ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ
«ВЫСШАЯ ШКОЛА ЭКОНОМИКИ»

На правах рукописи

Кодрян Максим Станиславович

ДИНАМИКА ОБУЧЕНИЯ И ЛАНДШАФТ ФУНКЦИИ ПОТЕРЬ НЕЙРОННЫХ СЕТЕЙ С МАСШТАБНО-ИНВАРИАНТНЫМИ ПАРАМЕТРАМИ

РЕЗЮМЕ

диссертации на соискание учёной степени

кандидата компьютерных наук

Москва — 2023

Диссертационная работа выполнена в федеральном государственном автономном образовательном учреждении высшего образования «Национальный исследовательский университет «Высшая школа экономики».

Научный руководитель: Ветров Дмитрий Петрович, к.ф.-м.н., Национальный исследовательский университет «Высшая школа экономики».

1 Тема диссертации

Масштабная инвариантность является одним из ключевых свойств, присущих параметрам большинства современных нейросетевых архитектур. Обеспеченная повсеместным наличием слоев нормализации промежуточных активаций и/или непосредственно весов, масштабная инвариантность, как следует из названия, заключается в неизменности реализуемой нейросетью функции при умножении ее параметров на произвольный положительный скаляр. В этой работе исследуются эффекты данного свойства на динамику обучения нейросетевых моделей, а также его влияние на внутреннее устройство ландшафта функции потерь.

В первой части работы рассматривается и разбирается эффект периодично повторяющихся циклов сходимости и дестабилизации при обучении нейронных сетей с использованием слоев нормализации и техники сокращения веса (weight decay; WD). Как показывает теоретический и эмпирический анализ, такое поведение является следствием конкурирующего влияния сокращения веса и масштабной инвариантности на норму параметров нейронной сети. Таким образом, происходит цикличное изменение сферы, на которой обучается модель, что приводит в конечном итоге к наблюдаемому периодическому поведению оптимизационного процесса.

Во второй части работы раскрывается внутреннее устройство ландшафта функции потерь нейронных сетей с масштабно-инвариантными параметрами путем фиксации сферы, на которой осуществляется обучение модели. Аналитически и экспериментально показано, что в такой постановке можно выделить три режима обучения: сходимость, хаотическое равновесие и расходимость. Каждый режим обладает рядом своих особенностей и позволяет выделить те или иные свойства истинного ландшафта функции потерь масштабно-инвариантных нейронных сетей, которые также нашли отражение в реальной практике обучения нейросетевых моделей, например, при проектировании расписания темпа обучения. Описанные эффекты обучения масштабно-инвариантных моделей на сфере исследуются в различных постановках с использованием как более классической кросс-энтропийной, так и квадратичной функции потерь (Mean Squared Error; MSE) на задачах классификации, которая показала свою перспективность в последних исследованиях [1].

Актуальность работы

Несмотря на колоссальный эмпирический прогресс в области глубокого обучения в последние десятилетия, поиск удовлетворительного обоснования принципов устройства и функционирования глубоких нейросетевых моделей все еще является крайне актуальной задачей [2]. Остаются нераскрытыми как частные вопросы, относящиеся к отдельным эффектам процесса обучения и свойствам финальных решений, к примеру двойной спуск функции потерь на контроле [3, 4] или так называемый гроккинг [5], так и глобальные проблемы, связанные с внутренним устройством ландшафта функции потерь и оптимизационной динамикой нейронных сетей, по типу способности современных нейронных сетей целиком запоминать обучающую выборку [6], наличия «минных полей» [7] и связности мод [8, 9] в ландшафте функции потерь и тому подобных феноменов перепараметризованного обучения [10]. Интерпретируемость и предсказуемость обучения и вывода моделей глубокого обучения являются необходимыми условиями не только для обеспечения более широкого и безопасного их применения, но и для разработки новых способов их усовершенствования, которые будут опираться не на практическую интуицию и эвристический подход, а на строгий научный метод.

Техники нормализации, такие как батч-нормализация (Batch Normalization; BN) [11], послойная нормализация (Layer Normalization) [12] или нормализация весов (Weight Normalization) [13], повсеместно используются в современных нейросетевых архитектурах, что эмпирически зачастую позволяет стабилизировать процесс обучения и улучшить итоговое качество моделей, однако дополнительно усложняет понимание происходящих в нейросетях процессов. Несмотря на определенный прогресс в понимании некоторых свойств, обеспеченных использованием нормализации в нейронных сетях, многие вопросы все еще остаются нераскрытыми [14, 15]. В частности, не до конца ясна роль нормализации в определении эффективной структуры поверхности функции потерь, а также как именно она влияет на динамику обучения современных нормализованных нейросетей. Данные вопросы в последние годы обретают особую актуальность в связи с обнаруженными эффектами сингулярности и нестабильности в определенных режимах применения техник нормализации [16, 17, 18, 19], что, вообще говоря, противоречит общепринятому ранее мнению о том, что нормализация, напротив, обеспечивает стабилизацию процесса обучения нейронных сетей.

Возможно, наиболее общим, а потому ключевым следствием использования произвольных техник нормализации в архитектуре нейронной сети является обеспечение свойства масштабной инвариантности весов данной сети, непосредственно предшествующих нормализационным слоям. В силу того, что обычно нормализация применяется практически после каждого внутреннего слоя нейросети, на практике оказывается, что подавляющее большинство параметров модели обладают указанным свойством. Данное обстоятельство подчеркивает основное отличие нормализованных нейросетей от сетей без использования техник нормализации, поэтому его нельзя игнорировать при изучении влияния нормализации на динамику обучения и ландшафт функции потерь. Таким образом, актуальное исследование и интерпретация нормализационных техник обязаны опираться на свойство масштабной инвариантности и его следствия, что демонстрируют последние работы в данной области [18, 19, 20, 21, 22, 23, 24, 25].

Первая часть данной диссертации посвящена раскрытию, исследованию и объяснению эффекта периодического поведения обучения нейронных сетей с использованием техник нормализации и сокращения веса. Сокращение веса является повсеместно использующимся приемом при обучении моделей машинного обучения, который заключается в скалярном умножении параметров на заданный положительный коэффициент меньше единицы после каждой итерации обучения и играет роль обобщенного классического L_2-регуляризатора [22, 26, 27]. Несмотря на то, что масштабно-инвариантные модели по определению не зависят от фактического значения нормы параметров, выясняется, что сокращение веса тем не менее существенно влияет на динамику обучения таких моделей за счет нетривиального изменения так называемого эффективного темпа обучения (effective learning rate; ELR). Предыдущие работы, исследующие данный эффект, пришли к некоторому противоречию по поводу того, как данное влияние определяет итоговое поведение оптимизационной динамики. Одни разделяют точку зрения о том, что обучение нормализованных моделей с использованием сокращения веса обязано в итоге прийти в состояние равновесия, когда все наблюдаемые метрики, включая величину эффективного темпа обучения, норму параметров, эмпирический риск и прочее, стабилизируются в каком-то фиксированном значении, что в целом благотворно влияет на результат обучения [19, 20, 28, 23]. Другие, наоборот, утверждают, что сокращение веса спустя определенное количество итераций обучения слишком сильно приблизит норму весов к нулю, что приведет к

численным нестабильностям и расхождению процесса оптимизации [17, 18, 19]. В данной работе описанное противоречие снимается и демонстрируется, что обе позиции в определенном смысле справедливы. С одной стороны, динамика обучения нормализованных моделей с сокращением веса в самом деле постоянно претерпевает нестабильности по указанной выше причине. С другой, такие нестабильности имеют последовательный характер, что приводит к периодическому поведению динамики обучения. Данное периодическое поведение имеет регулярную структуру, что позволяет в том числе рассматривать его как некое обобщение принципа равновесия. В данной работе проводится подробный экспериментальный и теоретический анализ, описывающий и обосновывающий механизмы, стоящие за таким периодическим поведением. В основной статье по данной теме также исследуются его следствия и эффекты в отношении обучения современных глубоких нейросетевых моделей.

