

От VC-размерности к законам масштабирования: эволюция теоретических подходов к оценке сложности и обобщающей способности

А. В. Грабовой*

В работе представлен обзор современного состояния исследований, посвященных теоретическим методам оценки сложности моделей машинного обучения и их связи с обобщающей способностью. Методологически работа оформлена как обзор предметной области. Указываются источники библиографического поиска, этапы дедупликации и полуавтоматического сужения корпуса документов, экспертный отбор и состав итогового аналитического набора публикаций. Основной вклад обзора состоит в анализе классических мер сложности из статистической теории обучения и их ограниченности для объяснения поведения современных перепараметризованных глубоких моделей машинного обучения. Показано, что VC-размерность, радемахеровские и PAC-байесовские оценки сохраняют значение как теоретический результат, однако характеризуют мощность класса гипотез в худшем случае и лишь ограниченно учитывают локальные геометрические характеристики решения, траекторию оптимизации и архитектурные особенности. Основное внимание уделено современным направлениям, включающим анализ ландшафта функции потерь, спектральных характеристик гессiana, методов оптимизации и эмпирических законов масштабирования. В результате предложена интерпретация текущего состояния предметной области, согласно которой современная теория сложности должна описывать не только класс моделей, но и эффективную сложность решений, учитывающих как локальные характеристики модели, так и данные, на которых она обучается.

Ключевые слова: теория сложности моделей, обобщающая способность, глубокое обучение, VC-размерность, радемахеровская сложность, ландшафт функции потерь, гессиан, двойной спуск, законы масштабирования.

* *Грабовой Андрей Валериевич* — кандидат физико-математических наук, старший научный сотрудник, Институт проблем управления им. В. А. Трапезникова РАН, Москва, Россия, e-mail: grabovoy@ipu.ru, ORCID: 0000-0002-4031-0025.

Grabovoy Andrey Valerievich — Candidate of Physical and Mathematical Sciences, Senior Researcher, Institute of Control Sciences of the Russian Academy of Sciences, Moscow, Russia.

1. Введение

Теория сложности моделей в машинном обучении — раздел статистической теории обучения и инструмент для анализа моделей машинного обучения. Теория отвечает на прикладные вопросы, связанные с анализом моделей. Например, какой объем данных нужен для обучения, почему архитектуры с близким числом параметров дают разное обобщение, какие методы оптимизации связаны с устойчивыми решениями и как интерпретировать рост качества при масштабировании моделей и данных. Для моделей глубокого обучения это существенно, так как современные модели сочетают сильную перепараметризацию с высокой эмпирической эффективностью. В данной работе под обзором предметной области (англ. *scoping review*) предполагается формат систематизированного описания области и сопоставления исследовательских трендов.

В рамках классической сложности основным объектом анализа выступает класс гипотез. В этом VC-размерность [1] и PAC-обучаемость [2] стали базовыми инструментами контроля обобщающей способности. Позднее это дополнили ростовые функции и оценки [3, 5], зависящие от выборки (англ. *data-dependent*). Однако развитие глубоких нейронных сетей показало, что одной лишь мощности класса моделей недостаточно для объяснения поведения моделей машинного обучения [30]. Модель с миллионами или миллиардами параметров может как переобучиться на заданный корпус данных, так и демонстрировать высокое качество обобщения [16]. При этом один и тот же класс функций способен вести себя по-разному в зависимости от алгоритма обучения, архитектурных ограничений и структуры данных [16, 30].

Наблюдается разрыв между числом параметров и эффективной сложностью решения. В рамках классической теории рост параметризации повышает риск переобучения [2]. Однако современные эмпирические результаты демонстрируют более сложную зависимость. Увеличение размера модели может ухудшать обобщение только при отдельных методах оптимизации, тогда как в более широком перепараметризованном режиме обучения качество нередко вновь улучшается. Следовательно, число параметров выступает лишь грубым методом оценки сложности и само по себе не определяет, какие функции реализуются после обучения. Существенными факторами становятся архитектурные ограничения, свойства данных, неявная регуляризация оптимизатора и геометрия найденного минимума [12].

По этой причине теория сложности для глубокого обучения постепенно смещается от анализа вместиимости класса гипотез к анализу эффективной сложности конкретных решений [13]. В современных иссле-

дованиях это выражается в геометрии ландшафта функции потерь [12] и спектральным характеристикам гессиана [14]. Параллельно усилилось внимание к режимам обучения [49] и эмпирическим законам масштабирования [13, 18]. В данном обзоре эта эволюция рассматривается как единый исследовательский тренд. От классических верхних оценок к подходам, ориентированным на описание свойств решений, достигаемых в процессе обучения.

В данной работе выводы строятся с использованием заранее зафиксированной процедуры построения литературного корпуса. В анализ включены классические источники статистической теории обучения и современные теоретические, методологические и обзорные публикации по геометрии ландшафта, методам обучения и масштабированию. Такой дизайн позволяет сопоставить классические и современные исследовательские тренды на широком и явно описанном наборе источников.

2. Методология обзора теории сложности моделей машинного обучения

Задача данной работы состоит в том, чтобы охватить основные направления теоретической литературы о сложности и обобщении в методах машинного обучения, зафиксировать процедуру построения корпуса и обеспечить воспроизводимость отбора на уровне источников данных и критериев включения. Данный обзор не заменяет систематический обзор с полной регистрацией протокола по стандартам PRISMA (англ. Preferred Reporting Items for Systematic Reviews and Meta-Analyses), однако позволяет явно разделить широкий поисковый охват и содержательный анализ.

Первичный библиографический сбор публикаций по тематике исследования выполняется по трем открытым агрегаторам научных публикаций — OpenAlex¹, arXiv² и Crossref³. Поиск выполняется посредством однотипных по смыслу поисковых запросов. Запросы охватывают классические и современные направления исследований. К классическим относятся VC-размерность и радемахеровская сложность. К современным — PAC-байесовские и информационно-теоретические границы обобщения, алгоритмическая устойчивость, ландшафт функции потерь, спектр гессиана,

¹OpenAlex API: <https://developers.openalex.org/api-reference/introduction#api-overview>.

²arXiv API: <https://info.arxiv.org/help/api/index.html>.

³Crossref REST API: <https://www.crossref.org/documentation/retrieve-metadata/rest-api/>.

двойной спуск, нейронное касательное ядро, режимы обучения, законы масштабирования, а также архитектуры трансформеров и диффузионных моделей. Для каждой формулировки запроса из каждого источника извлекается по 100 записей. Затем результаты объединяются в единый список публикаций с заголовками и, по возможности, аннотациями.

Записи с разных источников сопоставляются по идентификаторам DOI, совпадению нормализованных заголовков и близким полям описания. Дубликаты объединяются, причем для каждой уникальной работы сохраняется информация о числе поисковых попаданий и перечне запросов, которые ведут к включению работы в список. Такой подход учитывает устойчивость темы к различным формулировкам запроса, не подменяя содержательный отбор автоматически собранными показателями.

Полный дедуплицированный массив по объему не подлежит содержательному анализу целиком. Следовательно, на следующем этапе к каждой записи применяется правило-ориентированная процедура ранжирования, использующая данные заголовка и аннотации. Процедура учитывает совпадение с теоретически существенными лексическими шаблонами и принадлежность к заранее заданным тематическим группам, таким как классические методы, РАС-байесовские методы и информация, ландшафт и гессиян, режимы обучения нейронного касательного ядра (англ. Neural Tangent Kernel, NTK) и двойного спуска, масштабирование и архитектуры. Дополнительно учитывается число независимых поисковых попаданий и штрафы за доминирование узко прикладного контекста при слабой теоретической базе публикации. Все найденные публикации дополнительно распределяются по тематическим группам с ограничением перекоса, чтобы сокращенный перечень кандидатов сохранял представительность по исследовательским трендам, а не сводился к одной выделенной области исследований. В результате для экспертного разбора формируется сокращенный перечень из 500 релевантных записей.

На этапе ручного отбора из сокращенного списка кандидатов исключаются публикации, в которых термины обобщения и сложности использовались преимущественно в прикладном или узкоотраслевом значении без теоретического вклада в предметную область. Также исключаются работы, не добавляющие к обзору нового содержания относительно уже включенных источников. Предпочтение отдается публикациям, удовлетворяющим условиям: (1) формулируется фундаментальный теоретический результат или нетривиальная оценка сложности, (2) предлагается современный механизм, существенный для понимания сложности, (3) фиксируется устойчивый эмпирический результат, который теория сложности обязана учитывать.

Независимо от сокращенного списка кандидатов в итоговый перечень публикаций вручную включен набор фундаментальных источников по статистической теории обучения, устойчивости, ключевым архитектурам и режимам обучения. Такое дополнение методологически обосновано тем, что ранние канонические работы и отдельные базовые публикации по современным архитектурам не всегда устойчиво поднимаются в выдаче поиска по ключевым словам, ориентированного на недавнюю прикладную и вторичную литературу. Без явного включения этих источников корпус искажался бы в сторону современных обзоров.

Таблица 1. Тематическое распределение кандидатов и финального корпуса

Тематическая группа	В 500	Доля, %	Auto	Manual+	В финале
Классические границы	101	20.2	13	6	19
РАС-байесовские методы и информация	120	24.0	5	2	7
Ландшафт и гессиан	96	19.2	11	4	15
Режимы обучения	84	16.8	11	4	15
Масштабирование и архитектуры	45	9.0	7	4	11
Обзорные работы	54	10.8	9	0	9
Итого	500	100.0	56	20	76

Итоговый аналитический корпус включает 76 публикаций, из них 13 опорных работ по ключевым направлениям обзора, 44 дополнительных источников для контекста и сопоставления близких трендов и 19 зафиксированных фундаментальных работ. Из табл. 1 видно, что в списке из 500 кандидатов крупнейшие доли имеют вероятностно-информационная и классическая группы. Наблюдаемая неравномерность групп отражает свойства поисковой процедуры и текущую структуру доступной литературы, а не априорную иерархию важности направлений.

