ДОКЛАДЫ РОССИЙСКОЙ АКАДЕМИИ НАУК. МАТЕМАТИКА, ИНФОРМАТИКА, ПРОЦЕССЫ УПРАВЛЕНИЯ, 2023, том 514, № 2, с. 99–108

УДК 004.8

МЕТОДЫ, ИСПОЛЬЗУЮЩИЕ ГРАДИЕНТНЫЙ КЛИППИНГ, ДЛЯ ЗАДАЧ СТОХАСТИЧЕСКОЙ ОПТИМИЗАЦИИ С ТЯЖЕЛЫМ ШУМОМ

© 2023 г. М. Ю. Данилова[1,*]

Представлено академиком РАН А.А. Шананиным
Поступило 02.09.2023 г.
После доработки 08.10.2023 г.
Принято к публикации 15.10.2023 г.

Эта статья представляет собой обзор результатов ряда исследований [Gorbunov et al., 2020, 2021, 2022, Sadiev et al., 2023], в которых постепенно устранялись открытые вопросы, связанные с анализом сходимости с большой вероятностью стохастических методов оптимизации первого порядка при слабых предположениях о шуме. В начале мы представим концепцию градиентного клиппинга, которая играет ключевую роль в развитии стохастических методов для успешной работы в случае распределений с тяжелыми хвостами. Далее мы рассмотрим важность получения оценок сходимости методов в вероятностном контексте и их взаимосвязь с оценками сходимости по математическому ожиданию. Заключительные разделы статьи посвящены основным результатам в области задач минимизации и результатам численных экспериментов.

Ключевые слова: выпуклая оптимизация, стохастическая оптимизация, методы первого порядка

DOI: 10.31857/S2686954323601768, **EDN:** GFCAAP

1. ВВЕДЕНИЕ

Одной из основных задач в области искусственного интеллекта на данный момент является объединение теории и практики машинного обучения. В теории методов оптимизации, которые являются важным звеном в машинном обучении, обеспечивая эффективную настройку моделей и обработку данных, образование данной связи проявляется в исследованиях при более *слабых предположениях*, чем стандартные. Эти исследования направлены на анализ методов в более широком классе задач и способствуют более эффективному применению методов машинного обучения. Также стоит отметить, что некоторые явления, возникающие при реализации методов, не могут быть объяснены с помощью классического рассмотрения сходимости по математическому ожиданию [Gorbunov et al., 2020], что вызывает возрастающий интерес к рассмотрению *сходимости с большой вероятностью*. Таким образом, исследования сходимости методов с большой вероятностью при менее строгих условиях на целевую функцию действительно являются важной областью исследования в области искусственного интеллекта. Несмотря на значительный интерес к данной теме [Nazin et al., 2019, Davis et al., 2021, Cutkosky and Mehta, 2021, Nguyen et al., 2023, Liu and Zhou, 2023, Liu et al., 2023], некоторые важные аспекты оставались неисследованными на протяжении продолжительного времени. Это продолжалось до появления серии работ [Gorbunov et al., 2020, 2021, 2022, Sadiev et al., 2023], в которых были предложены и подробно проанализированы методы, основанные на применении градиентного клиппинга, для решения задач минизации и вариационных неравенств в предположении, что градиентный/операторный шум имеет ограниченный центральный α-й момент для $\alpha \in (1,2]$.

2. ГРАДИЕНТНЫЙ КЛИППИНГ И ТЯЖЕЛЫЕ ХВОСТЫ РАСПРЕДЕЛЕНИЯ

В рассматриваемых далее методах одним из ключевых аспектов для достижения желаемых теоретических результатов является градиентный клиппинг. Этот оператор играет важную роль в случаях, когда стохастические градиенты содержат шум с распределением, имеющим тяжелые хвосты. В данной главе мы более подробно изучим оператор клиппинга и определим понятие тяжелых хвостов распределений.