Во второй части диссертации упор делается на исследовании структуры ландшафта функции потерь масштабно-инвариантных нейронных сетей на их внутреннем домене, а именно, на сфере. Поскольку масштабно-инвариантные модели по своей сути не меняются при движении параметров вдоль радиального направления от начала координат, их естественной областью определения можно считать сферу, а не все пространство параметров. Соответственно, траекторию их обучения можно также эффективно рассматривать через проекцию на сферу, чтобы лучше понимать, как устроена оптимизационная динамика на истинном домене. Однако при стандартном обучении масштабно-инвариантных моделей во всем пространстве, особенно с использованием техники сокращения веса, эффективный темп обучения, который отвечает за скорость оптимизации на сфере единичного радиуса, нетривиально меняется, как, в частности, показано в предыдущей части работы. Это затрудняет изучение внутреннего ландшафта функции потерь, поскольку величину эффективного шага оптимизации не удается контролировать даже при фиксации обычного темпа обучения (learning rate; LR). В данной работе эта проблема решается путем перехода к оптимизации полностью масштабно-инвариантных нейронных сетей непосредственно на сфере методом проекции стохастического градиента. Такая процедура обучения нивелирует эффект динамически меняющегося эффективного темпа обучения и фиксирует его по построению, поскольку устраняет вариативность нормы параметров в ходе обучения и полностью переносит динамику на естественный домен. Это позволяет подробно и контролируемо исследовать истинное устройство ландшаф-

та функции потерь масштабно-инвариантных нейросетевых моделей. Выясняется, что обучение масштабно-инвариантных нейросетей на сфере может осуществляться в трех режимах в зависимости от заданного значения эффективного темпа обучения: сходимость, хаотическое равновесие и расходимость. Каждый режим обладает рядом отличительных черт и раскрывает те или иные особенности внутреннего устройства ландшафта функции потерь, к примеру, наличие целого спектра функционально и геометрически различных глобальных минимумов, соответствующих разным значениям эффективного темпа обучения в первом режиме, зон повышенной кривизны, препятствующих сходимости и отделяющей первый режим от второго, а также локальных и глобальных областей стабилизации оптимизационной динамики во втором режиме обучения. Изучению особенностей данных режимов и их следствий на динамику обучения и ландшафт функции потерь нейронных сетей с масштабно-инвариантными параметрами были посвящены две статьи: в первой основной упор делается на исследовании классической кросс-энтропийной функции потерь и впервые раскрываются основные свойства трех режимов обучения на сфере, во второй рассматривается случай квадратичной функции потерь в задачах классификации [1], а также дополняются и расширяются результаты предыдущей работы. Помимо прочего, в работах демонстрируется, как данные режимы проявляются при стандартном обучении актуальных архитектур глубокого обучения и как их можно использовать на практике, к примеру, для построения оптимальных расписаний темпа обучения.

Цель данной работы заключается в раскрытии и исследовании особенностей динамики обучения и устройства ландшафта функции потерь нейронных сетей с масштабно-инвариантными параметрами. Это позволит улучшить интерпретируемость современных нейросетевых моделей, использующих техники нормализации.

2 Основные результаты и выводы

Вклад. Основные результаты работы приведены ниже.

1. Исследована динамика обучения нормализованных нейронных сетей во всем пространстве параметров с использованием техники сокращения веса. Открыт и проанализирован как экспериментально, так и теоретически эффект периодического поведения такой динамики.

2. Через описанное периодическое поведение предложено разрешение сложившегося в литературе противоречия касаемо результата данной оптимизационной динамики: равновесие или нестабильность. Выведен принцип обобщенного равновесия.

3. Исследована динамика обучения полностью масштабно-инвариантных нейронных сетей на их естественном домене — сфере. Открыты и проанализированы как экспериментально, так и теоретически три режима такого обучения: сходимость, хаотическое равновесие и расходимость, в том числе выделены их основные характеристики.

4. Путем изучения данных режимов раскрыт ряд свойств внутреннего ландшафта функции потерь масштабно-инвариантных моделей, включая существование спектра различных глобальных минимумов, зон повышенной кривизны, а также регионов стабилизации оптимизационной динамики.

5. Дополнительно три режима исследованы для случая обучения с квадратичной функцией потерь на задачах классификации.

Теоретическая и практическая значимость. Данная работа продолжает общий актуальный тренд в области глубокого обучения на поиск и разработку удовлетворительных обоснований механизмов, стоящих за устройством и работой нейросетевых моделей. Упор в диссертации делается на принципе масштабной инвариантности, обусловленном применением техник нормализации, которые повсеместно используются в большинстве современных архитектур. Полученные результаты не только позволяют выявить и объяснить различные свойства динамики обучения и устройства ландшафта функции потерь нормализованных моделей, но и помогают обобщать предыдущие знания и разрабатывать более эффективные способы обучения нейронных сетей. В частности, с помощью открытого периодического поведения из первой части работы удалось разрешить сложившееся в литературе противоречие о динамике обучения нормализованных нейросетей с техникой сокращения веса, а изучение свойств выявленных трех режимов обучения на сфере из второй части послужило основой для интерпретации и выбора расписаний темпа обучения. Разработанные теоретические результаты позволяют усилить и формализовать получен-

ную эмпирическую интуицию, а также сами по себе представляют интерес в качестве рабочей математической модели, описывающей масштабно-инвариантную динамику.

Результаты, выносимые на защиту.

1. Открытое периодическое поведение динамики обучения нормализованных нейронных сетей с техникой сокращения веса, его экспериментальный и теоретический анализ.

2. Выведенный принцип обобщенного равновесия, разрешающий конфликт двух противоречащих друг другу позиций касаемо динамики такого обучения: равновесие против нестабильности.

3. Три открытых режима обучения полностью масштабно-инвариантных нейронных сетей на сфере с использованием как кросс-энтропийной, так и квадратичной функции потерь: сходимость, хаотическое равновесие и расходимость; их экспериментальный и теоретический анализ.

4. Выявленные свойства ландшафта функции потерь масштабно-инвариантных нейронных сетей на сфере: спектр различных глобальных минимумов, зоны повышенной кривизны, области стабилизации оптимизационной динамики и другие.

Личный вклад в результаты, выносимые на защиту. В первой работе автором были сформулированы и доказаны все представленные теоретические результаты. Автор внес основной вклад в обзор релевантной литературы, в частности, установил наличие противоречия по поводу результата исследуемой динамики обучения и предложил его разрешение через открытое периодическое поведение. Также автор участвовал в постановке экспериментов, анализе эмпирических результатов и написании текста совместно с Лобачевой Екатериной и другими соавторами.

Во второй работе автором также были сформулированы и доказаны все представленные теоретические результаты. Автор внес основной вклад в написание текста и обзор релевантной литературы. Совместно с другими соавторами участвовал в анализе и интерпретации эмпирических результатов, включая установление основных характеристик трех режимов обучения на сфере и их следствий в отношении ландшафта функции потерь. Автор также содействовал в постановке экспериментов, основную роль в которой приняли Лобачева Екатерина и Находнов Максим.

В третьей работе автор был одним из инициаторов исследования трех режимов обучения на квадратичной функции потерь, а также содействовал главному автору Находнову Максиму в интерпретации и систематизации полученных результатов, обзоре литературы и постановке экспериментов.

Публикации и апробация работы

Автор диссертации является главным автором в двух публикациях повышенного уровня и вторым автором в одной публикации стандартного уровня по теме диссертации.

* — авторы с равным вкладом в работу.

Публикации повышенного уровня:

1. Екатерина Лобачева*, **Максим Кодрян***, Надежда Чиркова, Андрей Малинин, Дмитрий Ветров. On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay (О периодическом поведении обучения нейронной сети с батч-нормализацией и сокращением веса). В материалах конференции Advances in Neural Information Processing Systems, 2021 (NeurIPS 2021). Том 34, стр. 21545-21556. Конференция ранга A* по рейтингу CORE.

2. **Максим Кодрян***, Екатерина Лобачева*, Максим Находнов*, Дмитрий Ветров. Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes (Обучение масштабно-инвариантных нейронных сетей на сфере может происходить в трех режимах). В материалах конференции Advances in Neural Information Processing Systems, 2022 (NeurIPS 2022). Том 35, стр. 14058-14070. Конференция ранга A* по рейтингу CORE.

Публикации стандартного уровня:

1. Максим Находнов, **Максим Кодрян**, Екатерина Лобачева, Дмитрий Ветров. Loss Function Dynamics and Landscape for Deep Neural Networks Trained with Quadratic Loss (Динамика и ландшафт функции потерь для глубоких нейронных сетей при обучении с квадратичной функцией потерь). Опубликовано в журнале Doklady Mathematics в 2022 году. Том 106, выпуск 1 (приложение), стр. 43-62. Журнал является англоязычной версией «Доклады Российской академии наук. Математика, информатика, процессы управления», индексируется в Scopus.

Доклады на научных конференциях и семинарах:

1. Конференция Neural Information Processing Systems, декабрь 2021. Тема: «On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay».

2. Семинар Mathematical Machine Learning Seminar MPI MIS + UCLA, декабрь 2021. Тема: «On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay».

3. Летняя Школа по Машинному Обучению EMINES School of Industrial Management, июль 2022. Тема: «On the Periodic Behavior of Neural Network Training with Batch Normalization and Weight Decay».

4. Семинар исследовательской группы байесовских методов, октябрь 2022. Тема: «Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes».

5. Конференция Fall into ML, ноябрь 2022. Тема: «Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes».

6. Конференция Neural Information Processing Systems, декабрь 2022. Тема: «Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes».

7. Семинар AIRI ИИшница, декабрь 2022. Тема: «Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes».

8. Конференция ФКН НИУ ВШЭ в Вороново, июнь 2023. Тема: «Training Scale-Invariant Neural Networks on the Sphere Can Happen in Three Regimes».

Объем и структура работы. Диссертация содержит введение, содержание публикаций и заключение. Полный объем диссертации 103 страницы.

3 Содержание работы

3.1 Периодическое поведение динамики обучения нормализованных нейронных сетей с техникой сокращения веса

Под нормализованными нейронными сетями понимаются такие, в архитектуре которых используются либо слои нормализации, такие как батч-нормализация [11] или послойная нормализация [12], либо непосредственно нормализация весов [13]. Нормализованными являются абсолютное большинство современных глубоких нейронных сетей, включая, к примеру, популярные архитектуры ResNet [29] и Transformer [30]. Современным стандартом обучения нейросетевых моделей стали стохастические методы с использованием техники сокращения веса, которая стабилизирует процесс оптимизации, а также играет роль регуляризатора, что повышает обобщающую способность финальных решений [22, 26, 27].