Несмотря на формализованную процедуру отбора, предложенная методология имеет ряд ограничений. Во-первых, поиск опирается на покрытие выбранных библиографических агрегаторов и может неполно отражать работы, индексируемые неравномерно или с задержкой. Во-вторых, ключевые запросы и лексические шаблоны неизбежно задают тематическое смещение, так как они преимущественно извлекают публикации, использующие устоявшийся словарь теории обобщения, по сравнению с работами с близким содержанием, но иной терминологией. В-третьих, правило-ориентированное ранжирование и последующий экспертный отбор снижают шум, но сохраняют интерпретационную зависимость от исследовательской рамки обзора. В-четвертых, обзор предметной области ориентирован на структурирование области и сравнительный анализ, а не

на метаанализ, поэтому выводы носят концептуальный, а не статистически агрегированный характер. Представленная процедура не претендует на полноту в смысле систематического обзора всех существующих публикаций по теме. Назначение работы — обеспечить прозрачный путь от широкого поиска к управляемому набору источников и зафиксировать границы обзорного анализа.

3. Классические меры сложности и их ограничения

Классическая теория обучения связывает обобщение с контролем мощности класса гипотез. VC-размерность измеряет способность класса реализовывать все возможные бинарные разметки на конечных выборках и тем самым служит критерием равномерной сходимости эмпирического риска к истинному [1, 3]. В PAC-постановке это приводит к оценкам требуемого объема выборки через параметры точности и сложности класса [2]. Для классов гипотез умеренной сложности такая картина остается концептуально строгой: конечная VC-размерность означает обучаемость, а рост сложности увеличивает требования к данным. Для бинарной задачи это выражается оценкой обобщения через VC-размерность:

$$R(h) \leq \hat{R}_S(h) + \sqrt{\frac{d_{VC} \left(\log \frac{2n}{d_{VC}} + 1 \right) + \log \frac{4}{\delta}}{n}}, \quad (1)$$

где $R(h)$ — истинный риск, $\hat{R}_S(h)$ — эмпирический риск на выборке S объема n , d_{VC} — VC-размерность класса, $\delta \in (0, 1)$ — уровень доверия.

Для моделей с большим числом параметров VC-оценки становятся завышенными и перестают быть информативными. Поэтому для глубоких моделей машинного обучения VC-размерность сохраняет значение как формальное понятие статистической теории сложности, но не дает ответа на вопрос, почему обученная модель демонстрирует обобщающую способность. Проблема в том, что если две модели имеют одинаковую архитектуру, то их VC-размерность совпадает с точностью до порядка величины, однако после обучения они ведут себя по-разному. Следовательно, сама мощность класса не определяет тот объект, который фактически определяет качество модели.

Отдельно стоит упомянуть комбинаторный подход [4], который учитывает структуру выборки, а не только грубую оценку в худшем случае класса. В рамках этого подхода объектом исследования остается класс

гипотез, а не локальные геометрические характеристики решения в параметрическом пространстве.

Дальнейшим развитием теории сложности является радемахеровская сложность [5, 6], которая зависит от рассматриваемой выборки и учитывает геометрию данных [31, 32]. Для линейных моделей и некоторых специальных классов моделей радемахеровский подход дает более точные оценки. Данный подход применяется в глубоком обучении, поскольку оценки, зависящие от выборки, частично отделяют реально наблюдаемую сложность задачи от чисто комбинаторной мощности класса моделей [31, 32]. Современные исследования развивают радемахеровскую сложность в нескольких направлениях. Например, для обобщения классических оценок [47], а также оценок для сложных классов функций [46]. Различные обзорные работы, также подчеркивают, что радемахеровский подход для глубоких сетей связан не столько с поиском одной универсальной оценки, сколько с выявлением границ применимости моделей [42, 75]. Эмпирическая радемахеровская сложность класса $\mathcal{H}$ на выборке $S = (x_i)_{i=1}^n$ определяется выражением:

$$\hat{\mathfrak{R}}_S(\mathcal{H}) = \mathbb{E}_\sigma \left[\sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^n \sigma_i h(x_i) \right], \quad (2)$$

где $\sigma_i \in \{-1, +1\}$ — независимые случайные знаки Радемахера. Для липшицевых потерь оценка принимает вид:

$$R(h) \leq \hat{R}_S(h) + 2 \hat{\mathfrak{R}}_S(\mathcal{H}) + 3 \sqrt{\frac{\log(2/\delta)}{2n}}. \quad (3)$$

Разница радемахеровского подхода с VC-размерностью (1) состоит в том, что сложность зависит от заданной выборки, а не только от мощности класса функций. Однако радемахеровские оценки по-прежнему характеризуют весь класс функций, тогда как обучение исследует лишь малую и структурированную его часть. Поэтому две модели одинаковой архитектуры, обученные разными оптимизаторами или с разными инициализациями, имеют сопоставимые оценки сложности сверху, но могут заметно отличаться по обобщающей способности.

С практической точки зрения важно, что оценки, зависящие от выборки, часто оказываются содержательными лишь после введения дополнительных структурных ограничений, таких как ограничение норм весов, липшицевых ограничений или специальных возмущений данных. Ввод специальных ограничений делает решение не полным, так как для получения оценки требуется заранее встраивать в модель информацию

о геометрии решения, то есть теория частично объясняет обобщение, а частично уже предполагает его.

С другой стороны, достоинство РАС-байесовских оценок состоит в том, что они связывают обобщение не только с функцией, но и с распределением над параметрами. Базовые постановки этого подхода даны в [7, 8], а современные обобщения и обзорные интерпретации представлены в [22, 33, 34, 51]. На практике эти оценки оказываются содержательными среди современных оценок. Внутри этого же тренда появляются и более специализированные приложения, например анализ стохастического градиентного спуска в РАС-байесовском подходе [56, 66]. Однако и эти оценки зависят от выбора априорного распределения, формы возмущений и параметризации и потому задают скорее гибкую математическую рамку, чем прямое объяснение того, почему заданная модель глубокого обучения достигает решения с высокой обобщающей способностью. В одной из канонических форм РАС-байесовская оценка записывается как

$$\mathbb{E}_{\theta \sim Q} R(\theta) \leq \mathbb{E}_{\theta \sim Q} \hat{R}_S(\theta) + \sqrt{\frac{\text{KL}(Q \| P) + \log \frac{2\sqrt{n}}{\delta}}{2(n-1)}}, \quad (4)$$

где P — априорное распределение, Q — апостериорное, $\text{KL}(Q \| P)$ — дивергенция Кульбака–Лейблера. В (4) видно, что сложность перенесена с класса функций на стоимость перехода от априорного распределения к апостериорному через KL-дивергенцию.

Задача в рамках информационно-теоретического подхода ставится иначе. Он связывает ошибку обобщения с количеством информации, извлекаемой алгоритмом из данных [9]. В глубоких моделях машинного обучения при этом сохраняются трудности с получением численных оценок. Примером такого результата является оценка:

$$|\mathbb{E}[R(W) - \hat{R}_S(W)]| \leq \sqrt{\frac{2\sigma^2}{n} I(W; S)}, \quad (5)$$

где W — параметры после обучения, $I(W; S)$ — взаимная информация между параметрами и выборкой, σ — параметр субгауссовости потерь. Неравенство (5) отражает алгоритмическую сложность, то есть чем меньше информации, записанной в параметрах выборки, тем ниже ожидаемая ошибка обобщения.

В совокупности эти подходы указывают на методологический сдвиг. Теория постепенно отказывается от предположения, что существует одна универсальная мера сложности, одинаково адекватная для всех методов обучения. Вместо этого формируется набор частичных, но более предметных описаний: одни применимы к анализу класса функций, другие —

к стохастической динамике алгоритма, третьи — к локальной устойчивости уже найденного решения. В литературе этот сдвиг описывается как переход от единой статистической меры сложности к семейству взаимодополняющих описаний. Этот тезис последовательно проводится как в обзорных работах [43, 44, 48], так и в теоретических работах [70, 71, 73, 74].

Одним из важных направлений исследований теории сложности является теория устойчивости моделей машинного обучения [10, 11]. Данная теория занимает промежуточное положение между классической статистической теорией и современной алгоритмической постановкой вопроса. С одной стороны, устойчивость по-прежнему связана с обобщением через формальные оценки сверху. С другой стороны, устойчивость напрямую зависит от свойств алгоритма и потому анализирует реальный процесс обучения, а не только класс гипотез. Для глубокого обучения это связано с тем, что стохастический градиентный спуск является шумным и траекторно-зависимым алгоритмом. Устойчивость тесно связана с размером батча, длительностью обучения, кривизной ландшафта и типом регуляризации. Поэтому устойчивость оказывается полезной не столько как самостоятельная универсальная мера сложности, сколько как часть более широкой картины, связывающей оптимизацию и геометрию. В этой постановке теория устойчивости помогает понять, почему современные исследования все чаще рассматривают сложность как свойство процесса обучения модели, а не только класса функций модели. Она не решает проблему полностью, но указывает направление развития: от емкостных описаний к алгоритмозависимым описаниям. В канонической постановке для β -устойчивого алгоритма используется оценка вида

$$|R(A(S)) - \hat{R}_S(A(S))| \leq \beta, \quad (6)$$

где A — алгоритм обучения, S — обучающая выборка, β — параметр устойчивости. Запись (6) фиксирует формальную связь между устойчивостью алгоритма и ошибкой обобщения.