В машинном обучении мы часто сталкиваемся с ситуацией, когда, решая задачу минимизации, у нас есть доступ только к стохастическому гради-

[1]*Московский физико-технический институт, Москва, Россия*

*****E-mail: danilovamarina15@gmail.com*

енту. В таком случае мы можем решать данную задачу, генерируя последовательность $\{x_k\}_{k\geqslant 0}$ с помощью стохастического градиентного спуска SGD, который является одним из наиболее распространенных методов оптимизации, применяемых в случае стохастической оптимизации. Классический стохастический градиентный спуск выглядит следующим образом

$$x^{k+1} = x^k - \gamma \cdot \nabla f(x^k, \xi^k), \qquad \text{(SGD)}$$

где $f(x)$ – целевая функция, γ – размер шага, $\nabla f(x^k, \xi^k)$ – стохастический градиент, т.е. *несмещенная* оценка $\nabla f(x^k)$: $\mathbb{E}_{\xi^k}[\nabla f(x^k, \xi^k)] = \nabla f(x^k)$.

Градиентный клиппинг

Путем взятия нелинейности от стохастического градиента мы можем получить новый метод с градиентным клиппингом, который применяется для контроля нормы градиента

$$x^{k+1} = x^k - \gamma \cdot \text{clip}\left(\nabla f(x^k, \xi^k), \lambda\right), \quad \text{(clipped-SGD)}$$

$$\text{clip}(x, \lambda) = \begin{cases} \min\left\{1, \dfrac{\lambda}{\|x\|}\right\} x, & \text{если } x \neq 0, \\ 0, & \text{иначе.} \end{cases} \qquad (1)$$

Данная техника полезна в случаях, когда градиент может быть слишком большим и вызывать проблемы с обучением, такие как взрывной градиент. Градиентный клиппинг ограничивает норму градиента до определенного предела $\lambda > 0$, чтобы избежать таких проблем. Это позволяет контролировать скорость обновления параметров модели и повышает стабильность процесса обучения. Отметим, что оператор $\text{clip}(\nabla f(x^k, \xi^k), \lambda)$ – *смещенная* оценка $\nabla f(x^k)$: $\mathbb{E}_{\xi^k}[\text{clip}(\nabla f(x^k, \xi^k), \lambda)] \neq \nabla f(x^k)$, усложняющая анализ clipped-SGD. Клиппинг набрал популярность в области искусственного интеллекта и машинного обучения с момента публикации статьи [Pascanu et al., 2013] в 2013 г. В этой статье было предложено использовать градиентный клиппинг для решения проблемы взрывающихся градиентов при обучении рекуррентных нейронных сетей. В случае метода без клиппинга значения целевой функции сильно колеблются на разных итерациях обучения, что указывает на нестабильность процесса обучения. В то же время метод с клиппингом градиента проявляет более стабильную сходимость. Это свидетельствует о робастности и надежности сходимости метода с клиппингом градиента. Клиппинг градиента и подобные подходы широко применяются в различных задачах обработки естественного языка (NLP). В 2017 г. было предложено использование клиппинга в LSTM моделях [Merity et al., 2018], а затем в двунаправленной языковой модели [Peters et al., 2017]. В 2020 г. он был применен для тонкой настройки модели `BERT` [Mosbach et al., 2020]. Эти исследования и применения свидетельствуют о тенденции использования клиппинга или подобных подходов в задачах NLP. Это обусловлено необходимостью обеспечения стабильности и робастности обучения моделей, особенно в случаях с большими и сложными наборами данных, типичными для NLP задач.

В работе [Zhang et al., 2020] представлены результаты эксперимента, целью которого было исследование влияния градиентного клиппинга. Для модели `ResNet50` на наборе данных `ImageNet` эффективно работает стандартный SGD с моментумом, что делает клиппинг необязательным. Но при обучении `BERT` на `Wikipedia+Books` градиенты имеют тяжелые хвосты, и метод `ADAM`, как адаптивный покомпонентный клиппинг, обеспечивает более эффективное обучение и сходимость. `ADAM`, адаптируя скорость обучения для каждого параметра индивидуально, позволяет более эффективно работать с тяжелыми хвостами распределений градиентов и обеспечивает более стабильное и быстрое обучение модели `BERT`. Правило обновления для ADAM согласно [Kingma and Ba, 2014] записывается в следующем виде