Как сказано ранее, нормализация индуцирует масштабную инвариантность весов нейросети, предшествующих слоям нормализации или непосредственно нормализованных. Из-за повсеместности применения техник нормализации в современных архитектурах таких весов оказывается большинство, потому изучение влияния масштабной инвариантности на оптимизационную динамику оказывается актуальной проблемой. Это приобретает особую важность в силу нетривиального и даже неожиданного взаимодействия масштабной инвариантности с техникой сокращения веса, как будет показано далее.

В данном разделе приводится исследование периодичной динамики обучения нормализованных моделей с сокращением веса (Рис. 1), которую в том числе можно рассматривать как обобщение двух конфликтующих точек зрения насчет итога такого обучения: равновесие против нестабильности. Для упрощения изложения и в качестве наиболее показательного случая здесь рассматриваются эксперименты со сверточными нейронными сетями, использующими батч-нормализацию и обучающимися с помощью алгоритма стохастического градиентного спуска (Stochastic Gradient Descent; SGD) с константным темпом обучения; тем не менее в основной работе показано, что приведенные результаты остаются также справедливы при использовании других архитектур, техник нормализации, алгоритмов оптимизации, включая классический градиентный спуск или оптимизатор Adam [31], и даже более общих масштабно-инвариантных моделей.

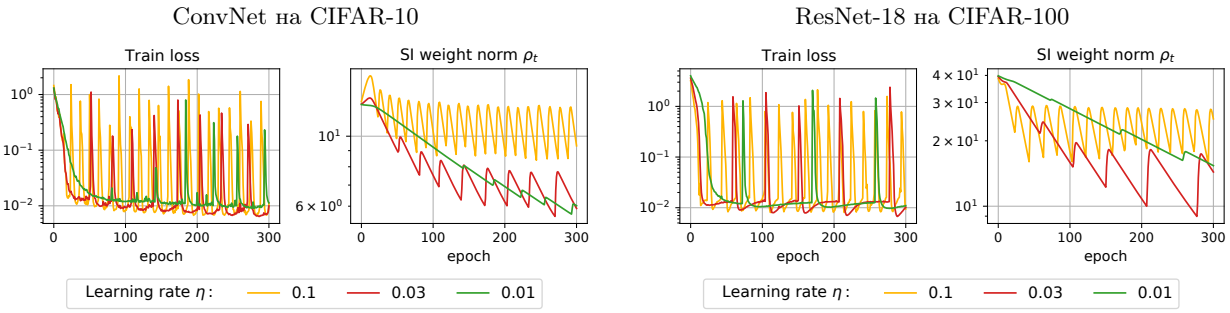


Рис. 1: Периодическое поведение ConvNet на CIFAR-10 и ResNet-18 на CIFAR-100, обученных с использованием SGD с коэффициентом сокращения веса 0.001 и разными значениями темпа обучения. На каждой панели слева приведено поведение функции потерь на обучении от эпохи, справа — нормы масштабно-инвариантных параметров от эпохи.

Основные понятия и постановка проблемы

Для прояснения постановки проблемы опишем основные следствия масштабной инвариантности на динамику обучения и ее взаимодействие с техникой сокращения веса. Рассмотрим произвольную масштабно-инвариантную функцию $f(x)$ такую, что

$$f(\alpha x) = f(x), \ \forall x, \forall \alpha > 0. \tag{1}$$

Уравнение (1) является по сути определением масштабной инвариантности. Путем дифференцирования обеих частей равенства (1) по x и по α можно получить следующие фундаментальные свойства градиента произвольных масштабно-инвариантных функций (см. Лемму 1.3 в Li and Arora [18]):

$$\begin{cases} \langle \nabla f(x), x \rangle = 0, \ \forall x & \text{(2a)} \\ \nabla f(\alpha x) = \dfrac{1}{\alpha} \nabla f(x), \ \forall x, \forall \alpha > 0. & \text{(2b)} \end{cases}$$

Рассмотрим оптимизацию $f(x)$ методом градиентного спуска с темпом обучения η и коэффициентом сокращения веса λ:

$$x_{t+1} = (1 - \eta\lambda)x_t - \eta \nabla f(x_t). \tag{3}$$

Приведенные выше свойства приводят к двум важным следствиям, касающимся динамики процесса оптимизации. Во-первых, согласно свойству (2a), сдвиг x в направлении $-\nabla f(x)$, т. е. шага градиентного спуска, всегда увеличивает $\|x\|$, а сокращение веса, напротив, уменьшает $\|x\|$ (см. Рис. 2 для иллюстрации). Взаимодействие данных «центростремительных» и «центробежных» сил может привести к нетривиальному изменению $\|x\|$ в процессе оптимизации. Во-вторых, согласно свойству (2b),

несмотря на то, что значение самой функции $f(x)$ инвариантно относительно умножения x на α, динамика существенно меняется, когда оптимизация выполняется на разных масштабах нормы параметров. Для меньших норм оптимизация делает более крупные шаги, что может привести к нестабильности, в то время как для больших норм шаги меньше, и процесс оптимизации может медленно сходиться.

Поскольку масштабно-инвариантную $f(x)$ можно рассматривать как функцию на сфере, ее оптимизационную динамику часто анализируют на единичной сфере, то есть при $\|x\| = 1$. Можно показать, что динамику обучения во всем пространстве параметров можно эквивалентно представить на единичной сфере, используя понятия эффективного градиента и эффективного темпа обучения вместо аналогичных стандартных понятий [18, 19, 20, 21, 22, 23, 24]. Эффективный градиент определяется как градиент в точке, спроецированной на единичную сферу, и может быть получен из (2b) как $\nabla f(x/\|x\|) = \nabla f(x)\|x\|$. Эффективный темп обучения определяется как $\tilde{\eta} = \eta/\|x\|^2$ [21, 24]. Изменение $\|x\|$ не влияет на эффективный градиент по определению и отражается только в эффективном темпе обучения: чем ниже норма, тем выше эффективный темп обучения и тем по сути больше реальные шаги оптимизации. В дальнейшем будут использоваться следующие обозначения для норм параметров, градиента, эффективного градиента и для эффективного темпа обучения на итерации t: $\rho_t \equiv \|x_t\|$, $g_t \equiv \|\nabla f(x_t)\|$, $\tilde{g}_t \equiv \|\nabla f(x_t/\|x_t\|)\| = \rho_t g_t$ и $\tilde{\eta}_t \equiv \eta/\rho_t^2$ соответственно.

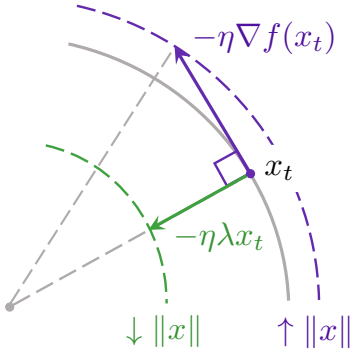


Рис. 2: Иллюстрация «центростремительной силы» сокращения веса и «центробежной силы» градиента при оптимизации масштабно-инвариантных функций.

Наличие противоречия в литературе

Кратко изложим суть сформировавшегося в сообществе противоречия по поводу обучения нормализованных моделей с масштабно-инвариантными параметрами с использованием техники сокращения веса (3). Как показано ранее, динамика такого обучения нетривиально изменяет норму параметров в силу взаимодействия «центробежной силы» градиента и «центростремительной силы» сокращения веса, что в свою очередь влияет на величину шага оптимизационного процесса. Таким образом,

в литературе сложились две противоречащих друг другу точки зрения по поводу результата соответствующей динамики.

С одной стороны, работы по типу Li et al. [19] или Wan et al. [23] утверждают, что такое обучение приводит к *равновесному* состоянию, где «центростремительная сила» компенсируется «центробежной силой» и в конечном итоге норма масштабно-инвариантных весов (вместе с другими статистиками обучения) будет стремиться к некоторому постоянному значению. Несколько других работ придерживаются аналогичного взгляда [20, 28].

С другой стороны, в ряде работ подчеркивается, что в связи с ослаблением роли градиента с прогрессом обучения использование сокращения веса может приблизить параметры слишком близко к нулевой точке, что приводит к *нестабильности* из-за чрезмерного повышения эффективного темпа обучения. В частности, в работе Li et al. [17] показывается, что приближение к нулевой точке в нормализованных нейронных сетях приводит к численным ошибкам после шага оптимизации и последующей остановке обучения. Li and Arora [18] также подчеркивают, что масштабно-инвариантные функции плохо обусловлены вблизи начала координат, и в упрощенном виде доказывают, что сходимость невозможна, если используются как техники нормализации, так и сокращения веса (однако гарантируется, если любая из них отсутствует). Более того, несмотря на приведенную точку зрения о равновесии, в работе Li et al. [19] эмпирически демонстрируется, что функция потерь на обучении претерпевает постоянные колебания между низкими и высокими значениями, когда используется градиентный спуск с сокращением веса.

Одним из ключевых результатов данной диссертационной работы является открытое периодическое поведение обучения нормализованных моделей с сокращением веса, которое, как показано далее, позволяет разрешить описанное фундаментальное противоречие через принцип обобщенного равновесия.