Ограниченность классических мер сложности проявляется не в том, что они неверны, а в том, что они описывают иной уровень абстракции по сравнению с тем, который требуется для анализа современных глубоких моделей. В перепараметризованных моделях важно не только то, насколько велик класс возможных функций, но и то, какое подмножество решений реально достижимо конкретным методом оптимизации, каковы локальные геометрические характеристики этих решений и какие индуктивные смещения встроены в архитектуру. Эти идеи проявились после работ, показавших, что глубокие сети способны идеально настраиваться на случайные метки, оставаясь при этом успешными на реальных

данных [30]. Отсюда следует вывод, что теория сложности для глубокого обучения не ограничивается контролем емкости класса гипотез. Она учитывает, по меньшей мере, три дополнительных аспекта: *алгоритм обучения, локальные геометрические характеристики решений и структуру данных, согласованную с архитектурой*. Это и мотивирует переход к современным геометрическим и ландшафтным подходам.

4. Ландшафт функции потерь и геометрические характеристики решений

Если классические меры оценивают пространство всех потенциально допустимых решений, то ландшафтные подходы ориентированы на анализ свойств локальных минимумов, к которым приводит обучение. Интерес к данному тренду усилился после работ по визуализации ландшафта потерь, показавших, что решения с высокой обобщающей способностью часто располагаются в более широких областях пространства параметров [12, 40, 62]. Идея плоских минимумов как возможного объяснения этого эффекта восходит к ранним работам по нейронным сетям [13]. В дальнейшем она была переосмыслена в терминах спектра гессиана, локальной устойчивости и неявной регуляризации алгоритмов обучения, а также в терминах нормализованных и параметризационно-устойчивых характеристик плоскости [57].

В данном тренде понятие сложности получает иную интерпретацию. Речь идет не о размере класса реализуемых функций, а о свойствах найденного решения в локальной геометрии. В этом смысле ландшафтная сложность оказывается ближе к эффективной сложности обучения, чем к абстрактной мощности параметрического семейства.

В этом направлении часто используется анализ гессиана функции потерь и его спектра. Эмпирические работы показывают типичную структуру спектра, где мало доминирующих направлений кривизны и много почти плоских направлений [14, 58, 60]. Это согласуется с гипотезой, что решения с высоким обобщением чаще лежат не в острых изолированных минимумах, а в более широких областях параметрического пространства. Локальная геометрия задается матрицей гессиана

$$H(\theta) = \nabla_{\theta}^2 \mathcal{L}(\theta), \quad (7)$$

а ее спектральные характеристики описываются собственными значениями

$$\lambda_1(\theta) \geq \lambda_2(\theta) \geq \dots \geq \lambda_p(\theta) \geq 0. \quad (8)$$

Важными характеристиками ландшафта являются: спектральный радиус $\lambda_{\max}(H)$ и след $\text{tr}(H) = \sum_i \lambda_i$, которые соответственно отражают максимальную локальную кривизну и интегральную кривизну в окрестности найденного решения.

С этой точки зрения плоские минимумы представляют интерес по двум причинам. Во-первых, они более устойчивы к малым возмущениям параметров и потому потенциально демонстрируют более стабильный перенос на новые данные. Во-вторых, они отражают не только свойства класса моделей, но и свойства траектории, посредством которой алгоритм достиг найденного решения. Поэтому связь между ландшафтом и обобщением тесно переплетается с анализом оптимизации.

Существуют и попытки формализовать эту интуицию в виде количественных мер. В частности, в недавних работах на основе гессиана предлагаются величины, характеризующие вариабельность кривизны по примерам и связывающие локальную геометрию с объемом выборки и устойчивостью обучения [15]. Значение таких подходов состоит в переводе обсуждения плоских и острых решений из качественной плоскости в количественную. В более широком контексте этот тренд включает также оценки, использующие представления о сложности через внутренние представления и предпочтение более простых функций в параметрическом отображении [63, 64].

Наряду с этим были предложены и алгоритмические методы, специально ориентированные на предпочтение широких минимумов. К таким работам относятся как анализ влияния размера батча и остроты минимума [37], так и методы, явно управляющие локальными геометрическими характеристиками решения, включая метод минимизации с учетом остроты (англ. Sharpness-Aware Minimization, SAM) и его более поздние обобщения на режим Edge of Stability (границы устойчивости) [38, 59]. Дополнительно есть работы, где предпочтение широких минимумов связывается с фильтрацией высокочастотных компонент стохастической динамики [67]. Это важно, поскольку влияние оптимизатора на локальные геометрические характеристики решения означает необходимость учитывать в теории сложности не только архитектуру модели, но и динамику алгоритма. Для SAM целевая функция имеет характерную мини-макс форму:

$$\min_{\theta} \max_{\|\varepsilon\|_2 \leq \rho} \mathcal{L}(\theta + \varepsilon), \quad (9)$$

где ρ задает радиус локального возмущения. Выражение (9) формализует идею избегания острых минимумов в явной оптимизационной постановке и связывает выбор алгоритма с геометрией получаемого решения.

На стыке геометрии и оптимизации проявляется различие между классической и современной постановкой вопроса. В статистической теории обучения алгоритм часто рассматривался как вторичный элемент. Сначала задавался класс гипотез, затем для него выводились общие оценки. В глубоком обучении ситуация принципиально иная. Один и тот же класс гипотез, обученный разными вариантами стохастического градиентного спуска, может приводить к решениям с различной локальной кривизной и различным качеством обобщения. Поэтому шум стохастической оптимизации, размер батча и свойства траектории становятся частью самой теории сложности, а не внешней технической деталью [10, 11].

С этой точки зрения, оптимизатор действует как механизм отбора решений внутри большого перепараметризованного пространства. Даже если класс допустимых функций чрезвычайно богат, реальная траектория SGD не исследует его равномерно. Вместо этого она систематически приводит к решениям с определенными геометрическими характеристиками, чувствительностью к возмущениям и структурой признаков. Следовательно, эффективная сложность оказывается результатом совместного действия архитектуры, данных и алгоритма оптимизации. Современные работы описывают этот механизм на разных уровнях: от диффузионной аппроксимации, в которой SGD экспоненциально предпочитает плоские минимумы [39], до анализа динамической устойчивости, границы устойчивости и детальных оценок устойчивости самого стохастического градиентного спуска [61, 65, 72].

Этот подход связывает локальный анализ гессиана с более глобальными вопросами стабильности обучения. Если решение лежит в области, устойчивой к шуму оптимизации и малым возмущениям параметров, то его свойства могут согласовываться с более высоким обобщением. Тем самым ландшафтный анализ связывает геометрию минимумов и алгоритмическую теорию устойчивости. Такая интерпретация согласуется с работами о фрактальной структуре стохастической оптимизации, где свойства траектории и свойства обобщения рассматриваются совместно [68].

В анализе плоских минимумов необходимо учитывать методологическую проблему. Геометрия в пространстве параметров не всегда напрямую соответствует геометрии в пространстве функций. Для нейронных сетей характерны перенормировки, симметрии слоев и другие преобразования параметров, которые могут заметно менять локальную кривизну без изменения реализуемой функции. Поэтому простая интерпретация вида «более плоский минимум означает более высокое обобщение» требует оговорок. Отсюда вытекают два следствия. Во-первых, ландшафтные меры нужно по возможности строить так, чтобы они были более устой-

чивы к параметризационным артефактам или хотя бы явно учитывали их наличие. Во-вторых, эмпирическая полезность таких мер не отменяет необходимости осторожной причинной интерпретации. Если плоскость зависит от выбора параметризации, то она, вероятно, является не самостоятельной причиной высокого обобщения, а наблюдаемым маркером более глубокой структуры. Практическое направление решения этой проблемы связано с инвариантными или квази-инвариантными мерами. В литературе используются нормализованные варианты остроты минимума, в которых локальная кривизна масштабируется параметрами слоя или нормы весов [57]. Указанные ограничения не отменяют использование ландшафтного подхода, а уточняют границы его применимости. В наиболее строгой постановке речь идет не о тезисе плоских минимумов вообще, а о связи обобщения с устойчивыми геометрическими свойствами решений, достигаемых конкретной динамикой обучения. Такая формулировка согласуется с современными эмпирическими данными и необходимостью учитывать архитектурную специфику.

Основное достоинство ландшафтного подхода состоит в том, что он описывает свойства уже найденного решения. Это позволяет частично компенсировать основной недостаток классических мер сложности — их слабую чувствительность к алгоритму и режиму обучения. Кроме того, геометрические характеристики естественно связываются с такими современными явлениями, как неявная регуляризация, влияние размера батча и различия между оптимизаторами [10, 11]. С другой стороны, у данного направления есть ряд недостатков. Во-первых, вычисление гессиана или даже его устойчивых спектральных характеристик остается ресурсоемкой задачей для больших моделей. Во-вторых, причинная интерпретация плоских минимумов далека от окончательной ясности: решения с высоким обобщением могут быть плоскими не потому, что плоскость сама по себе вызывает обобщение, а потому, что и плоскость, и хорошее обобщение являются следствиями более глубокой структуры данных и оптимизации. В-третьих, локальная геометрия фиксирует свойства конкретной точки, но не заменяет более широкую теорию методов обучения и масштабирования. Поэтому ландшафтный анализ следует понимать не как окончательную замену классических мер сложности, а как компонент более общей теории, в которой учитываются геометрия решения, алгоритм его поиска и структура данных.