$$m_k = \beta_1 m_{k-1} + (1-\beta_1)\nabla f(x^k, \xi^k),$$

$$V_k = \beta_2 V_{k-1} + (1-\beta_2)(\nabla f(x^k, \xi^k))^2,$$

$$x^{k+1} = x^k - \frac{\gamma}{\sqrt{V_k + \delta}} m^k,$$

где все операции с векторами (возведение в квадрат, извлечение корня, деление на вектор) происходят покомпонентно $\odot$. Когда $\beta_1 = 0$ `Adam`(`RMSprop`) можно рассматривать как `clipped-SGD` с “адаптивным” и покомпонентным λ_k.

Тяжелые хвосты распределения

Далее давайте определим, в каком случае мы будем считать, что распределение имеет тяжелый хвост. Мы будем говорить, что случайный вектор X имеет легкие (субгауссовские) хвосты распределения, если выполняет следующее неравенство:

$$\mathbb{P}\{\|X - \mathbb{E}[X]\| \geqslant b\} \leqslant 2\exp\left(-\frac{b^2}{2\sigma^2}\right) \quad \forall b > 0, \qquad (2)$$

которое эквивалентно (с точностью до числового коэффициента в σ)

$$\mathbb{E}\left[\exp\left(\frac{\|X - \mathbb{E}[X]\|^2}{\sigma^2}\right)\right] \leqslant \exp(1). \qquad (3)$$

В общем случае мы говорим, что случайный вектор X имеет тяжелые хвосты распределения, если не выполнено условие (2) и существует конечная дисперсия. Однако были получены результаты при более общем условии, что он имеет ограниченный центральный α-й момент для некоторого $\alpha \in (1,2]$:

$$\mathbb{E}\left[\|X - \mathbb{E}[X]\|^{\alpha}\right] \leqslant \sigma^{\alpha}. \quad (4)$$

Когда $\alpha = 2$, приведенное выше предположение восстанавливает стандартное предположение о равномерно ограниченной дисперсии. Однако предположение (4) допускает, чтобы дисперсия была неограниченной, когда $\alpha \in (1,2)$.

3. СХОДИМОСТЬ С БОЛЬШОЙ ВЕРОЯТНОСТЬЮ И СХОДИМОСТЬ ПО МАТЕМАТИЧЕСКОМУ ОЖИДАНИЮ

В этом обзоре мы рассматриваем исследования, в которых были получены результаты, основанные на сходимости с большой вероятностью. В данной главе мы обсудим важность обеспечения гарантий сходимости с большой вероятностью и недостатки оценок, связанных со сходимостью по математическому ожиданию.

Задача и ограничения

В первую очередь давайте формально определим рассматриваемую нами задачу и уточним предположения, с которыми мы работаем. Мы рассматриваем задачу безусловной стохастической оптимизации, где целевая функция представляет собой математическое ожидание относительно переменной ξ

$$\min_{x\in\mathbb{R}^n}\{f(x) = \mathbb{E}_{\xi}[f(x,\xi)]\}, \quad (5)$$

где $f: \mathbb{R}^n \to \mathbb{R}$ выпуклая и L-гладкая, т.е. $\forall x, y \in \mathbb{R}^n$

$$f(x) \geqslant f(y) + \langle \nabla f(y), x - y\rangle, \quad (6)$$

$$\|\nabla f(x) - \nabla f(y)\| \leqslant L\|x - y\|. \quad (7)$$

Стохастический градиент $\nabla f(x,\xi)$ с ограниченным центральным α-м моментом ($\alpha \in (1,2]$), т.е. $\forall x \in \mathbb{R}^n$

$$\begin{gathered}\mathbb{E}_{\xi}[\nabla f(x,\xi)] = \nabla f(x),\\ \mathbb{E}_{\xi}\left[\|\nabla f(x,\xi) - \nabla f(x)\|^{\alpha}\right] \leqslant \sigma^{\alpha}.\end{gathered} \quad (8)$$

Сходимость итеративного метода

Теперь давайте уточним, что мы понимаем под сходимостью итеративного метода. Какие гарантии на сходимость могут быть предоставлены? Мы будем рассматривать два подхода к анализу этих методов.