Периодическое поведение и лежащие в его основе механизмы

Для чистоты изложения приведенные здесь и далее экспериментальные результаты затрагивают случай полностью масштабно-инвариантных сверточных нейронных сетей (ConvNet, ResNet-18), обучаемых с помощью стохастического градиентного спуска с сокращением веса на задачах классификации натуральных изображений CIFAR [32]. Для обеспечения полной масштабной инвариантности после каждого

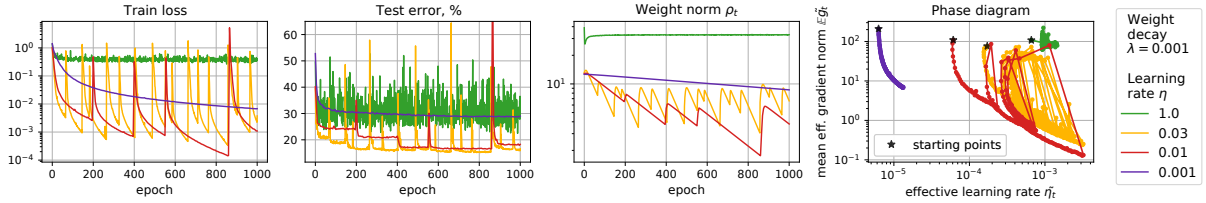


Рис. 3: Периодическое поведение обучения масштабно-инвариантной нейронной сети ConvNet на CIFAR-10.

сверточного слоя в архитектуру добавляется слой батч-нормализации с необучаемыми аффинными параметрами, а также фиксируется последний линейный слой. Это гарантирует масштабную инвариантность всех обучаемых параметров модели, при этом практически не снижает ее качество [33, 18]. В основной работе специально дополнительно исследуется случай стандартных архитектур с присутствием немасштабно-инвариантных параметров, а также обучаемых в более нетривиальной постановке с использованием инерции, расписания темпа обучения и аугментации данных; вкратце, все результаты, включая периодическое поведение, остаются справедливыми при условии достаточно долгого обучения и не слишком агрессивного расписания темпа обучения (см. также Рис. 1).

На Рисунке 3 продемонстрировано периодическое поведение обучения масштабно-инвариантной нейронной сети ConvNet на наборе данных CIFAR-10 для различных значений темпа обучения. В процессе оптимизации явно встречаются нестабильности, которые, однако, не приводят к полной расходимости, но вызывают новый цикл обучения, причем наблюдаемая периодичность дестабилизаций, как и поведение динамики обучения внутри каждого цикла, является регулярным и подчиняющимся некоторому обобщенному закону равновесия. Таким образом, эмпирически можно заключить, что приведенные ранее позиции равновесия и нестабильности оказываются одновременно справедливыми в той или иной мере. Более конкретные теоретические результаты, подкрепляющие такой вывод, приведены далее.

Наблюдаемое периодическое поведение возникает из-за взаимодействия между нормализацией и сокращением веса, а именно, из-за их конкурирующего влияния на норму масштабно-инвариантных весов нейросети. Как обсуждалось ранее, сокращение веса направлено на уменьшение нормы параметров, в то время как градиенты от функции потерь направлены на ее увеличение (см. Рис. 2). Эти две силы попере-

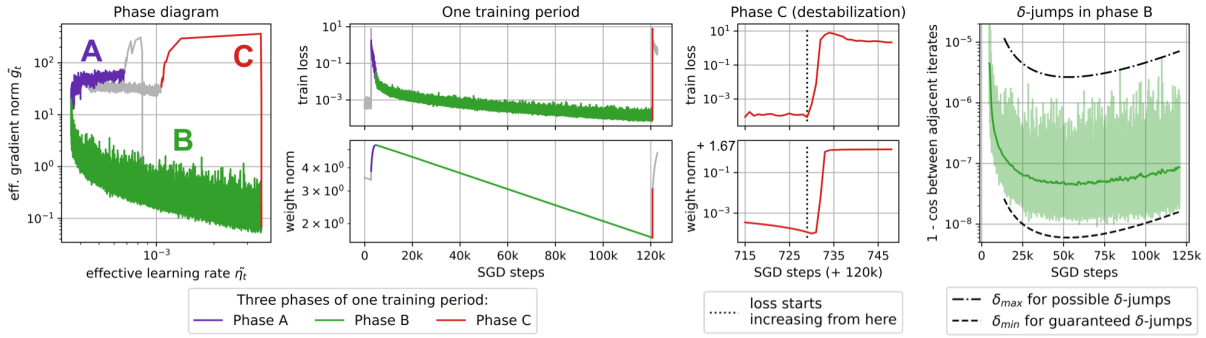


Рис. 4: Один период обучения масштабно-инвариантной сети ConvNet на CIFAR-10: можно выделить три фазы. На крайнем правом графике сравнивается эмпирически наблюдаемое косинусное расстояние между весами на соседних итерациях с теоретически полученными границами.

менно перевешивают друг друга в течение довольно длительных циклов обучения, что и приводит к периодическому поведению.

Для уточнения деталей рассмотрим более подробно один период обучения, проанализировав Рисунок 4. В начале периода наблюдаются высокие значения функции потерь на обучении, а следовательно, большие градиенты перевешивают влияние сокращения веса. Это приводит к быстрому снижению потерь на обучении, увеличению нормы параметров и уменьшению эффективного темпа обучения. Описанная фаза обозначена буквой A на графиках. Далее, по мере снижения потерь на обучении наступает момент, когда градиенты становятся малыми и перевешиваются сокращением веса. В результате норма весов начинает уменьшаться, а эффективный темп обучения увеличиваться, что обозначено как фаза B на рисунках. Наконец, когда норма параметров становится слишком малой, а эффективный темп обучения, напротив, слишком высоким, оптимизация совершает несколько больших шагов и покидает окрестность точки минимума. Эффективные градиенты резко возрастают вместе со значением функции потерь на обучении и, помноженные на высокий эффективный темп обучения, приводят к быстрому росту нормы весов (фаза C). Подробные графики для фазы C на третьей слева панели Рисунка 4 подтверждают, что потери на обучении начинают возрастать раньше нормы параметров. В конечном итоге, когда норма весов становится большой, эффективный темп обучения снижается и останавливает процесс расхождения, что приводит к новому циклу обучения.

В основной работе также приводятся дополнительные абляционные исследования, демонстрирующие, что периодическое поведение в самом деле удается устранить пу-

тем фиксации нормы параметров (также см. следующий раздел данной диссертации). Это дополнительно подтверждает предложенное обоснование механизма периодического поведения через динамику нормы весов, а следовательно, и эффективной скорости оптимизации.

Теоретический анализ

В данном подразделе приводятся основные теоретические результаты, касающиеся динамики оптимизации масштабно-инвариантных функций с сокращением веса (3), которые дополняют и подкрепляют приведенные выше эмпирические наблюдения. Главным образом, формулируются понятия и результаты, позволяющие объяснить причины дестабилизации между фазами B и C периода, а также приводится теорема, формально обобщающая позицию Li et al. [19] о равновесии. Точная формулировка утверждений вместе с доказательствами приводится в основной работе.

В первую очередь, динамика (3) переформулируется в терминах динамики норм параметров и градиентов:

$$\rho_{t+1}^2 = (1 - \eta\lambda)^2 \rho_t^2 + \eta^2 \tilde{g}_t^2 / \rho_t^2. \tag{4}$$

Далее, вводится понятие *δ-скачка*, отражающее необходимое условие для возникновение дестабилизации при анализе динамики (3) при помощи (4).

Определение 1. *Говорится, что динамика* (3) *совершает* ***δ-скачок****, если косинусное расстояние между соседними итерациями превышает заданный порог $\delta > 0$:*

$$1 - \cos(x_t, x_{t+1}) > \delta.$$

Первый результат о необходимых и достаточных условиях для возникновения δ-скачка формулируется следующим образом.

Утверждение 1. *Положим эффективную норму градиента локально ограниченной: $\ell \leq \tilde{g}_t \leq L$. Тогда выполняются следующие приближенные условия для δ-скачка:*

$$\begin{cases} \rho_t^2 \lessapprox \dfrac{\eta L}{\sqrt{2\delta}} \implies \delta\text{-скачок возможен}, & \text{(5a)} \\ \rho_t^2 \lessapprox \dfrac{\eta \ell}{\sqrt{2\delta}} \implies \delta\text{-скачок гарантирован}. & \text{(5b)} \end{cases}$$

Выполнение данных условий на практике можно видеть на правой панели Рисунка 4. Видно, что реальная динамика косинусного расстояния между соседними

итерациями остается в пределах границ на δ-скачок, полученными из Утверждения 1. Также видно, что общий тренд данной динамики демонстрирует увеличение δ-скачков ближе к концу фазы B, что свидетельствует о возрастающей нестабильности.

Следующее утверждение позволяет спрогнозировать время возникновения δ-скачка, а значит, и дестабилизации в зависимости от выбора гиперпараметров темпа обучения η и коэффициента сокращения веса λ.

Утверждение 2. *Обозначим* $\kappa = \sqrt{\frac{\eta}{2\lambda}}$. *При условиях Утверждения 1:*

- $\rho_0^2 > \kappa\ell \wedge \delta < \eta\lambda\frac{L^2}{\ell^2} \Rightarrow$ ***минимальное*** *время до* δ*-скачка:* $t_{\min} = \mathcal{O}(1/4\eta\lambda)$*;*
- $\rho_0^2 > \kappa L \wedge \delta < \eta\lambda\frac{\ell^2}{L^2} \Rightarrow$ ***максимальное*** *время до* δ*-скачка:* $t_{\max} = \mathcal{O}(1/2\eta\lambda)$.