5. Эмпирические оценки сложности и сходимости

Эмпирически показано, что простая монотонная связь между сложностью и ошибкой обобщения не выполняется. Это видно в классических работах по двойному спуску [16, 49], а также в последующих уточнениях для различных режимов и моделей [41, 45, 50]. Если в классических мерах сложности рост сложности модели после некоторой точки должен лишь ухудшать обобщение, то в реальности часто наблюдается другое. Сначала ошибка падает, затем возрастает, а затем снова уменьшается. Явление двойного спуска не образует завершенной теории, однако остается эмпирическим фактом. Оно показывает, что сложность в глубоком обучении не сводится к единственной скалярной величине вроде числа параметров или грубой мощности класса гипотез. Причем в разных режимах обучения одна и та же архитектура может функционировать по-разному, а качество обобщения определяется тем, в каком режиме обучения и при какой параметризации она работает. Для описания феномена двойного спуска используется кусочно-заданная форма тестовой ошибки

$$\mathcal{E}_{\text{test}}(m) \approx \begin{cases} \downarrow, & m < m_{\text{int}}, \\ \uparrow, & m \approx m_{\text{int}}, \\ \downarrow, & m > m_{\text{int}}, \end{cases} \quad (10)$$

где m — параметр сложности модели, m_{int} — интерполяционный порог. Запись (10) формализует немонотонность зависимости «сложность–ошибка».

В рамках режимной интерпретации двойного спуска обычно выделяются три фазы. До интерполяционного порога доминирует режим недообученной аппроксимации, и рост сложности чаще снижает смещение. Вблизи порога m_{int} усиливается чувствительность к шуму и нестабильности оптимизации, что приводит к локальному росту тестовой ошибки. После порога в сильно перепараметризованной области начинает доминировать режим селективной оптимизации, в котором алгоритм чаще находит решения с более благоприятной локальной геометрией.

Современная интерпретация этой проблемы часто формулируется в терминах двух режимов обучения. В достаточно широких нейросетевых моделях обучение может быть близко к линеаризованному режиму, где динамика хорошо описывается через нейронное касательное ядро [17, 36, 53]. В этом случае модель ведет себя почти как ядерный метод, а часть классических гарантий становится более применимой. В этом режиме используется линеаризованная динамика предсказаний вида

$$f_{t+1}(x) \approx f_t(x) - \eta K_{\theta_0}(x, X) \nabla_f \mathcal{L}(f_t(X), y), \quad (11)$$

где K_{θ_0} — НТК в точке инициализации θ_0 , η — шаг обучения. Формула (11) задает формальный переход к ядерной аппроксимации в широких сетях.

Однако на практике в моделях глубокого обучения доминирует другой режим — обучение признаков, в котором внутренняя репрезентация данных существенно меняется по ходу оптимизации. Этот режим в большей степени согласуется с современными эмпирическими результатами, но в меньшей степени поддается классическому анализу. Ограничения НТК-приближений и роль эволюции признаков обсуждаются в [35, 55], а также в более поздних работах [54, 76]. В таком режиме эффективная сложность определяется не только архитектурой, но и траекторией перестройки признакового пространства, а значит снова выводит нас к геометрии решения и динамике оптимизации. Режимы обучения служат связующим звеном между классической статистической теорией и ландшафтным анализом. Они позволяют объяснить, почему одна и та же модель может подчиняться различным схемам в зависимости от размера модели, данных и метода оптимизации.

Формально различие между режимами можно зафиксировать через динамику касательного ядра. В НТК-режиме выполняется приближение

$$\|K_{\theta_t} - K_{\theta_0}\| \ll \|K_{\theta_0}\|, \quad (12)$$

поэтому обучение близко к линеаризованной модели с почти фиксированными признаками. В режиме обучения признаков существенной является противоположная ситуация

$$\|K_{\theta_t} - K_{\theta_0}\| = \mathcal{O}(1), \quad (13)$$

и тогда ключевым объектом анализа становится дрейф представлений. Соотношения (12)–(13) задают математический критерий, почему одна и та же архитектура требует разных мер сложности в разных режимах обучения.

Режимный анализ важен не только как теоретическая классификация, но и как средство интерпретации экспериментальных результатов. Когда модель находится ближе к ядерному режиму, ее поведение оказывается сравнительно более предсказуемым, а роль оптимизации сводится к подбору коэффициентов в почти фиксированном признаковом пространстве. Когда же обучение протекает в режиме обучения признаков, ключевым объектом анализа становится эволюция представлений, а следовательно, сложность должна описывать уже не только выходную функцию, но и динамику внутренней репрезентации.

Здесь проявляется причина, по которой одна универсальная мера сложности не получает убедительного теоретического обоснования. Показатель, эффективный в почти линеаризованном режиме, может оказаться малоинформативным в режиме активного обучения признаков. Следовательно, корректнее говорить не о единственной численной сложности, а о семействе взаимодополняющих описаний, чья применимость зависит от фазы обучения и архитектурного режима. Режимы обучения рассматриваются как объяснительный слой между общими теоретическими оценками и практикой больших нейронных сетей.

Еще одно современное направление связано с эмпирическими законами масштабирования. Базовые результаты для языковых моделей даны в [18, 19], а последующие обобщения и поправки обсуждаются в [29, 69]. Эти результаты показывают, что качество моделей часто изменяется по степенным законам при росте числа параметров, объема данных и вычислительного бюджета. Соответствующие зависимости оказываются практически полезными для проектирования больших моделей, однако сами по себе не дают объяснения механизма обобщения. В литературе этот блок рассматривается как отдельный макроуровень описания, связывающий теорию обобщения с инженерной практикой построения больших моделей [43, 44, 48]. В упрощенной форме такие зависимости записываются как

$$\mathcal{L}(N, D, C) \approx \mathcal{L}_{\infty} + aN^{-\alpha} + bD^{-\beta} + cC^{-\gamma}, \quad (14)$$

где N — число параметров, D — объем данных, C — вычислительный бюджет, $\alpha, \beta, \gamma > 0$ — эмпирические показатели масштабирования. Формула (14) удобна как макроописание, но не задает сама по себе механизма, через который именно эти показатели возникают в разных режимах обучения.

Для трансформеров и диффузионных моделей наблюдается следующая картина. Эти архитектуры продемонстрировали масштабируемость, имеющую прикладное значение [20, 21]. В то же время для них отсутствует единая теория сложности, которая одновременно учитывала бы архитектурную специфику, оптимизацию, геометрию решения и режимы обучения. В этой постановке законы масштабирования задают описательный макроуровень, тогда как ландшафтные и режимные подходы ориентированы на объяснительный уровень. В работе [23] масштабирование связывается с анализом кривизны и полных спектральных характеристик гессиана для архитектуры трансформера. Статистико-аппроксимационные и архитектурно-специфические версии представлены в работах [24–26]. Эта траектория исследований соединяет макроэмпирические зависимости с локальной геометрией архитектурно-специфических моделей.

Для трансформеров одна из трудностей состоит в том, что их успешность определяется не только размером модели, но и специфическими индуктивными смещениями механизма внимания, глубиной сети, структурой токенизации и масштабом предварительного обучения. Поэтому перенос классических оценок обучаемости в этот контекст редко дает содержательное объяснение. На уровне практики законы масштабирования помогают выбирать баланс между параметрами, данными и вычислениями, но дают ограниченную информацию о том, почему архитектура механизма внимания оказывается столь эффективной в режиме обучения признаков.

Для большей детализации полезно разделять архитектурные механизмы по их влиянию на сложность. Для трансформеров это, прежде всего, механизм самовнимания, остаточные связи и нормализация слоев: вместе они стабилизируют оптимизацию глубоких стеков и меняют эффективную кривизну задачи обучения. На уровне анализа сложности это означает, что параметр N в законах масштабирования (14) агрегирует разные по роли степени свободы. Практически эквивалентный рост числа параметров может сопровождаться разной динамикой обобщения в зависимости от того, где добавляется емкость: в ширине блоков внимания, в числе слоев, в MLP-части или в словаре токенизации. Этот вывод согласуется с работами, где законы масштабирования выводятся из статистико-аппроксимационного анализа и из динамики обучения признаков [24, 25, 27].

Для диффузионных моделей проблема имеет иную структуру. Здесь качество зависит не только от выразительности аппроксимирующей сети, но и от самой многошаговой стохастической процедуры генерации из шума. Следовательно, сложность должна описывать не статическую функцию, а целую динамическую схему восстановления сигнала. Это демонстрирует ограниченность единой скалярной меры сложности. Даже если параметризация сети фиксирована, изменение режима шума, числа шагов и процедуры обучения способно существенно менять как стабильность, так и качество генерации [21].

Для диффузионных архитектур это особенно заметно: эффективная сложность определяется совместно параметризацией денойзера и расписанием шума. Поэтому даже при близких кривых законов масштабирования финальная генеративная способность может различаться из-за разных траекторий обучения по временным шагам и разной чувствительности к ошибкам на ранних и поздних уровнях шума [28].

Связь геометрического и масштабного уровней удобно описывать как трехуровневую цепочку: локальные метрики формируют режим обучения, а режим обучения определяет вид наблюдаемой кривой закона масштаби-

рования при росте N, D, C . Ключевое методологическое следствие состоит в том, что одинаковая макрокривая качества не гарантирует одинаковый механизм обобщения.

Из этого следует вывод, что теория сложности глубокого обучения не может ограничиться ни одним из этих методов по отдельности. Классические оценки задают базовый язык, ландшафтные меры описывают локальные геометрические характеристики решений, анализ режимов обучения объясняет качественные переходы между динамическими фазами, а законы масштабирования фиксируют крупномасштабные эмпирические закономерности.