Сходимость в среднем. Сходимость в среднем (по математическому ожиданию) подразумевает определение числа итераций N (или вызовов оракула), необходимых для достижения такой точки x^N, при которой выполняется одно из следующих условий: $\mathbb{E}[\|x^N - x^*\|^2] \leqslant \varepsilon$, $\mathbb{E}[f(x^N) - f(x^*)] \leqslant \varepsilon$, $\mathbb{E}[\|\nabla f(x^N)\|^2] \leqslant \varepsilon$. Обычно такие гарантии зависят только от определенных моментов стохастического градиента, таких как дисперсия.

Сходимость с большой вероятностью. Сходимость с большой вероятностью подразумевает определение числа итераций N (или вызовов оракула), необходимых для достижения такой точки x^N, при которой выполняется одно из следующих условий: $\mathbb{P}\{\|x^N - x^*\|^2 \leqslant \varepsilon\} \geqslant 1 - \beta$, $\mathbb{P}\{f(x^N) - f(x^*) \leqslant \varepsilon\} \geqslant 1 - \beta$, $\mathbb{P}\{\|\nabla f(x^N)\|^2 \leqslant \varepsilon\} \geqslant 1 - \beta$. Такие гарантии чувствительны к распределению шума в стохастических градиентах.

Недостатки сходимости по математическому ожиданию

Для того, чтобы продемонстрировать, что гарантии, связанные с математическим ожиданием, нечувствительны к распределению шума, давайте рассмотрим простой пример. Возьмем следующую задачу минимизации и применим алгоритм SGD с постоянным шагом

$$\min_{x\in\mathbb{R}^n}\{f(x) = \mathbb{E}_{\xi}[f(x,\xi)]\}, \quad f(x,\xi) = \frac{1}{2}\|x\|^2 + \langle \xi, x\rangle, \quad (9)$$

где $\mathbb{E}[\xi] = 0$ и $\mathbb{E}[\|\xi\|^2] = \sigma^2$. Нас интересует случай ограниченной дисперсии в (4) ($\alpha = 2$), так как согласно примеру в работе [Zhang et al., 2020] в случае не субгауссовского распределения случайной величины ξ, SGD может не сходиться по математическому ожиданию, когда $\alpha < 2$. В данном примере рассмотрим 3 варианта распределения ξ: нормальное (гауссовское) распределение, распределение Вейбулла (не субгауссовское), распределение Бёрра XII (не субгауссовоское). Оказывается, что все три случая не будут отличимы для гарантий по математическому ожиданию, поскольку известные гарантии для алгоритма SGD [Ghadimi and Lan, 2013] не зависят от распределения шума

$$\begin{gathered}\mathbb{E}[f(x^k) - f(x^*)] \leqslant\\ \leqslant (1-\gamma)^k(f(x^0) - f(x^*)) + \frac{\gamma\sigma^2}{2}.\end{gathered} \quad (10)$$

В представленной оценке действительно нет явной информации о распределении. Однако как

