложений,
- владение общим подходом к решению широкого класса прикладных задач анализа данных, допускающих математическую формализацию.
- подготовка слушателей к изучению смежных математических дисциплин, связанных с анализом данных, машинным обучением, оптимизацией.
- приобретение навыков приложения концентрации меры и матричных разложений в других естественнонаучных дисциплинах.

## 2. Перечень формируемых компетенций

Освоение дисциплины направлено на формирование следующих компетенций:

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки

ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности
---	---

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

В результате освоения дисциплины обучающиеся должны

знать:

- фундаментальные подходы к решению задач концентрации меры и матричных разложений;
- 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.

уметь:

- понять поставленную задачу;
- использовать свои знания для решения фундаментальных и прикладных задач;
- 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;
- 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 математики больших данных;
- 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	Концентрация меры на сфере (около экватора)	5	5		5
2	Примеры концентрации меры (случайные графы, группа поворотов, случайные перестановки и т.д.)	5			10
3	Теорема Джонсона-Линденштраусса.	5	5		5
4	Теоремы Клартага.	5			10
5	Неравенства концентрации меры.	5	5		5
6	Малоранговые приближения матриц и векторов.	5			10
Итого часов		30	15		45
Подготовка к экзамену		0 час.			

Общая трудоёмкость	90 час., 2 зач.ед.
--------------------	--------------------

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

Семестр: 7 (Осенний)

##### 1. Концентрация меры на сфере (около экватора)

Общая постановка задачи. Теорема Максвелла о скорости распределения молекул газа в сосуде. Неравенства Леви и Пуанкаре. Примеры концентрации равномерной меры на других множествах. Приложения к теории информации.

##### 2. Примеры концентрации меры (случайные графы, группа поворотов, случайные перестановки и т.д.)

Модели случайных графов Эрдёша-Реньи, группы перестановок, поворотов и концентрация равномерной меры на таких дискретных множествах. Неравенство Талаграна.

##### 3. Теорема Джонсона-Линденштраусса.

Сжатие информации с помощью теоремы Джонсона-Линденштраусса. Приложения к построению RIP-матриц в L1-оптимизации.

##### 4. Теоремы Клартага.

Понимание теоремы Клартага как обобщение теоремы Максвелла. Обзор результатов теории концентрации меры (по В.Д. Мильману).

##### 5. Неравенства концентрации меры.

Неравенства Азума-Хефдинга, Немировского, Бернштейна-Фридмана, неравенства для случайных матриц (Тропп, Колчинский и др.). Приложения неравенств концентрации меры к задачам стохастической оптимизации.

##### 6. Малоранговые приближения матриц и векторов.

Матричные нормы. Сингулярное разложение (SVD) и теорема Эккарта-Янга-Мирского. Принцип наибольшего объема. CGR разложение и его приложения. Алгоритмы построения малоранговых приближений. Тензорные разложения: каноническое разложение и разложение Таккера, higher-order SVD.

#### 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. Математические основы машинного обучения и прогнозирования [Текст] / В. В. Вьюгин ; Моск. физ.-техн. ин-т (гос. ун-т), Лаб. структурных методов анализа данных в предсказательном моделировании (ПреМоЛаб), Ин-т проблем передачи информации им. А. А. Харкевича РАН, М., МЦНМО, 2013

#### Дополнительная литература

1. Оптимальность в играх и решениях [Текст]/Э. Й. Вилкас, -М., Наука, 1990
2. Глубокое обучение, Электрон. версия печ. публикации / Я. Гудфеллоу, И. Бенджио, А. Курвилль . — Москва, ДМК Пресс, 2018

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

<http://dm.fizteh.ru/>

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

<https://www.cs.cornell.edu/jeh/book%20no%20solutions%20March%202019.pdf>

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

Студент, изучающий курс "Математика больших данных", 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 дифференцированному зачету.

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.

**ОЦЕНОЧНЫЕ МАТЕРИАЛЫ ПО ДИСЦИПЛИНЕ (МОДУЛЮ)**

<b>по направлению:</b>	Информатика и вычислительная техника
<b>профиль подготовки:</b>	Математическое моделирование и компьютерные технологии Физтех-школа Прикладной Математики и Информатики кафедра математических основ управления
<b>курс:</b>	4
<b>квалификация:</b>	бакалавр

Семестр, формы промежуточной аттестации: 7 (осенний) - Дифференцированный зачет

**Разработчики:**

Э.А. Горбунов, ассистент  
Г.А. Кабатянский, д-р физ.-мат. наук

## 1. Компетенции, формируемые в процессе изучения дисциплины

Код и наименование компетенции	Индикаторы достижения компетенции
ОПК-1 Способен применять фундаментальные знания, полученные в области физико-математических и (или) естественных наук, и использовать их в профессиональной деятельности	ОПК-1.2 Способен строить математические модели, производить количественные расчеты и оценки
ОПК-2 Способен использовать современные информационные технологии и программные средства при решении задач профессиональной деятельности, соблюдая требования информационной безопасности	ОПК-2.1 Способен применять современные вычислительную технику и сервисы сети Интернет в области (сфере) профессиональной деятельности

## 2. Показатели оценивания компетенций

В результате изучения дисциплины «Математика больших данных» обучающийся должен:

### знать:

- фундаментальные подходы к решению задач концентрации меры и матричных разложений;
- 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.

### уметь:

- понять поставленную задачу;
- использовать свои знания для решения фундаментальных и прикладных задач;
- 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;
- 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 математики больших данных;
- 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. Концентрация меры на сфере (около экватора);
2. Примеры концентрации меры (случайные графы, группа поворотов, случайные перестановки и т.д.);
3. Теорема Джонсона-Линденштраусса;
4. Теоремы Клартага;

5. Приложения к кодированию и теории информации;
6. Неравенства концентрации меры. Азума-Хефдинга, Бернштейна-Фридмана;
7. Матричные нормы;
8. Сингулярное разложение (SVD) и теорема Эккарта-Янга-Мирского;
9. Принцип наибольшего объема;
10. CGR-разложение и его приложения;
11. Алгоритмы построения малоранговых приближений;
12. Тензорные разложения: каноническое разложение и разложение Таккера, higher-order SVD;
13. Стохастический градиентный спуск;
14. Онлайн-оптимизация.

#### Билет 1

1. Явление концентрации меры на сфере. Изопериметрическое неравенство (задача Дидоны). Теорема Максвелла. Теорема Клартага.
2. Сингулярное разложение (SVD) и теорема Эккарта-Янга-Мирского

#### Билет 2

1. Теорема Джонсона-Линденштраусса (с доказательством). Пример приложения в комбинаторной оптимизации (RIP матрицы).
2. Тензорные разложения: каноническое разложение и разложение Таккера, higher-order SVD

#### 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 дифференцированного зачета обучающиеся могут пользоваться программой дисциплины, а также справочной литературой, вычислительной техникой, конспектами лекций.

Дифференцированный зачет проводится путем организации специального опроса, проводимого в устной форме.