Следствие 1. *Таким образом, можно заключить, что нестабильности, а следовательно, и периоды возникают с частотой, прямо пропорциональной произведению темпа обучения на коэффициент сокращения веса* $\eta\lambda$.

Наконец, финальный и центральный теоретический результат — *принцип обобщенного равновесия* — обобщает позицию равновесия из работ [19, 23] и таким образом закрывает вопрос о противоречии с позицией нестабильности: в ходе обучения периодическое поведение обязано стабилизироваться в определенных границах.

Теорема 1. *При условиях Утверждения 2 и выполнения* $2\eta\lambda L \leq \ell$ *справедливы следующие границы на норму параметров:*

$$\kappa\ell \leq \rho_t^2 \leq \kappa L,\ t \gg 1.$$

Если $\rho_0^2 > \kappa L$, *то* ρ_t^2 *линейно сходится к интервалу* $[\kappa\ell, \kappa L]$ *за время* $\mathcal{O}\left(1/\eta\lambda\right)$.

Эмпирическое подтверждение приведенных утверждений подробно приводится в основной работе, однако его также можно видеть на Рисунке 3. Видно, что с увеличением гиперпараметра η при фиксированном λ периоды становятся чаще в соответствии с Утверждением 2 и Следствием 1. Также можно заметить, что, несмотря на периодическое поведение, с определенной эпохи обучения норма параметров (как и прочие метрики) четко лежит в определенных границах, подтверждая тезис, сформулированный в Теореме 1.

Заключение и прочие результаты

В данной части диссертационной работы подробно разобран феномен периодического поведения динамики обучения нормализованных нейросетевых моделей с техникой сокращения веса. Приведено разъяснение механизмов, стоящим за данным периодическим поведением, а также приведены теоретические результаты, подкрепляющие эмпирическую интуицию. Наконец, через принцип обобщенного равновесия периодического поведения было разрешено противоречие по поводу такой динамики обучения: равновесие или нестабильность.

В основной работе также приведено множество дополнительных результатов, опущенных здесь, включая эмпирический анализ следствий периодического поведения, таких как стадия прогрева (warm-up stage) и минимумы на разных периодах обучения, и абляционные исследования периодического поведения в разных постановках стандартного обучения нейронных сетей.

3.2 Три режима обучения масштабно-инвариантных нейронных сетей на сфере

В предыдущем разделе диссертации упоминалось, что масштабно-инвариантные функции (1) определены по своей сути на сфере, что является их естественным доменом. Как правило, подразумевается единичная сфера, для которой специально вводятся понятия эффективного градиента и эффективного темпа обучения, однако в данной работе рассматривается более широкое определение и под естественной, или внутренней, областью определения масштабно-инвариантных функций будет пониматься сфера произвольного фиксированного радиуса в пространстве параметров. При изучении нейронных сетей с масштабно-инвариантными весами, таким образом, возникает вопрос об устройстве ландшафта функции потерь на указанном естественном домене, дабы не учитывать побочные симметрии, не влияющие по сути на реализуемую моделью функцию.

Обыкновенно для исследования структуры ландшафта функции потерь нейронных сетей и/или динамики их обучения прибегают к контролируемым экспериментам, когда все гиперпараметры, включая темп обучения, выставляются фиксированными или хотя бы подконтрольными величинами [34, 35, 36, 37, 38]. Как было подробно разобрано ранее, в случае обучения нейросетей с масштабно-инвариантными

параметрами во всем пространстве эффективный темп обучения оказывается нетривиально меняющейся величиной, даже если все гиперпараметры обучения зафиксированы. Это усложняет и даже искажает понимание внутреннего устройства ландшафта функции потерь таких моделей.

Для того чтобы исправить этот недостаток, в данном разделе рассматривается обучение полностью масштабно-инвариантных моделей на сфере фиксированного радиуса методом проекции (стохастического) градиента. В такой постановке величина эффективного темпа обучения оказывается полностью контролируемой на этапе постановки эксперимента, в частности, возможна ее фиксация в заданное константное значение. Выясняется, что в зависимости от значения эффективного темпа обучения такая оптимизация на сфере может осуществляться в трех режимах: сходимость, хаотическое равновесие и расходимость (см. Рис. 5). Первый режим (низкие значения эффективного темпа обучения) можно рассматривать как типичный случай сходимости к минимуму с монотонно убывающим значением функции потерь на обучении. Второй режим (средние значения эффективного темпа обучения) демонстрирует устойчивое колебательное поведение функции потерь вокруг некоторого значения, отделенного как от глобального минимума, так и от случайного угадывания. Данный режим носит название *хаотическое равновесие*, поскольку напоминает состояние равновесия, описанное в некоторых предыдущих работах, как упоминалось до этого [19, 23]. Последний третий режим (высокие значения эффективного темпа обучения) представляет собой дестабилизированное, расходящееся состояние обучения, связанное с чрезмерно большим размером шага оптимизации. Каждый режим позволяет выявить определенные особенности устройства ландшафта функции потерь на сфере, которые расширяют и углубляют предыдущие результаты.

Далее в данном разделе рассматривается обучение именно в такой постановке с использованием полностью масштабно-инвариантных нейронных сетей, которые получаются из стандартных архитектур методом, описанным в предыдущем разделе. В двух основных работах по теме данного раздела дополнительно исследуется случай обучения масштабно-инвариантных моделей во всем пространстве, а также обыкновенного обучения нейронных сетей, в том числе с расписанием темпа обучения, с точки зрения трех открытых режимов.

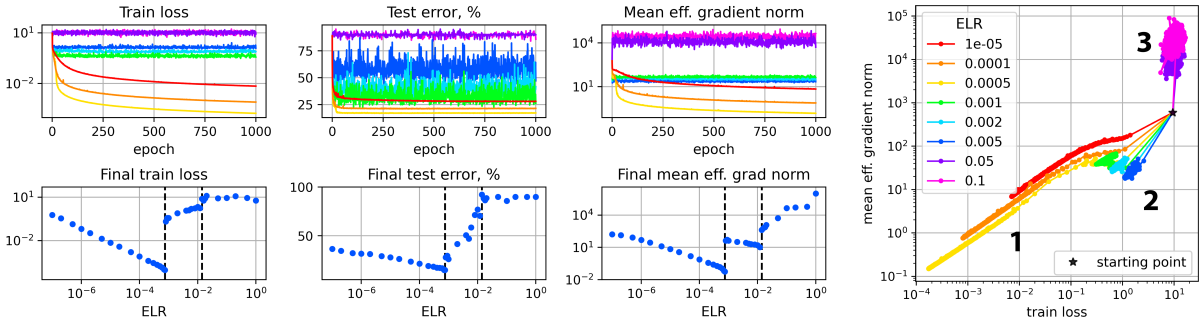


Рис. 5: Три режима обучения масштабно-инвариантной нейронной сети на сфере: (1) сходимость для низких, (2) хаотическое равновесие для средних и (3) расходимость для высоких значений эффективного темпа обучения. ConvNet на CIFAR-10. Штриховые линии обозначают границы между режимами.

Теоретический анализ

Возникновение обозначенных выше режимов обучения моделей с масштабно-инвариантными параметрами на сфере методом проекции градиента является аналитически подтвержденным фактом. Для обоснования данного тезиса были выведены последующие теоретические результаты, проясняющие общие свойства подобной оптимизационной динамики и подробно разбирающие пример конкретной масштабно-инвариантной функции. Конкретные формулировки и доказательства приводимых утверждений описаны во второй из трех указанных работ автора, относящихся к данной диссертации.

Рассмотрим функцию $F(\boldsymbol{\theta})$ от вектора параметров $\boldsymbol{\theta} \in \mathbb{R}^P$, который можно условно разбить на n групп: $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n)$, где каждый $\boldsymbol{\theta}_i \in \mathbb{R}^{p_i}$ и $\sum_{i=1}^{n} p_i = P$. Будем полагать, что каждая из этих групп масштабно-инвариантна, т. е. умножение любого из $\boldsymbol{\theta}_i$ на положительный скаляр при фиксированных остальных не меняет значение функции F. Заметим, что такая ситуация типична для нейронных сетей с несколькими нормализованными слоями, поскольку каждый из этих слоев в отдельности масштабно-инвариантен. Естественно, если функция масштабно-инвариантна относительно нескольких групп параметров, то она также масштабно-инвариантна по отношению к их объединению, поэтому весь вектор параметров $\boldsymbol{\theta}$ тоже масштабно-инвариантен.