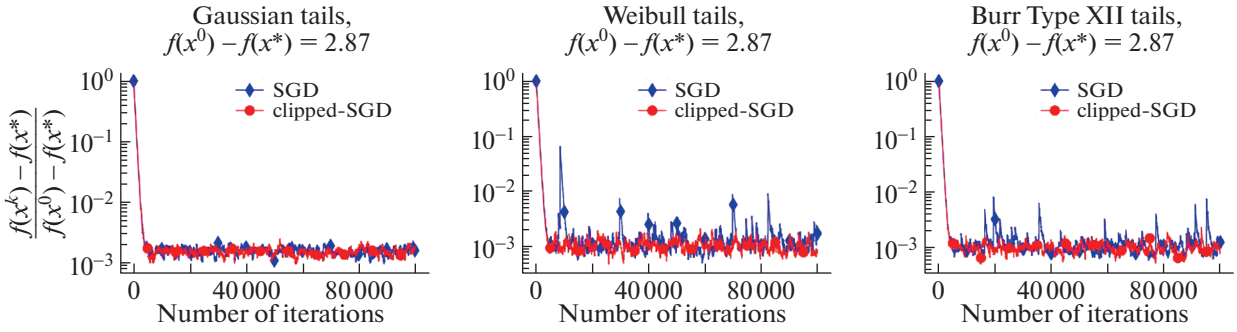

Рис. 1. Траектории SGD и clipped-SGD, применяемые для решения задачи (9) с тремя различными распределениями случайной величиной ξ (нормальным, Вейбулла и Бёрра XII) [Gorbunov et al., 2020].

показано на рис. 1, на практике поведение метода может зависеть от распределения шума.

Анализируя графики 1, мы приходим к выводу, что при запуске метода нам бы хотелось уже после первого запуска иметь гарантии с высокой вероятностью получить хорошее решение. В случае обычного SGD мы не можем гарантировать это без уменьшения шага для сдерживания всплесков, что приводит к потери скорости. Однако в случае clipped-SGD с правильным уровнем клиппинга, мы можем избежать необходимости уменьшения шага. Чтобы оценить эту разницу, мы будем рассматривать оценки сходимости с большой вероятностью.

Результаты о сходимости с большой вероятностью в предположении легких хвостов

Прежде чем приступить к обсуждению новейших результатов, давайте посмотрим, что было известно на момент начала обозреваемого цикла работ, т.е. к 2020 г. При предположении о легких хвостах распределения (3), а также при условии выпуклости и L-гладкости функции f, было известно следующее Devolder et al. [2011] показали, что алгоритм SGD достигает точки $\hat{x}$, для которой выполняется неравенство $f(\hat{x}) - f(x^*) \leqslant \varepsilon$ с вероятностью не менее $1 - \beta$, используя

$$\mathcal{O}\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln^2\left(\frac{1}{\beta}\right)\right\}\right)$$

вызовов оракула.

Ghadimi and Lan [2012] показали, что алгоритм SGD достигает точки $\hat{x}$, для которой выполняется неравенство $f(\hat{x}) - f(x^*) \leqslant \varepsilon$ с вероятностью не менее $1 - \beta$, используя

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}, \frac{\sigma^2 R_0^2}{\varepsilon^2}\ln^2\left(\frac{1}{\beta}\right)\right\}\right)$$

вызовов оракула.

Неравенство Маркова

На текущем этапе важно отметить, что нельзя просто применить неравенство Маркова и получить гарантии сходимости с большой вероятностью на основе сходимости по математическому ожиданию. Используя неравенство Маркова

$$\mathbb{P}\{f(\hat{x}) - f(x^*) > \varepsilon\} < \frac{\mathbb{E}[f(\hat{x}) - f(x^*)]}{\varepsilon} \qquad (11)$$

и сделав достаточное количество шагов в алгоритме SGD, мы можем обеспечить гарантию $\mathbb{E}[f(\hat{x}) - f(x^*)] \leqslant \varepsilon\beta$, что даст нам $\mathbb{P}\{f(\hat{x}) - f(x^*) > \varepsilon\} \leqslant \beta$ или $\mathbb{P}\{f(\hat{x}) - f(x^*) \leqslant \varepsilon\} \geqslant 1 - \beta$. Для достижения точности $\mathbb{E}[f(\hat{x}) - f(x^*)] \leqslant \varepsilon\beta$ методу SGD необходимо

$$\mathcal{O}\left(\max\left\{\frac{LR_0^2}{\varepsilon\beta}, \frac{\sigma^2 R_0^2}{\varepsilon^2\beta^2}\right\}\right) \quad \text{вызовов оракула.}$$

К сожалению, зависимость от β^{-a}, $a > 0$ не подходит, так как необходимо получить логарифмическую зависимость от $\frac{1}{\beta}$, потому что β должно быть маленьким. Поэтому неравенство Маркова не позволяет получить нужные гарантии о сходимости с большой вероятностью. Однако возникает вопрос: можно ли более точно проанализировать сходимость алгоритма SGD с большой вероятностью? Да, следующая теорема содержит результаты о сходимости SGD в предположении тяжелых хвостов, но при этом подчеркивает необходимость модификации классического алгоритма SGD. Градиентный клиппинг, описанный в предыдущей главе, является подходящим решением этой задачи.