6. Сравнительный анализ и открытые вопросы

Рассмотренные работы в табл. 2 дают сравнительную картину современных подходов к сложности в глубоком обучении. Ключевой вывод состоит в том, что разные меры сложности отвечают на разные исследовательские вопросы. Классические меры описывают вместимость класса гипотез. Вероятностные и информационные методы частично учитывают данные и алгоритм. Геометрические и ландшафтные методы описывают свойства найденного решения. Режимные и масштабные подходы фиксируют качественные переходы и крупномасштабные закономерности.

Видно, что сохраняется разрыв между глобальной сложностью класса моделей и эффективной сложностью реально найденного решения. Первая необходима для формальной теории обучаемости, однако в глубоких сетях она недостаточна для объяснения практического обобщения. Вторая пока не имеет общепринятого формального описания, при этом современные геометрические и алгоритмические подходы в основном ориентированы на ее анализ. Классическая теория отвечает на вопрос о том, когда обучение в принципе возможно, тогда как современная практика требует ответа на вопрос почему из множества допустимых решений обучение систематически приводит к подмножеству решений с благоприятными эмпирическими свойствами. Если первый вопрос относится к глобальной структуре класса гипотез, то второй — к селективности алгоритма и геометрии достижимых минимумов. Поэтому современные исследования движутся к многокомпонентному описанию сложности, а не к поиску единственной универсальной формулы.

Современная теория должна быть не только формально строгой, но и эмпирически релевантной. Иначе говоря, качество меры сложности сегодня определяется не только ее математической строгостью, но и тем, насколько она позволяет согласовать классические гарантии с реальными

Таблица 2. Сравнение подходов к анализу сложности моделей глубокого обучения

Подход	Что измеряет	Сильные стороны	Ограничения для глубокого обучения
VC- и PAC-подходы	Мощность класса гипотез в худшем случае	Строгие критерии обучаемости и ясный статистический язык	Классическая оценка сложности через мощность класса функций и гарантии обучаемости в вероятностной постановке; пессимистичны в перепараметризованных режимах, не учитывают алгоритм и найденное решение
Радемахеровские и нормовые оценки	Емкость класса функций, зависящая от данных	Более тонкие оценки сверху, чувствительность к выборке и нормам	Уточняют обобщение за счет геометрии выборки и ограничений на параметры; по-прежнему описывают весь класс, а не подмножество реально достижимых решений
PAC-байесовские и информационные границы	Связь обобщения с распределением параметров и информацией об обучении	Учитывают алгоритмический аспект и допускают адаптивный анализ	Формируют вероятностно-информационную рамку, связывающую обобщение с распределением параметров и информацией из данных; зависят от выбора априорных распределений, параметризации и часто остаются численно слабыми
Ландшафтные и гессиановые меры	Локальные геометрические характеристики решения	Отражают свойства минимумов, устойчивость и связь с оптимизацией	Геометрический анализ свойств найденного решения и его устойчивости к возмущениям; дороги вычислительно, причинная интерпретация плоскости и спектра не всегда однозначна
Режимы обучения и законы масштабирования	Качественные фазы обучения и крупномасштабные эмпирические зависимости	Описывают современную практику больших моделей	Описывают фазы обучения и закономерности качества при росте параметров, данных и вычислений; дают неполное объяснение механизма обобщения и требуют стыковки с геометрией и оптимизацией

режимами обучения больших сетей. Ландшафт и оптимизация в этом случае выступают как аналитические компоненты, так как позволяют анализировать сложность не только теоретически, но и применительно к решениям, возникающим в современных вычислительных режимах.

Таблица 3. Уровни объяснения и типы теоретических утверждений

Уровень объяснения	Тип утверждений	Типичный вопрос	Основной риск неверной интерпретации
Глобальный, статистический	Гарантии обучаемости, оценки сверху на ошибку	Когда обучение в принципе возможно?	Перенос оценок худшего случая на конкретную обученную модель без учета алгоритма
Алгоритмический	Устойчивость, неявная регуляризация, селективность траектории	Почему из многих решений выбираются именно эти?	Подмена причинного механизма постфактум корреляциями
Геометрический	Спектральные и ландшафтные характеристики найденного минимума	Какие свойства решения связаны с обобщением?	Игнорирование параметризационных инвариантностей и вычислительной чувствительности метрик
Эмпирический	Режимные переходы и законы масштабирования	Как меняется качество при росте ресурсов?	Объяснение механизма только по форме масштабной кривой

В табл. 3 представлены ключевые различия между различными направлениями исследований. Статистические оценки нередко интерпретируются так, будто они обязаны предсказывать локальную геометрию решения, а ландшафтные наблюдения интерпретируются так, как будто они автоматически дают универсальные гарантии обучаемости. Методологически корректно рассматривать эти уровни как взаимодополняющие. С этой точки зрения содержательный вклад обзора состоит не только в перечислении направлений, но и в реконструкции их совместимой архитектуры. Классические оценки ограничивают область допустимых объяснений, алгоритмические и геометрические подходы описывают механизм селекции решений внутри этой области, режимный и масштабный анализ тестируют устойчивость механизма при изменении масштаба моделей, данных и вычислений. Такая многоуровневая схема дает более строгий язык для формулировки открытых задач.

Из этого анализа выделяется четыре вывода. Во-первых, классические меры, основанные на емкости класса, не исчезают из современной теории, но они дают базовые ориентиры, а не окончательные оценки. Во-вторых, алгоритм обучения в глубоких сетях является частью самой сложности. В-третьих, геометрические свойства найденного решения сопоставимы

по роли со свойствами класса допустимых функций. В-четвертых, крупномасштабные эмпирические закономерности, включая двойной спуск и законы масштабирования, выступают не альтернативой строгой теории, а ее тестом. Современные обзорные и синтетические работы подтверждают многоуровневую картину, в которой неявная регуляризация, геометрия ландшафта и режимы масштабирования рассматриваются как взаимодополняющие части одного исследовательского тренда. Это видно в обзорных публикациях [43, 44, 48] и в работах, ориентированных на теоретический анализ [52, 70, 73].

Исходя из анализа, выделяются следующие направления исследований теории сложности моделей машинного обучения.

1. Эффективная сложность достижимых решений. Требуется теория, описывающая не только весь класс параметрически допустимых функций, но и подмножество решений, которое достигается обучением из заданных инициализаций при заданных алгоритмах. На этом уровне соединяются язык классической емкости класса и язык геометрии найденных минимумов.

2. Совместное описание оптимизации и ландшафта. Ландшафтные меры рассматриваются после обучения, а теория оптимизации — отдельно. Для глубокого обучения этого уже недостаточно. Требуется теория, в которой свойства спектра, плоскости минимумов, размер батча, шум и выбор оптимизатора описываются как части единого механизма формирования обобщающего решения.

3. Теория для современных больших архитектур. Трансформеры, диффузионные модели и другие современные архитектуры демонстрируют устойчивые эмпирические закономерности, однако соответствующая теория сложности остается фрагментарной. Особенно важна задача согласовать законы масштабирования с более локальными геометрическими и динамическими характеристиками.

4. Согласование уровней объяснения и критериев верификации. Открытой проблемой остается построение единого метода сопоставления результатов разных типов. От статистических оценок до геометрических и режимных индикаторов. Для прогресса необходимы работы, в которых для одной и той же постановки задачи одновременно фиксируются глобальные оценочные характеристики, алгоритмические показатели

устойчивости, ландшафтные метрики и эмпирические зависимости качества.

5. Переход от ретроспективного объяснения к прогностической теории. Значительная часть современных методов успешно объясняет уже наблюдаемые эффекты, но в меньшей степени дает проверяемые прогнозы до проведения полного вычислительного эксперимента. Соответственно, одним из направлений дальнейших исследований является разработка предсказательных критериев, позволяющих на ранних этапах обучения оценивать вероятный класс обобщающего решения и ожидаемую чувствительность к изменению масштаба. Это согласуется с результатами, где динамика спектра гессiana в раннем и среднем обучении коррелирует с последующей устойчивостью решения [58, 60]. Она также согласуется с работами, связывающими стабильность траектории SGD и итоговое обобщение [61, 65].

Остается открытым вопрос о форме будущей объединяющей теории. Вероятно, она не будет одной формулой или одной универсальной оценкой. Более обоснованным представляется многоуровневый подход: на верхнем уровне — укрупненные законы масштабирования и фазовые переходы, на среднем — режимы обучения и роль оптимизатора, на локальном — геометрия минимумов, спектр гессiana и устойчивость. Такое расслоение в большей степени соответствует наблюдаемой практике глубокого обучения, чем попытка свести все эффекты к одной численной характеристике. Эти проблемы показывают, что тренд развития теории сложности в глубоком обучении лежит не в отказе от классических результатов, а в их переинтерпретации и расширении. Классическая теория остается необходимым основанием, но ее нужно дополнять анализом эффективной геометрии решений и алгоритмических режимов.

7. Заключение

Теория сложности моделей глубокого обучения смещается от анализа абстрактной мощности класса гипотез к анализу эффективной сложности реально достигаемых решений. Статья опирается на воспроизводимую процедуру построения корпуса публикаций и сопоставление классических, вероятностно-информационных, ландшафтных и режимно-масштабных линий исследований. Классические меры — VC-размерность, радемахеровские оценки, PAC-байесовские и информационные границы — остаются фундаментальной частью языка теории обучения, однако в перепараметризованных режимах уже не дают единственного объяснения обобщающей

способности. Современные подходы описывают сложность через совместный анализ геометрии ландшафта, динамики оптимизации, режимов обучения и масштабных закономерностей. Ключевой вывод состоит в том, что архитектура, данные и алгоритм обучения должны анализироваться совместно. Только в этом случае теория сохраняет и формальную строгость, и эмпирическую релевантность для современных больших моделей.