Выпишем алгоритм минимизации $F(\boldsymbol{\theta})$ на сфере радиуса ρ с фиксированным темпом обучения η:

$$\begin{cases} \hat{\boldsymbol{\theta}}^{(t)} \leftarrow \boldsymbol{\theta}^{(t)} - \eta \nabla F(\boldsymbol{\theta}^{(t)}), \\ \boldsymbol{\theta}^{(t+1)} \leftarrow \hat{\boldsymbol{\theta}}^{(t)} \cdot \frac{\rho}{\left\|\hat{\boldsymbol{\theta}}^{(t)}\right\|}. \end{cases} \tag{6}$$

По аналогии с ранее рассмотренным случаем единственной масштабно-инвариантной группы вводятся понятия индивидуального эффективного градиента и индивидуального эффективного темпа обучения для каждой группы $\boldsymbol{\theta}_i$ как аналоги обыкновенного градиента и темпа обучения, но рассчитанные в точке $\tilde{\boldsymbol{\theta}} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_i/\rho_i, \ldots, \boldsymbol{\theta}_n)$, где $\rho_i \equiv \|\boldsymbol{\theta}_i\|$. Обозначим норму эффективного градиента для группы $\boldsymbol{\theta}_i$ как $\tilde{g}_i = g_i \rho_i$, где $g_i \equiv \|\nabla_{\boldsymbol{\theta}_i} F(\boldsymbol{\theta})\|$, а соответствующий эффективный темп обучения как $\tilde{\eta}_i = \eta/\rho_i^2$. Также обозначим норму общего эффективного градиента и общий (фиксированный) эффективный темп обучения как $\tilde{g} = g\rho$, где $g \equiv \|\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta})\|$, и $\tilde{\eta} = \eta/\rho^2$.

Из соотношения $\sum_{i=1}^n \rho_i^2 = \rho^2$ выводится следующее фундаментальное соотношение, связывающее индивидуальные эффективные темпы обучения с общим:

$$\sum_{i=1}^{n} \frac{1}{\tilde{\eta}_i} = \frac{1}{\tilde{\eta}}. \tag{7}$$

Далее, вводится понятие эффективного размера шага оптимизации (effective step size; ESS), который является полным аналогом обыкновенного размера шага оптимизации, но рассчитанный для единичной сферы. По определению эффективный размера шага оптимизации равен произведению эффективного темпа обучения на норму эффективного градиента: $\tilde{\eta}_i \tilde{g}_i$ для отдельной группы $\boldsymbol{\theta}_i$ и $\tilde{\eta}\tilde{g}$ для полного вектора параметров $\boldsymbol{\theta}$. Данная величина указывает, насколько в действительности изменяются параметры после шага оптимизации с учетом их масштабной инвариантности. Оказывается, общий эффективный размер шага оптимизации является выпуклой комбинацией индивидуальных после возведения в квадрат:

$$(\tilde{\eta}\tilde{g})^2 = \sum_{i=1}^{n} \omega_i (\tilde{\eta}_i \tilde{g}_i)^2, \ \sum_{i=1}^{n} \omega_i = 1, \ \omega_i \propto \frac{1}{\tilde{\eta}_i}. \tag{8}$$

Благодаря полученным соотношениям удается вывести динамику обновления индивидуальных эффективных темпов обучения в ходе процесса (6):

$$\tilde{\eta}_i^{(t+1)} \leftarrow \tilde{\eta}_i^{(t)} \frac{1 + (\tilde{\eta}\tilde{g}^{(t)})^2}{1 + (\tilde{\eta}_i^{(t)} \tilde{g}_i^{(t)})^2}. \tag{9}$$

Данный центральный результат анализа динамики обучения на сфере позволяет заключить, что при высоком/низком значении индивидуального эффективного разме-

ра шага оптимизации на данной итерации значение эффективного темпа обучения на следующей итерации обязано понизиться/повыситься. Таким образом, поскольку величины эффективного размера шага и эффективного темпа обучения тесно связаны по определению, возникает принцип *отрицательной обратной связи*, когда большие эффективные темпы обучения должны становиться меньше, а меньшие больше. Данный принцип является общим и ключевым для анализа режимов обучения.

Для наглядного объяснения различий между тремя режимами обучения на сфере был построен следующий пример функции с несколькими группами масштабно-инвариантных параметров, для которой были изучены оптимизационные свойства в зависимости от значения общего эффективного темпа обучения:

$$F(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \alpha_i f(x_i, y_i) = \sum_{i=1}^{n} \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}, \tag{10}$$

где $\boldsymbol{x} = (x_1, \ldots, x_n)$, $\boldsymbol{y} = (y_1, \ldots, y_n)$, $\alpha_i > 0$. Таким образом, каждая из n пар (x_i, y_i) образует отдельную масштабно-инвариантную группу параметров.

Для простоты положим $\rho = 1$, то есть оптимизация (6) осуществляется на единичной сфере, а значит, $\tilde{\eta} = \eta$. Для облегчения анализа оптимизационной динамики общей функции (10) был получен следующий результат о сходимости каждой из составляющих ее подфункций $\alpha_i f(x_i, y_i) = \alpha_i \frac{x_i^2}{x_i^2 + y_i^2}$.

Утверждение 3. *Для функции $f_\alpha(x, y) = \alpha \frac{x^2}{x^2+y^2}$, оптимизируемой на сфере с эффективным темпом обучения $\tilde{\eta}$:*

1. *при $\tilde{\eta} < \frac{1}{\alpha}$ наблюдается линейная сходимость в минимум;*
2. *при $\tilde{\eta} > \frac{1}{\alpha}$ наблюдается стабилизация на уровне $\frac{1}{2}\left(\alpha - \frac{1}{\tilde{\eta}}\right)$.*

Теперь, зная условия сходимости составляющих F подфункций, а именно, когда индивидуальный эффективный темп обучения оказывается меньше $1/\alpha_i$, с учетом соотношения (7) можно заключить, что первый режим — режим сходимости — будет наблюдаться при выполнении следующего условия:

$$\tilde{\eta} < \frac{1}{\sum_{i=1}^{n} \alpha_i}. \tag{11}$$

В противном случае ситуация, при которой все индивидуальные эффективные темпы обучения окажутся ниже порога сходимости невозможна, а поэтому динамика их обновления (9) окажется незатухающей и обязана сойтись к некоторому равновесию согласно принципу отрицательной обратной связи. Натурально, состоянием

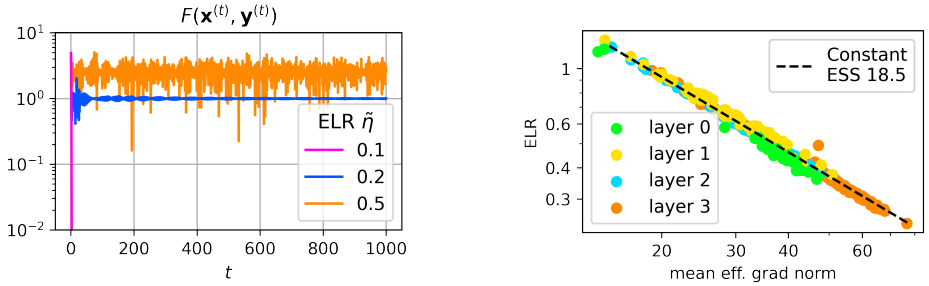


Рис. 6: Проверка теоретических результатов. Слева: динамика оптимизации функции (10) с использованием разных значений эффективного темпа обучения, соответствующих трем режимам. Справа: выравнивание индивидуальных эффективных размеров шага во втором режиме при обучении масштабно-инвариантной нейронной сети на сфере.

равновесия такой динамики является ситуация, при которой все индивидуальные эффективные размеры шага оптимизации сравняются, что эквивалентно достижению следующих уровней для индивидуальных эффективных размеров темпа обучения:

$$\tilde{\eta}_i^* \equiv \frac{\tilde{\eta} \sum_{j=1}^{n} \alpha_j}{\alpha_i}. \tag{12}$$

Таким образом, система обязана будет войти в состояние хаотического равновесия, то есть во второй режим обучения. Наконец, при слишком высоком значении $\tilde{\eta}$ хаос будет доминировать над равновесием и оптимизация перейдет в тертий режим расходимости.

На Рисунке 6 приведены некоторые результаты экспериментальной проверки полученных теоретических выкладок. Слева изображено поведение функции (10) при оптимизации на единичной сфере с тремя различными эффективными темпами обучения, соответствующих трем режимам согласно пороговому правилу (11). Видно, что наименьшее значение приводит к быстрой сходимости функции, среднее заставляет функцию стабилизироваться на определенном уровне, а самое большое вызывает наиболее хаотичное поведение. Справа изображено предсказанное выравнивание эффективных размеров шага оптимизации во втором режиме для отдельных масштабно-инвариантных групп реальной нейросетевой модели ConvNet, обучаемой на сфере. Можно видеть, что эффективные размеры шага — произведение эффективных темпов обучения на норму эффективных градиентов — действительно концентрируются около одного значения, обозначенного пунктирной черной линией на графике.

Эмпирические результаты

В данный подраздел вынесены основные экспериментальные результаты касаемо устройства ландшафта функции потерь масштабно-инвариантных нейронных сетей на сфере, полученные путем анализа трех режимов обучения. Детали постановки экспериментов, а также дополнительные результаты приведены во второй и третьей работе приложенного списка статей автора данной диссертации. Здесь для упрощения изложения будет рассматриваться масштабно-инвариантная сверточная нейронная сеть ConvNet, обученная на задаче классификации натуральных изображений CIFAR-10 указанным методом проекции стохастического градиента на сферу вида (6). Основные отображаемые метрики: значение функции потерь на обучении, ошибка на контрольной выборке, а также средняя норма эффективного стохастического градиента модели, которая одновременно играет роль меры прогресса оптимизации и кривизны найденного минимума, как обосновано в основном тексте работ. Все запуски проводились из одной точки инициализации и одним и тем же порядком батчей данных для обеспечения полной фиксации постановки эксперимента и нивелирования эффекта стохастичности обучения при сравнении результатов. Для начала разбирается случай классической кросс-энтропийной функции потерь.