Теорема 3.1. *Для любых $\varepsilon > 0$, $\beta \in (0,1)$ и алгоритма* SGD *параметризованного количеством шагов K и длиной шага γ, существует μ-сильно выпуклая L-гладкая задача $\min\limits_{x \in Q} f(x)$, а также стохастический оракул с шумом, имеющим ограниченный α-й момент с $\alpha = 2$, при условии $0 < \mu \leqslant L$ такие,*

что для алгоритма SGD *с длиной шага* $0 \leqslant \gamma \leqslant \frac{1}{\mu}$ *выполнено следующее*

$$\mathbb{P}\left\{\left\|x^K - x^*\right\|^2 \geq \varepsilon\right\} \leq \beta \Rightarrow K = \Omega\left(\frac{\sigma}{\mu\sqrt{\beta\varepsilon}}\right). \qquad (12)$$

Результаты о сходимости с большой вероятностью в предположении тяжелых хвостов

Давайте рассмотрим известные результаты о гарантиях сходимости алгоритма SGD с большой вероятностью. В случае, когда распределение имеет тяжелые хвосты, т.е. не выполнено условие (2), и дисперсия равномерно ограничена (4) ($\alpha = 2$), первые результаты были получены в 2019 г. Nazin et al. [2019] предложили Robust Stochastic Mirror Descent (RSMD), напоминающий clipped-SGD, и доказали следующую оценку

$$\mathbb{O}\left(\max\left\{\frac{LD^2}{\varepsilon}, \frac{\sigma^2 D^2}{\varepsilon^2}\right\}\ln\left(\frac{1}{\beta}\right)\right).$$

Это была первая работа в данной области. Из недостатков стоит отметить, что доказательство опирается на диаметр множества $D < +\infty$ и нет ускорения метода. В скором времени был получен новый результат для другого метода. Davis et al. [2021] предложили proxBoost, основанный на робастном оценивании расстояния и методе проксимальной точки. Авторы данной работы доказали следующую оценку (в сильно выпуклом случае):

$$\mathbb{O}\left(\max\left\{\sqrt{\frac{L}{\mu}}\ln\left(\frac{LR_0^2\ln\frac{L}{\mu}}{\varepsilon}\right), \frac{\sigma^2\ln\frac{L}{\mu}}{\mu\varepsilon}\right\}\ln\left(\frac{L}{\mu}\right)\ln\left(\frac{\ln\frac{L}{\mu}}{\beta}\right)\right).$$

Данный результат ускоренный и получен для любой выпуклой замкнутой области (ограниченной/неограниченной). При этом стоит отметить, что предложенный метод требует решения вспомогательной задачи на каждой итерации, а так же в оценке появился дополнительный логарифм числа обусловленности.

4. ОСНОВНЫЕ РЕЗУЛЬТАТЫ ДЛЯ ЗАДАЧ МИНИМИЗАЦИИ

В данной главе мы обсудим основные результаты для задач минимизации, методы и сложности, связанные с их анализом. Давайте начнем с неускоренного метода clipped-SGD. В связи с тем, что его проще анализировать, чем другие методы, на его примере мы сможем понять, с какими сложностями можно столкнуться при анализе методов с клиппингом. Шаг метода clipped-SGD

$$x^{k+1} = x^k - \gamma \cdot \underbrace{\texttt{clip}(\nabla f(x^k, \xi^k), \lambda)}_{\widetilde{\nabla} f(x^k, \xi^k)}.$$