Список литературы

- [1] V. N. Vapnik, A. Y. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities”, *Theory of Probability and its Applications*, **16:2** (1971), 264–280. DOI: 10.1137/1116025.
- [2] L. G. Valiant, “A theory of the learnable”, *Communications of the ACM*, **27:11** (1984), 1134–1142. DOI: 10.1145/1968.1972.
- [3] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
- [4] К. В. Воронцов, “Комбинаторный подход к оценке качества обучаемых алгоритмов”, *Математические вопросы кибернетики*, **13** (2004), 5–51.
- [5] P. L. Bartlett, S. Mendelson, “Rademacher and Gaussian complexities: Risk bounds and structural results”, *Journal of Machine Learning Research*, **3** (2002), 463–482.
- [6] V. Koltchinskii, “Rademacher penalties and structural risk minimization”, *IEEE Transactions on Information Theory*, **47:5** (2001), 1902–1914. DOI: 10.1109/18.930926.
- [7] D. A. McAllester, “Some PAC-Bayesian theorems”, *Machine Learning*, **37:3** (1999), 355–363. DOI: 10.1023/A:1007618624809.
- [8] G. K. Dziugaite, D. M. Roy, “Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data”, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.
- [9] A. Xu, M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms”, *Advances in Neural Information Processing Systems*, **30**, 2017.
- [10] O. Bousquet, A. Elisseeff, “Stability and generalization”, *Journal of Machine Learning Research*, **2** (2002), 499–526.
- [11] M. Hardt, B. Recht, Y. Singer, “Train faster, generalize better: stability of stochastic gradient descent”, *Proceedings of the 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, **48**, 2016, 1225–1234.
- [12] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, “Visualizing the Loss Landscape of Neural Nets”, *Advances in Neural Information Processing Systems*, **31**, 2018, 6389–6399.
- [13] S. Hochreiter, J. Schmidhuber, “Flat minima”, *Neural Computation*, **9:1** (1997), 1–42. DOI: 10.1162/neco.1997.9.1.1.
- [14] L. Sagun, L. Bottou, Y. LeCun, “Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond”, arXiv: 1611.07476.

- [15] N. S. Kiselev, A. V. Grabovoy, “Unraveling the Hessian: A Key to Smooth Convergence in Loss Function Landscapes”, *Doklady Mathematics*, **110**:Suppl. 1 (2024), S49–S61. DOI: 10.1134/S1064562424601987.
- [16] M. Belkin, D. Hsu, S. Ma, S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off”, *Proceedings of the National Academy of Sciences*, **116**:32 (2019), 15849–15854. DOI: 10.1073/pnas.1903070116.
- [17] A. Jacot, F. Gabriel, C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks”, *Advances in Neural Information Processing Systems*, **31**, 2018.
- [18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, “Scaling laws for neural language models”, arXiv: 2001.08361.
- [19] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, L. Sifre, “Training Compute-Optimal Large Language Models”, *Advances in Neural Information Processing Systems*, **35**, 2022, 30016–30030. DOI: 10.52202/068431–2176.
- [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need”, *Advances in Neural Information Processing Systems*, **30**, 2017, 5998–6008.
- [21] J. Ho, A. Jain, P. Abbeel, “Denoising diffusion probabilistic models”, *Advances in Neural Information Processing Systems*, **33**, 2020, 6840–6851.
- [22] F. Hellström, G. Durisi, B. Guedj, M. Raginsky, “Generalization Bounds: Perspectives from Information Theory and PAC-Bayes”, *Foundations and Trends in Machine Learning*, **18**:1 (2025), 1–223. DOI: 10.1561/22000000112.
- [23] E. Petrov, N. Kiselev, V. Meshkov, A. Grabovoy, “Closing the Curvature Gap: Full Transformer Hessians and Their Implications for Scaling Laws”, arXiv: 2504.12516.
- [24] A. Havrilla, W. Liao, “Understanding Scaling Laws with Statistical and Approximation Theory for Transformer Neural Networks on Intrinsically Low-dimensional Data”, arXiv: 2411.06646.
- [25] M. K. Pasupuleti, “Scaling Laws for Transformers on Low-Dimensional Data: A Statistical and Approximation Theory Perspective”, *International Journal of Academic and Industrial Research Innovations*, **2**:1 (2022), 191–202. DOI: 10.62311/nesx/rp-5-01-2022.
- [26] M. Mayaki, “Generalization and Scaling Laws for Mixture-of-Experts Transformers”, arXiv: 2604.09175.
- [27] Z. Yao, R. Wu, T. Gao, “Understanding Scaling Laws in Deep Neural Networks via Feature Learning Dynamics”, arXiv: 2512.21075.
- [28] Z. Liang, H. He, C. Yang, B. Dai, “Scaling Laws For Diffusion Transformers”, arXiv: 2410.08184.
- [29] N. Sardana, J. Portes, S. Doubov, J. Frankle, “Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws”, arXiv: 2401.00448.
- [30] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, “Understanding deep learning (still) requires rethinking generalization”, *Communications of the ACM*, **64**:3 (2021), 107–115. DOI: 10.1145/3446776.

- [31] L. V. Truong, “On Rademacher Complexity-based Generalization Bounds for Deep Learning”, arXiv: 2208.04284.
- [32] S. Sachs, T. van Erven, L. Hodgkinson, R. Khanna, U. Şimşekli, “Generalization Guarantees via Algorithm-dependent Rademacher Complexity”, arXiv: 2307.02501.
- [33] W. Zhou, V. Veitch, M. Austern, R. Adams, P. Orbanz, “Non-Vacuous Generalization Bounds at the ImageNet Scale: A PAC-Bayesian Compression Approach”, arXiv: 1804.05862.
- [34] L. Hodgkinson, C. Heide, R. Salomone, F. Roosta, M. Mahoney, “A PAC-Bayesian Perspective on the Interpolating Information Criterion”, arXiv: 2311.07013.
- [35] S. Arora, S. Du, W. Hu, Z. Li, R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks”, *Proceedings of the 36th International Conference on Machine Learning*, 2019, 322–332.
- [36] B. Adlam, J. Pennington, “The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization”, arXiv: 2008.06786.
- [37] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. Tang, “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima”, arXiv: 1609.04836.
- [38] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, “Sharpness-Aware Minimization for Efficiently Improving Generalization”, *International Conference on Learning Representations*, 2021, arXiv: 2010.01412.
- [39] Z. Xie, I. Sato, M. Sugiyama, “A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima”, arXiv: 2002.03495.
- [40] L. Wu, Z. Zhu, W. E, “Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes”, arXiv: 1706.10239.
- [41] S. Mei, A. Montanari, “The generalization error of random features regression: Precise asymptotics and double descent curve”, arXiv: 1908.05355.
- [42] G. Valle-Perez, A. Louis, “Generalization bounds for deep learning”, arXiv: 2012.04115.
- [43] X. Hu, L. Chu, J. Pei, W. Liu, J. Bian, “Model Complexity of Deep Learning: A Survey”, arXiv: 2103.05127.
- [44] F. He, D. Tao, “Recent advances in deep learning theory”, arXiv: 2012.10931.
- [45] M. Advani, A. Saxe, H. Sompolinsky, “High-dimensional dynamics of generalization error in neural networks”, *Neural Networks*, **132** (2020), 428–446. DOI: 10.1016/j.neunet.2020.08.022.
- [46] V. Cherkassky, X. Shao, F. Mulier, V. Vapnik, “Model complexity control for regression using VC generalization bounds”, *IEEE Transactions on Neural Networks*, **10**:5 (1999), 1075–1089. DOI: 10.1109/72.788648.
- [47] A. Hole, “Vapnik-Chervonenkis Generalization Bounds for Real Valued Neural Networks”, *Neural Computation*, **8**:6 (1996), 1277–1299. DOI: 10.1162/neco.1996.8.6.1277.
- [48] Y. Dar, V. Muthukumar, R. Baraniuk, “A Farewell to the Bias-Variance Trade-off? An Overview of the Theory of Overparameterized Machine Learning”, arXiv: 2109.02355.