Обучение с малыми значениями эффективного темпа приводит к первому режиму обучения, называемому *режимом сходимости*. Оптимизация демонстрирует типичное поведение сходимости (см. Рис. 5): после нескольких эпох модель способна достичь области с очень низким значением потерь на обучении и продолжает сходиться до точки минимума. Скорость сходимости напрямую зависит от значения эффективного темпа обучения. Кроме того, обучение с разными значениями эффективного темпа приводит к решениям с разной кривизной (средняя норма эффективных градиентов) и генерализацией (ошибка на контроле): более высокие значения приводят к решению с меньшей кривизной и с лучшей обобщающей способностью. Дополнительные результаты из основных работ автора также подтверждают, что оптимумы, достигнутые после обучения с разными значениями эффективного темпа обучения, не только различаются по описанным характеристикам, но и геометрически находятся в разных линейно несвязных регионах ландшафта функции потерь.

Выбранное значение эффективного темпа обучения влияет не только на скорость сходимости и свойства итогового решения: вся оптимизационная траектория цели-

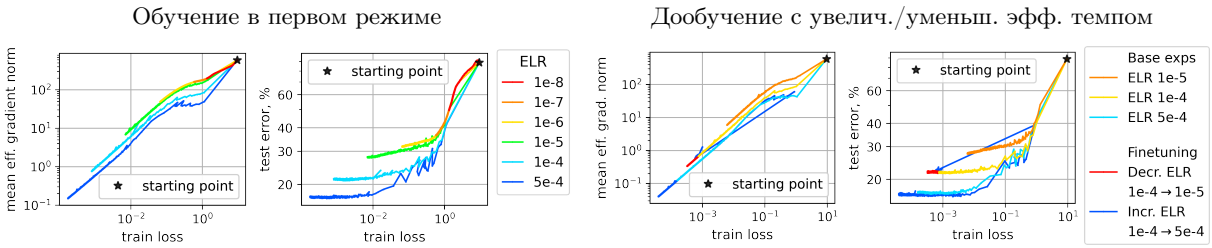


Рис. 7: Первый режим обучения. Слева: обучение с фиксированным эффективным темпом сходится к областям с разными характеристиками кривизны и генерализации. Справа: дообучение с уменьшенным темпом остается на той же траектории, а с увеличенным выскакивает и сходится к более оптимальной области.

ком существенно отличается для запусков с разными значениями. Дабы проанализировать траектории обучения независимо от скорости оптимизации, рассмотрим эволюцию кривизны и генерализации в сравнении с потерями на обучении. Соответствующие графики продемонстрированы для различных значений эффективного темпа на Рисунке 7, слева. Для самых низких значений траектории совпадают, при этом обучение слишком медленное, чтобы сойтись в оптимум. Для остальных значений по мере увеличения эффективного темпа траектории начинают смещаться вниз и влево на диаграммах, при этом в области низких значений функции потерь они выглядят как параллельные линии. Такие различия демонстрируют, что обучение с более высоким эффективным темпом не только приводит к более оптимальной конечной точке по характеристикам кривизны и обобщающей способности, но и лучше обуславливает всю траекторию оптимизации. Также был проведен эксперимент с дообучением из конца траектории, присущей данному значению эффективного темпа, с другим значением (см. Рис. 7, справа). Дообучение с более низким значением остается в том же регионе и продолжает двигаться по той же траектории, не находя других минимумов. Дообучение с более высоким значением через короткое время выходит за пределы окрестности текущего оптимума и сходится в новую область, характеристики которой соответствуют новому эффективному темпу обучения.

Таким образом, анализ первого режима позволяет сделать вывод о том, что внутренняя структура ландшафта функции потерь масштабно-инвариантных нейронных сетей содержит множество оптимумов, различающихся по характеристикам кривизны и обобщающей способности, которые образуют взаимно однозначное соответствие с выбранным значением эффективного темпа обучения.

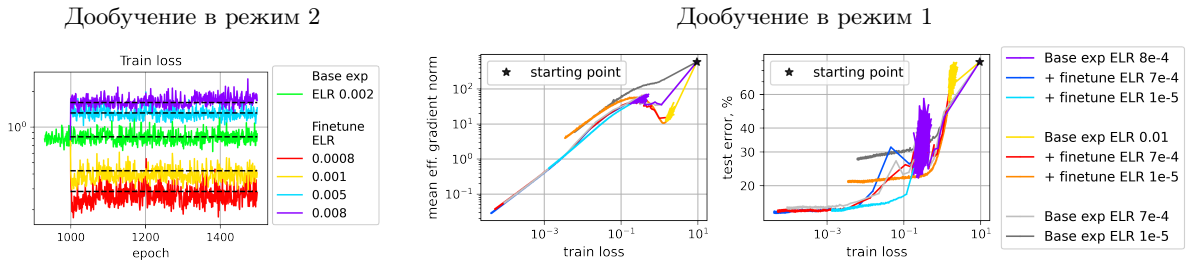


Рис. 8: Дообучение из второго режима с другим значением эффективного темпа обучения. Слева: во второй режим приводит к стабилизации на новом уровне. Справа: в первый режим приводит к сходимости в различные оптимумы при старте из высокого значения и в один оптимум при старте из низкого значения.

Во втором режиме оптимизация шумно стабилизируется вокруг определенного значения по различным метрикам обучения, к примеру, по функции потерь на обучении (см. Рис. 5), что получило название *хаотическое равновесие*. В данном режиме значение эффективного темпа обучения уже слишком высокое, чтобы сойтись в оптимум, но все еще недостаточно, чтобы разойтись окончательно. Таким образом, процесс оптимизации стабилизируется в определенном регионе ландшафта функции потерь с практически фиксированным уровнем метрик обучения, который определяется выбранным значением эффективного темпа. Дополнительные эксперименты из основных работ показывают, что этот регион локально выпуклый для умеренных значений эффективного темпа обучения во втором режиме.

Чтобы продемонстрировать, что значение эффективного темпа обучения однозначно определяет уровень стабилизации функции потерь, был проведен эксперимент с дообучением с другим значением эффективного темпа из второго режима (см. Рис. 8, слева). Изменение значения эффективного темпа обучения соответствующим образом изменяет оптимизационную динамику, а именно, выводит ее на уровень, соответствующий новому значению. Также были проведены эксперименты с дообучением из разных стартовых значений эффективного темпа второго режима с новым значением из первого. Как видно из правой части Рисунка 8, итоговые оптимумы сильно зависят от стартового значения. Дообучение моделей, предварительно обученных с низкими значениями эффективного темпа во втором режиме, всегда сходится в точки с одинаковыми характеристиками кривизны и генерализации. Для

больших стартовых значений дообучение приводит к множеству траекторий в зависимости от выбранного эффективного темпа обучения из первого режима.

Таким образом, проанализировав второй режим обучения, можно заключить, что в ландшафте функции потерь на сфере выделяются регионы стабилизации, в которых оптимизационная динамика фиксируется на определенном уровне в зависимости от выбранного значения эффективного темпа обучения. Такие регионы могут быть как локальными (выпуклые, позволяют сойтись в единственный тип минимума), так и глобальными (невыпуклые, содержат множество различных минимумов).

В дополнение к вышесказанному можно предположить, что основное различие между первым и вторым режимами оптимизации связано с наличием зон повышенной кривизны в траекториях оптимизации нейронных сетей. На диаграмме нормы эффективного градиента против потерь на обучении на Рисунке 5, справа, можно наблюдать, что кривизна достигает своего пика именно в точке перехода между первыми двумя режимами. Это позволяет заключить, что обучение лишь с достаточно малыми значениями эффективного темпа позволяет преодолеть это узкое место и войти в режим сходимости. В основных работах данный тезис обосновывается дополнительно с точки зрения анализа перехода между режимами, а также взаимосвязи между этим переходом и двойным спуском ошибки по эпохам обучения [4].

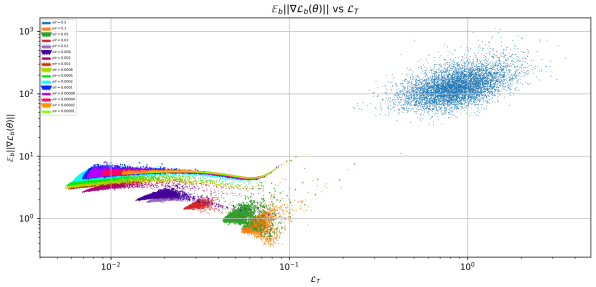
Для наиболее высоких значений эффективного темпа обучения наблюдается наиболее нестабильное поведение оптимизации, соответствующее случайному угадыванию (см. Рис. 5). Так проявляется третий режим обучения, называемый *режимом расходимости*. В основных работах автора диссертации данный режим напрямую сравнивается со случайным блужданием и методом градиентного подъема: во всех трех случаях соседние итерации оказываются некоррелированными, однако случайное блуждание дает оценку снизу по значению функции потерь на обучении, а градиентный подъем — сверху. Таким образом, поведение модели а третьем режиме можно рассматривать как нечто среднее между случайным угадыванием и расходимостью.

Случай квадратичной функции потерь

В заключение данного раздела диссертации кратко выделим основные результаты, полученные при обучении на сфере с использованием квадратичной функции потерь.