Смещенность оценки $\mathbb{E}[\widetilde{\nabla} f(x^k, \xi^k) | x^k] \neq \nabla f(x^k)$ – основная причина сложностей, возникающих при теоретическом анализе. Для того, чтобы понять, где именно возникает эта проблема, давайте посмотрим классическое доказательство сходимости SGD

$$\|x^{k+1} - x^*\|^2 = \|x^k - x^*\|^2 - 2\gamma\langle x^k - x^*, \widetilde{\nabla} f(x^k, \xi^k)\rangle + \gamma^2\|\widetilde{\nabla} f(x^k, \xi^k)\|^2.$$

Используя выпуклость гладкость f и перегруппировывая слагаемые, получается следующее выражение:

$$\frac{2\gamma(1-2\gamma L)}{N}\sum_{k=0}^{N-1}\Delta_k \leqslant \frac{1}{N}(R_0^2 - R_N^2) + \frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2,$$

где $\Delta_k = f(x^k) - f(x^*)$, $R_k = \|x^k - x^*\|$, $\theta_k = \widetilde{\nabla} f(x^k, \xi^k) - \nabla f(x^k)$. Красным цветом выделена стохастическая часть суммы, но как оценить сверху эти слагаемые? Это достигается с помощью применения неравенства Бернштейна.

Неравенство Бернштейна для мартингальных разностей

Лемма 4.1. [Bennett, 1962, Dzhaparidze and Van Zanten, 2001, Freedman et al., 1975] *Пусть последовательность случайных величин* $\{X_i\}_{i\geq 1}$ *образует последовательность мартингальных разностей, т.е.* $\mathbb{E}[X_i | X_{i-1}, \ldots, X_1] = 0$ *для любых* $i \geq 1$. *Предположим, что существуют условные ограниченные дисперсии* $\sigma_i^2 \overset{\text{def}}{=} \mathbb{E}[X_i^2 | X_{i-1}, \ldots, X_1]$, *а также существует константа* $c > 0$ *такая, что* $|X_i| \leq c$ *для любых* $i \geq 1$. *Тогда для любых* $b > 0$, $G > 0$ *и* $N \geq 1$

$$\mathbb{P}\left\{\left|\sum_{i=1}^N X_i\right| > b \text{ и } \sum_{i=1}^N \sigma_i^2 \leq G\right\} \leq 2\exp\left(-\frac{b^2}{2G + 2cb/3}\right).$$

Для того, чтобы применить лемму (4.1) для сумм $\frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2$, необходимо ограничить смещение, дисперсию, $\|x^k - x^*\|$ и $\|\theta_k\|$ с большой вероятностью. Следующая лемма была доказана для работы со смещением, дисперсией и т.д.

Лемма 4.2. *Пусть* X *случайный вектор в* $\mathbb{R}^n$ *и* $\widetilde{X} = \text{clip}(X, \lambda)$. *Тогда,*

$$\left\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\right\| \leqslant 2\lambda. \tag{13}$$

Более того, если для некоторых $\sigma \geqslant 0$ *и* $\sigma \in [1,2)$ *выполняется*

$$\mathbb{E}[X] = x \in \mathbb{R}^n, \quad \mathbb{E}[\|X - x\|^\alpha] \leqslant \sigma^\alpha \tag{14}$$

и $\|x\| \leqslant \lambda/2$, *тогда*

$$\left\|\mathbb{E}[\widetilde{X}] - x\right\| \leqslant \frac{2^\alpha \sigma^\alpha}{\lambda^{\alpha-1}}, \tag{15}$$

$$\mathbb{E}[\|\widetilde{X} - \mathbb{E}[\widetilde{X}]\|^2] \leqslant 18\lambda^{2-\alpha}\sigma^\alpha. \tag{16}$$

Остается ограничить расстояние до решения. В силу того, что $\Delta_k \geqslant 0\ \forall k \geqslant 0$, полученное ранее неравенство

$$\frac{2\gamma(1-2\gamma L)}{N}\sum_{k=0}^{N-1}\Delta_k \leqslant \frac{1}{N}(R_0^2 - R_N^2) + \\ + \frac{2\gamma}{N}\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + \frac{2\gamma^2}{N}\sum_{k=0}^{N-1}\|\theta_k\|^2$$