- [49] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, I. Sutskever, “Deep double descent: Where bigger models and more data hurt”, *Journal of Statistical Mechanics: Theory and Experiment*, **2021**:12 (2021), 124003. DOI: 10.1088/1742-5468/ac3a74.
- [50] E. H. Lee, V. Cherkassky, “Understanding Double Descent Using VC-Theoretical Framework”, *IEEE Transactions on Neural Networks and Learning Systems*, **35**:12 (2024), 18838–18847. DOI: 10.1109/TNNLS.2024.3388873.
- [51] B. Ghojogh, A. Ghodsi, *PAC Learnability and Information Bottleneck in Deep Learning: Tutorial and Survey*, HAL preprint hal-04705310, 2024.
- [52] Z. Johannsen, “Implicit Regularization and Generalization in Overparameterized Neural Networks”, arXiv: 2604.07603.
- [53] S. Vakili, M. Bromberg, J. Garcia, D.-s. Shiu, A. Bernacchia, “Uniform Generalization Bounds for Overparameterized Neural Networks”, arXiv: 2109.06099.
- [54] G. Bachmann, T. Hofmann, A. Lucchi, “Generalization Through The Lens Of Leave-One-Out Error”, arXiv: 2203.03443.
- [55] T. Suzuki, S. Akiyama, “Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods”, arXiv: 2012.03224.
- [56] K. Sakamoto, I. Sato, “Analyzing Lottery Ticket Hypothesis from PAC-Bayesian Theory Perspective”, arXiv: 2205.07320.
- [57] Y. Tsuzuku, I. Sato, M. Sugiyama, “Normalized Flat Minima: Exploring Scale Invariant Definition of Flat Minima for Neural Networks using PAC-Bayesian Analysis”, arXiv: 1901.04653.
- [58] M. Wei, D. Schwab, “How noise affects the Hessian spectrum in overparameterized neural networks”, arXiv: 1910.00195.
- [59] P. Long, P. Bartlett, “Sharpness-Aware Minimization and the Edge of Stability”, arXiv: 2309.12488.
- [60] Z. Xie, Q.-Y. Tang, Y. Cai, M. Sun, P. Li, “On the Power-Law Hessian Spectrums in Deep Learning”, arXiv: 2201.13011.
- [61] L. Wu, W. Su, “The Implicit Regularization of Dynamical Stability in Stochastic Gradient Descent”, arXiv: 2305.17490.
- [62] P. Zhou, J. Feng, “The Landscape of Deep Learning Algorithms”, arXiv: 1705.07038.
- [63] P. Natekar, M. Sharma, “Representation Based Complexity Measures for Predicting Generalization in Deep Learning”, arXiv: 2012.02775.
- [64] G. Valle-Perez, C. Camargo, A. Louis, “Deep learning generalizes because the parameter-function map is biased towards simple functions”, arXiv: 1805.08522.
- [65] Y. Lei, Y. Ying, “Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent”, arXiv: 2006.08157.
- [66] B. London, “A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent”, *Advances in Neural Information Processing Systems*, **30**, 2017.
- [67] D. Bisla, J. Wang, A. Choromanska, “Low-Pass Filtering SGD for Recovering Flat Optima in the Deep Learning Optimization Landscape”, arXiv: 2201.08025.
- [68] A. Camuto, G. Deligiannidis, M. Erdogdu, M. Gurbuzbalaban, U. Simsekli, L. Zhu, “Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms”, arXiv: 2106.04881.

- [69] J. Wang, A. Choromanska, “A Survey of Optimization Methods for Training DL Models: Theoretical Perspective on Convergence and Generalization”, arXiv: 2501.14458.
- [70] V. Nagarajan, “Explaining generalization in deep learning: progress and fundamental limits”, arXiv: 2110.08922.
- [71] K. Kawaguchi, L. Kaelbling, Y. Bengio, “Generalization in Deep Learning”, arXiv: 1710.05468.
- [72] S. Arora, Z. Li, A. Panigrahi, “Understanding Gradient Descent on Edge of Stability in Deep Learning”, arXiv: 2205.09745.
- [73] T. Poggio, A. Banburski, Q. Liao, “Theoretical Issues in Deep Networks: Approximation, Optimization and Generalization”, arXiv: 1908.09375.
- [74] T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, H. Mhaskar, “Theory of Deep Learning III: explaining the non-overfitting puzzle”, arXiv: 1801.00173.
- [75] N. Golowich, A. Rakhlin, O. Shamir, “Size-Independent Sample Complexity of Neural Networks”, *Information and Inference: A Journal of the IMA*, **9**:2 (2020), 473–504. DOI: 10.1093/imaiai/iaz007.
- [76] N. Vyas, Y. Bansal, P. Nakkiran, “Limitations of the NTK for Understanding Generalization in Deep Learning”, arXiv: 2206.10012 .

Статья поступила 5 мая 2026 г.

From VC Dimension to Scaling Laws: The Evolution of Theoretical Approaches to Complexity and Generalization

A. V. Grabovoy

This paper presents a review of the current state of research on theoretical approaches to complexity in deep learning and their relation to generalization. Methodologically, the review is structured as a scoping review: it documents bibliographic sources, deduplication and semi-automated corpus narrowing, expert screening, and the composition of a final analytical publication set. The central question is why classical complexity measures, despite their foundational role in statistical learning theory, are insufficient for explaining the behavior of modern overparameterized neural networks. VC dimension, Rademacher complexity, PAC-Bayesian bounds, and information-theoretic approaches remain indispensable as a baseline analytical language, yet they mainly characterize worst-case hypothesis-class capacity and only partially reflect the geometry of the learned solution, the optimization trajectory, and architectural inductive biases. The main focus is therefore placed on contemporary approaches based on the loss landscape, Hessian spectra, optimization methods, and empirical scaling laws. The review develops a synthetic interpretation of the field according to which a useful theory of complexity for deep learning should describe not only the expressive

capacity of a class, but also the effective complexity of the solutions actually reached by concrete algorithms on concrete data.

Keywords: complexity theory, generalization, deep learning, VC dimension, Rademacher complexity, loss landscape, Hessian, double descent, scaling laws.

References

- [1] V. N. Vapnik, A. Y. Chervonenkis, “On the uniform convergence of relative frequencies of events to their probabilities”, *Theory of Probability and its Applications*, **16**:2 (1971), 264–280. DOI: 10.1137/1116025.
- [2] L. G. Valiant, “A theory of the learnable”, *Communications of the ACM*, **27**:11 (1984), 1134–1142. DOI: 10.1145/1968.1972.
- [3] V. N. Vapnik, *Statistical Learning Theory*, John Wiley & Sons, 1998.
- [4] K. V. Vorontsov, “A combinatorial approach to assessing the quality of learning algorithms”, *Mathematical Questions of Cybernetics*, **13** (2004), 5–51 (In Russian).
- [5] P. L. Bartlett, S. Mendelson, “Rademacher and Gaussian complexities: Risk bounds and structural results”, *Journal of Machine Learning Research*, **3** (2002), 463–482.
- [6] V. Koltchinskii, “Rademacher penalties and structural risk minimization”, *IEEE Transactions on Information Theory*, **47**:5 (2001), 1902–1914. DOI: 10.1109/18.930926.
- [7] D. A. McAllester, “Some PAC-Bayesian theorems”, *Machine Learning*, **37**:3 (1999), 355–363. DOI: 10.1023/A:1007618624809.
- [8] G. K. Dziugaite, D. M. Roy, “Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data”, *Proceedings of the Thirty-Third Conference on Uncertainty in Artificial Intelligence*, 2017.
- [9] A. Xu, M. Raginsky, “Information-theoretic analysis of generalization capability of learning algorithms”, *Advances in Neural Information Processing Systems*, **30**, 2017.
- [10] O. Bousquet, A. Elisseeff, “Stability and generalization”, *Journal of Machine Learning Research*, **2** (2002), 499–526.
- [11] M. Hardt, B. Recht, Y. Singer, “Train faster, generalize better: stability of stochastic gradient descent”, *Proceedings of the 33rd International Conference on Machine Learning*, Proceedings of Machine Learning Research, **48**, 2016, 1225–1234.

-
- [12] H. Li, Z. Xu, G. Taylor, C. Studer, T. Goldstein, “Visualizing the Loss Landscape of Neural Nets”, *Advances in Neural Information Processing Systems*, **31**, 2018, 6389–6399.
 - [13] S. Hochreiter, J. Schmidhuber, “Flat minima”, *Neural Computation*, **9**:1 (1997), 1–42. DOI: 10.1162/neco.1997.9.1.1.
 - [14] L. Sagun, L. Bottou, Y. LeCun, “Eigenvalues of the Hessian in Deep Learning: Singularity and Beyond”, arXiv: 1611.07476.
 - [15] N. S. Kiselev, A. V. Grabovoy, “Unraveling the Hessian: A Key to Smooth Convergence in Loss Function Landscapes”, *Doklady Mathematics*, **110**:Suppl. 1 (2024), S49–S61. DOI: 10.1134/S1064562424601987.
 - [16] M. Belkin, D. Hsu, S. Ma, S. Mandal, “Reconciling modern machine-learning practice and the classical bias–variance trade-off”, *Proceedings of the National Academy of Sciences*, **116**:32 (2019), 15849–15854. DOI: 10.1073/pnas.1903070116.
 - [17] A. Jacot, F. Gabriel, C. Hongler, “Neural tangent kernel: Convergence and generalization in neural networks”, *Advances in Neural Information Processing Systems*, **31**, 2018.
 - [18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, D. Amodei, “Scaling laws for neural language models”, arXiv: 2001.08361.
 - [19] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, O. Vinyals, J. W. Rae, L. Sifre, “Training Compute-Optimal Large Language Models”, *Advances in Neural Information Processing Systems*, **35**, 2022, 30016–30030. DOI: 10.52202/068431–2176.
 - [20] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, “Attention is all you need”, *Advances in Neural Information Processing Systems*, **30**, 2017, 5998–6008.
 - [21] J. Ho, A. Jain, P. Abbeel, “Denoising diffusion probabilistic models”, *Advances in Neural Information Processing Systems*, **33**, 2020, 6840–6851.
 - [22] F. Hellström, G. Durisi, B. Guedj, M. Raginsky, “Generalization Bounds: Perspectives from Information Theory and PAC-Bayes”, *Foundations and Trends in Machine Learning*, **18**:1 (2025), 1–223. DOI: 10.1561/22000000112.
 - [23] E. Petrov, N. Kiselev, V. Meshkov, A. Grabovoy, “Closing the Curvature Gap: Full Transformer Hessians and Their Implications for Scaling Laws”, arXiv: 2504.12516.
 - [24] A. Havrilla, W. Liao, “Understanding Scaling Laws with Statistical and Approximation Theory for Transformer Neural Networks on Intrinsically Low-dimensional Data”, arXiv: 2411.06646.