Как видно из Рисунка 9, обучение масштабно-инвариантной модели ConvNet на сфере с использованием квадратичной функции потерь не позволило сойтись в оп-

Обучение на полной выборке CIFAR-10



Обучение на подвыборке CIFAR-10

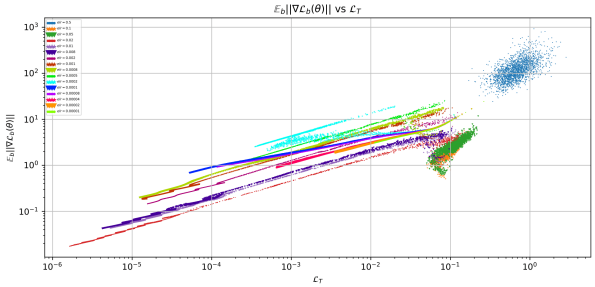


Рис. 9: Диаграмма функции потерь на обучении против средней нормы эффективного градиента. Обучение с квадратичной функцией потерь на сфере. Слева: обучение на полной выборке данных не позволяет сойтись в оптимум и обнаружить первый режим обучения. Справа: обучение на подвыборке из 5000 объектов позволяет выделить все три режима обучения.

тимум за отведенный бюджет по итерациям. Таким образом, можно четко выделить лишь второй и третий режимы оптимизации. Однако после сокращения объема выборки до 5000 объектов модель оказалась способна сойтись при определенных значениях эффективного темпа обучения и продемонстрировать наличие также первого режима в полном соответствии со случаем кросс-энтропийной функции потерь. Во второй работе автора по теме данного раздела приводятся дополнительные эксперименты на основе анализа косинусных расстояний между итерациями, которые позволяют более тонко провести различие между режимами и показать наличие аналога первого режима даже без непосредственно сходимости.

Заключение и прочие результаты

В данном разделе был разобран случай обучения полностью масштабно-инвариантных моделей на сфере методом проекции стохастического градиента с фиксированным значением эффективного темпа обучения. Было выделено три режима такого обучения в зависимости от конкретного значения эффективного темпа: сходимость, хаотическое равновесие и расходимость. Данные режимы были проанализированы теоретически и экспериментально, выделены их основные характеристики, а также получены сведения о внутреннем устройстве ландшафта функции потерь масштабно-инвариантных нейронных сетей. Среди них наличие целого спектра различных глобальных оптимумов, зон повышенной кривизны и областей стабилизации оптимизационной динамики — как локальных, так и глобальных. Помимо классической кросс-

энтропийной был также исследован случай использования квадратичной функции потерь на задачах классификации с точки зрения трех режимов обучения.

В двух основных статьях автора по данной теме (второй и третьей по порядку в общем списке) приводится множество дополнительных результатов, опущенных в данном разделе, включая проявление трех режимов при стандартном обучении нейронных сетей, интерпретацию и поиск с их помощью оптимальных расписаний темпа обучения, линейную связность в разных режимах, анализ переходов между режимами, расписания эффективного темпа обучения и прочее.

4 Заключение

В заключительном разделе подытоживаются основные результаты работы.

1. Проведено исследование динамики стандартного обучения нейронных сетей с использованием техник нормализации и сокращения веса. Открыто, объяснено и проанализировано как экспериментально, так и теоретически возникающее периодическое поведение при таком обучении.

2. Предложено разрешение сложившегося в литературе противоречия об итоге такого обучения (равновесие против нестабильности) через принцип обобщенного равновесия в периодическом поведении. Данный тезис обоснован как экспериментально, так и теоретически.

3. Проведено исследование динамики обучения полностью масштабно-инвариантных нейронных сетей на сфере фиксированного радиуса с использованием как кросс-энтропийной, так и квадратичной функции потерь. Открыты и изучены с использованием как теоретического, так и экспериментального анализа три режима такого обучения: сходимость, хаотическое равновесие и расходимость.

4. Благодаря исследованию данных режимов более детально раскрыто устройство ландшафта функции потерь масштабно-инвариантных моделей на естественном домене. В частности, показано наличие целого спектра различных по своим свойствам глобальных минимумов, зон повышенной кривизны, локальных и глобальных областей стабилизации оптимизационной динамики.

Список литературы

[1] Like Hui and Mikhail Belkin. Evaluation of neural architectures trained with square loss vs cross-entropy in classification tasks. In *International Conference on Learning Representations*, 2021.

[2] Leo Breiman. Reflections after refereeing papers for nips. In *The Mathematics of Generalization*, pages 11–15. CRC Press, 2018.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Science*, 116(32), 2019.

[4] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. In *International Conference on Learning Representations*, 2020.

[5] Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

[6] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[7] W Ronny Huang, Zeyad Ali Sami Emam, Micah Goldblum, Liam H Fowl, Justin K Terry, Furong Huang, and Tom Goldstein. Understanding generalization through visualizations. In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*, 2020.

[8] Felix Draxler, Kambis Veschgini, Manfred Salmhofer, and Fred Hamprecht. Essentially no barriers in neural network energy landscape. In *International Conference on Machine Learning*. PMLR, 2018.

[9] Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry P Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Advances in Neural Information Processing Systems*, 2018.

[10] Mikhail Belkin. Fit without fear: remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

[11] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. PMLR, 2015.

[12] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016.

[13] Tim Salimans and Diederik P Kingma. Weight normalization: a simple reparameterization to accelerate training of deep neural networks. *Advances in Neural Information Processing Systems*, 29, 2016.

[14] Nils Bjorck, Carla P Gomes, Bart Selman, and Kilian Q Weinberger. Understanding batch normalization. *Advances in Neural Information Processing Systems*, 31, 2018.

[15] Shibani Santurkar, Dimitris Tsipras, Andrew Ilyas, and Aleksander Madry. How does batch normalization help optimization? *Advances in Neural Information Processing Systems*, 31, 2018.

[16] Greg Yang, Jeffrey Pennington, Vinay Rao, Jascha Sohl-Dickstein, and Samuel S. Schoenholz. A mean field theory of batch normalization. In *International Conference on Learning Representations*, 2019.

[17] Xiang Li, Shuo Chen, and Jian Yang. Understanding the disharmony between weight normalization family and weight decay. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4715–4722, 2020.

[18] Zhiyuan Li and Sanjeev Arora. An exponential learning rate schedule for deep learning. In *International Conference on Learning Representations*, 2020.

[19] Zhiyuan Li, Kaifeng Lyu, and Sanjeev Arora. Reconciling modern deep learning with traditional optimization analyses: The intrinsic learning rate. *Advances in Neural Information Processing Systems*, 33, 2020.

[20] Twan Van Laarhoven. L2 regularization versus batch and weight normalization. *arXiv preprint arXiv:1706.05350*, 2017.

[21] Elad Hoffer, Ron Banner, Itay Golan, and Daniel Soudry. Norm matters: efficient and accurate normalization schemes in deep networks. *Advances in Neural Information Processing Systems*, 31, 2018.

[22] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. In *International Conference on Learning Representations*, 2019.

[23] Ruosi Wan, Zhanxing Zhu, Xiangyu Zhang, and Jian Sun. Spherical motion dynamics: Learning dynamics of normalized neural network using sgd and weight decay. In *Advances in Neural Information Processing Systems*, volume 34, 2021.

[24] Simon Roburin, Yann de Mont-Marin, Andrei Bursuc, Renaud Marlet, Patrick Pérez, and Mathieu Aubry. A spherical analysis of adam with batch normalization. *arXiv preprint arXiv:2006.13382*, 2020.

[25] Kaifeng Lyu, Zhiyuan Li, and Sanjeev Arora. Understanding the generalization benefit of normalization layers: Sharpness reduction. In *Advances in Neural Information Processing Systems*, volume 35, 2022.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[27] Aitor Lewkowycz and Guy Gur-Ari. On the training dynamics of deep networks with l_2 regularization. *Advances in Neural Information Processing Systems*, 33, 2020.

[28] Vitaliy Chiley, Ilya Sharapov, Atli Kosson, Urs Koster, Ryan Reece, Sofia Samaniego de la Fuente, Vishal Subbiah, and Michael James. Online normalization for training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition*, 2015.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[32] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[33] Elad Hoffer, Itay Hubara, and Daniel Soudry. Fix your classifier: the marginal value of training the last weight layer. In *International Conference on Learning Representations*, 2018.

[34] Yuanzhi Li, Colin Wei, and Tengyu Ma. Towards explaining the regularization effect of initial large learning rate in training neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[35] Nikhil Iyer, V Thejas, Nipun Kwatra, Ramachandran Ramjee, and Muthian Sivathanu. Wide-minima density hypothesis and the explore-exploit learning rate schedule. *arXiv preprint arXiv:2003.03977*, 2020.

[36] Aitor Lewkowycz, Yasaman Bahri, Ethan Dyer, Jascha Sohl-Dickstein, and Guy Gur-Ari. The large learning rate phase of deep learning: the catapult mechanism. *arXiv preprint arXiv:2003.02218*, 2020.

[37] Jeremy Cohen, Simran Kaur, Yuanzhi Li, J Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability. In *International Conference on Learning Representations*, 2021.

[38] Justin Gilmer, Behrooz Ghorbani, Ankush Garg, Sneha Kudugunta, Behnam Neyshabur, David Cardoze, George Edward Dahl, Zachary Nado, and Orhan Firat. A loss curvature perspective on training instabilities of deep learning models. In *International Conference on Learning Representations*, 2022.