- [25] M. K. Pasupuleti, “Scaling Laws for Transformers on Low-Dimensional Data: A Statistical and Approximation Theory Perspective”, *International Journal of Academic and Industrial Research Innovations*, **2:1** (2022), 191–202. DOI: 10.62311/nesx/rp-5-01-2022.
- [26] M. Mayaki, “Generalization and Scaling Laws for Mixture-of-Experts Transformers”, arXiv: 2604.09175.
- [27] Z. Yao, R. Wu, T. Gao, “Understanding Scaling Laws in Deep Neural Networks via Feature Learning Dynamics”, arXiv: 2512.21075.
- [28] Z. Liang, H. He, C. Yang, B. Dai, “Scaling Laws For Diffusion Transformers”, arXiv: 2410.08184.
- [29] N. Sardana, J. Portes, S. Doubov, J. Frankle, “Beyond Chinchilla-Optimal: Accounting for Inference in Language Model Scaling Laws”, arXiv: 2401.00448.
- [30] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, “Understanding deep learning (still) requires rethinking generalization”, *Communications of the ACM*, **64:3** (2021), 107–115. DOI: 10.1145/3446776.
- [31] L. V. Truong, “On Rademacher Complexity-based Generalization Bounds for Deep Learning”, arXiv: 2208.04284.
- [32] S. Sachs, T. van Erven, L. Hodgkinson, R. Khanna, U. Şimşekli, “Generalization Guarantees via Algorithm-dependent Rademacher Complexity”, arXiv: 2307.02501.
- [33] W. Zhou, V. Veitch, M. Austern, R. Adams, P. Orbanz, “Non-Vacuous Generalization Bounds at the ImageNet Scale: A PAC-Bayesian Compression Approach”, arXiv: 1804.05862.
- [34] L. Hodgkinson, C. Heide, R. Salomone, F. Roosta, M. Mahoney, “A PAC-Bayesian Perspective on the Interpolating Information Criterion”, arXiv: 2311.07013.
- [35] S. Arora, S. Du, W. Hu, Z. Li, R. Wang, “Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks”, *Proceedings of the 36th International Conference on Machine Learning*, 2019, 322–332.
- [36] B. Adlam, J. Pennington, “The Neural Tangent Kernel in High Dimensions: Triple Descent and a Multi-Scale Theory of Generalization”, arXiv: 2008.06786.
- [37] N. S. Keskar, D. Mudigere, J. Nocedal, M. Smelyanskiy, P. Tang, “On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima”, arXiv: 1609.04836.
- [38] P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, “Sharpness-Aware Minimization for Efficiently Improving Generalization”, *International Conference on Learning Representations*, 2021, arXiv: 2010.01412.
- [39] Z. Xie, I. Sato, M. Sugiyama, “A Diffusion Theory For Deep Learning Dynamics: Stochastic Gradient Descent Exponentially Favors Flat Minima”, arXiv: 2002.03495.

-
- [40] L. Wu, Z. Zhu, W. E, “Towards Understanding Generalization of Deep Learning: Perspective of Loss Landscapes”, arXiv: 1706.10239.
 - [41] S. Mei, A. Montanari, “The generalization error of random features regression: Precise asymptotics and double descent curve”, arXiv: 1908.05355.
 - [42] G. Valle-Perez, A. Louis, “Generalization bounds for deep learning”, arXiv: 2012.04115.
 - [43] X. Hu, L. Chu, J. Pei, W. Liu, J. Bian, “Model Complexity of Deep Learning: A Survey”, arXiv: 2103.05127.
 - [44] F. He, D. Tao, “Recent advances in deep learning theory”, arXiv: 2012.10931.
 - [45] M. Advani, A. Saxe, H. Sompolinsky, “High-dimensional dynamics of generalization error in neural networks”, *Neural Networks*, **132** (2020), 428–446. DOI: 10.1016/j.neunet.2020.08.022.
 - [46] V. Cherkassky, X. Shao, F. Mulier, V. Vapnik, “Model complexity control for regression using VC generalization bounds”, *IEEE Transactions on Neural Networks*, **10**:5 (1999), 1075–1089. DOI: 10.1109/72.788648.
 - [47] A. Hole, “Vapnik-Chervonenkis Generalization Bounds for Real Valued Neural Networks”, *Neural Computation*, **8**:6 (1996), 1277–1299. DOI: 10.1162/neco.1996.8.6.1277.
 - [48] Y. Dar, V. Muthukumar, R. Baraniuk, “A Farewell to the Bias-Variance Trade-off? An Overview of the Theory of Overparameterized Machine Learning”, arXiv: 2109.02355.
 - [49] P. Nakkiran, G. Kaplun, Y. Bansal, T. Yang, B. Barak, I. Sutskever, “Deep double descent: Where bigger models and more data hurt”, *Journal of Statistical Mechanics: Theory and Experiment*, **2021**:12 (2021), 124003. DOI: 10.1088/1742-5468/ac3a74.
 - [50] E. H. Lee, V. Cherkassky, “Understanding Double Descent Using VC-Theoretical Framework”, *IEEE Transactions on Neural Networks and Learning Systems*, **35**:12 (2024), 18838–18847. DOI: 10.1109/TNNLS.2024.3388873.
 - [51] B. Ghogho, A. Ghodsi, *PAC Learnability and Information Bottleneck in Deep Learning: Tutorial and Survey*, HAL preprint hal-04705310, 2024.
 - [52] Z. Johannsen, “Implicit Regularization and Generalization in Overparameterized Neural Networks”, arXiv: 2604.07603.
 - [53] S. Vakili, M. Bromberg, J. Garcia, D.-s. Shiu, A. Bernacchia, “Uniform Generalization Bounds for Overparameterized Neural Networks”, arXiv: 2109.06099.
 - [54] G. Bachmann, T. Hofmann, A. Lucchi, “Generalization Through The Lens Of Leave-One-Out Error”, arXiv: 2203.03443.
 - [55] T. Suzuki, S. Akiyama, “Benefit of deep learning with non-convex noisy gradient descent: Provable excess risk bound and superiority to kernel methods”, arXiv: 2012.03224.

-
- [56] K. Sakamoto, I. Sato, “Analyzing Lottery Ticket Hypothesis from PAC-Bayesian Theory Perspective”, arXiv: 2205.07320.
 - [57] Y. Tsuzuku, I. Sato, M. Sugiyama, “Normalized Flat Minima: Exploring Scale Invariant Definition of Flat Minima for Neural Networks using PAC-Bayesian Analysis”, arXiv: 1901.04653.
 - [58] M. Wei, D. Schwab, “How noise affects the Hessian spectrum in overparameterized neural networks”, arXiv: 1910.00195.
 - [59] P. Long, P. Bartlett, “Sharpness-Aware Minimization and the Edge of Stability”, arXiv: 2309.12488.
 - [60] Z. Xie, Q.-Y. Tang, Y. Cai, M. Sun, P. Li, “On the Power-Law Hessian Spectrums in Deep Learning”, arXiv: 2201.13011.
 - [61] L. Wu, W. Su, “The Implicit Regularization of Dynamical Stability in Stochastic Gradient Descent”, arXiv: 2305.17490.
 - [62] P. Zhou, J. Feng, “The Landscape of Deep Learning Algorithms”, arXiv: 1705.07038.
 - [63] P. Natekar, M. Sharma, “Representation Based Complexity Measures for Predicting Generalization in Deep Learning”, arXiv: 2012.02775.
 - [64] G. Valle-Perez, C. Camargo, A. Louis, “Deep learning generalizes because the parameter-function map is biased towards simple functions”, arXiv: 1805.08522.
 - [65] Y. Lei, Y. Ying, “Fine-Grained Analysis of Stability and Generalization for Stochastic Gradient Descent”, arXiv: 2006.08157.
 - [66] B. London, “A PAC-Bayesian Analysis of Randomized Learning with Application to Stochastic Gradient Descent”, *Advances in Neural Information Processing Systems*, **30**, 2017.
 - [67] D. Bisla, J. Wang, A. Choromanska, “Low-Pass Filtering SGD for Recovering Flat Optima in the Deep Learning Optimization Landscape”, arXiv: 2201.08025.
 - [68] A. Camuto, G. Deligiannidis, M. Erdogdu, M. Gurbuzbalaban, U. Simsekli, L. Zhu, “Fractal Structure and Generalization Properties of Stochastic Optimization Algorithms”, arXiv: 2106.04881.
 - [69] J. Wang, A. Choromanska, “A Survey of Optimization Methods for Training DL Models: Theoretical Perspective on Convergence and Generalization”, arXiv: 2501.14458.
 - [70] V. Nagarajan, “Explaining generalization in deep learning: progress and fundamental limits”, arXiv: 2110.08922.
 - [71] K. Kawaguchi, L. Kaelbling, Y. Bengio, “Generalization in Deep Learning”, arXiv: 1710.05468.
 - [72] S. Arora, Z. Li, A. Panigrahi, “Understanding Gradient Descent on Edge of Stability in Deep Learning”, arXiv: 2205.09745.

-
- [73] T. Poggio, A. Banburski, Q. Liao, “Theoretical Issues in Deep Networks: Approximation, Optimization and Generalization”, arXiv: 1908.09375.
 - [74] T. Poggio, K. Kawaguchi, Q. Liao, B. Miranda, L. Rosasco, X. Boix, J. Hidary, H. Mhaskar, “Theory of Deep Learning III: explaining the non-overfitting puzzle”, arXiv: 1801.00173.
 - [75] N. Golowich, A. Rakhlin, O. Shamir, “Size-Independent Sample Complexity of Neural Networks”, *Information and Inference: A Journal of the IMA*, **9:2** (2020), 473–504. DOI: 10.1093/imaiai/iaz007.
 - [76] N. Vyas, Y. Bansal, P. Nakkiran, “Limitations of the NTK for Understanding Generalization in Deep Learning”, arXiv: 2206.10012.

Received on May 5, 2026