можно переписать следующим образом

$$R_N^2 \leqslant R_0^2 + 2\gamma\sum_{k=0}^{N-1}\langle x^* - x^k, \theta_k\rangle + 2\gamma^2\sum_{k=0}^{N-1}\|\theta_k\|^2.$$

Основная идея: доказать по индукции $R_N \leqslant CR_0$ с большой вероятностью для некоторой числовой константы C. Это означает, что точки, генерируемые методом, остаются в некотором шаре вокруг решения. Мы представили общий подход к получению вероятностных оценок сходимости методов, использующих градиентный клиппинг в предположении, что шум имеет ограниченный центральный α-й момент для $\alpha \in (1,2]$.

Сходимость clipped-SGD с большой вероятностью

Теорема 4.1. *Пусть* f *выпуклая и* L*-гладкая на* $B_{7R_0}(x^*) = \{x \in \mathbb{R}^n \,|\, \|x - x^*\| \leqslant 7R_0\}$ *и* (8) *выполняется* на $B_{7R_0}(x^*)$. *Тогда для любых* $\beta \in (0,1)$, $\varepsilon \geqslant 0$ *таких, что* $\ln(LR_0^2/\varepsilon\beta) \geqslant 2$ *существует такой выбор* γ, *что* clipped-SGD *с уровнем клиппинга* $\lambda \sim 1/\gamma$ *достигает* $\overline{x}^N$, *удовлетворяющую* $f(\overline{x}^N) - f(x^*) \leqslant \varepsilon$ *с вероятностью не менее* $1-\beta$, *используя*

$$\mathbb{O}\left(\max\left\{\frac{LR_0^2}{\varepsilon}, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right)$$

итераций/вызовов оракула.

Полученный результат соответствует (с точностью до логарифмических множителей) результатам для SGD в случае легких хвостов и для RSMD в случае тяжелых хвостов, но для безусловной задачи. Отметим, что все предположения достаточно требовать на шаре вокруг решения.

Ускоренный clipped-SGD: clipped-SSTM

Теперь, имея представление о рассмотренных ранее идеях доказательства, мы готовы перейти к рассмотрению ускоренного метода. На основе метода подобных треугольников (Stochastic Similar Triangles Method) предложенного Gasnikov and Nesterov [2016], и с использованием градиентного клиппинга был получен ускоренный вариант clipped-SGD–clipped-SSTM

$$\alpha_{k+1} = \frac{k+2}{2aL}, \quad A_{k+1} = A_k + \alpha_{k+1}, \quad \lambda_{k+1} = \frac{B}{\alpha_{k+1}},$$
$$A_0 = a_0 = 0,\ y^0 = z^0 = x^0,$$
$$x^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^k}{A_{k+1}},$$
$$z^{k+1} = z^k - \alpha_{k+1}\underbrace{\widetilde{\nabla} f(x^{k+1}, \xi^k)}_{\mathrm{clip}(\nabla f(x^{k+1},\xi^k),\lambda_{k+1})}, \qquad \text{(clipped-SSTM)}$$
$$y^{k+1} = \frac{A_k y^k + \alpha_{k+1} z^{k+1}}{A_{k+1}}.$$

Параметр a в clipped-SSTM используется для уменьшения размера шага алгоритма, что позволяет избежать необходимости использовать большие батчи. Что касается основной идеи доказательства, она аналогична неускоренному случаю: с помощью индукции доказывается, что с высокой вероятностью выполнено неравенство $R_N \leqslant CR_0$. Так же стоит отметить, что в ускоренном случае уровень клиппинга λ_{k+1} уменьшающийся. Из-за чувствительности ускоренных методов к неточностям, включая стохастичность, использование постоянного уровня клиппинга не является достаточным, так как это может привести к значительному смещению. Для метода подобных треугольников можно доказать, что $\|\nabla f(x^{k+1})\| = \mathbb{O}(1/\alpha_{k+1})$. По аналогии с детерминированным случаем выбирается $\lambda_{k+1} \sim 1/\alpha_{k+1}$, и удается доказать, что условие $\|\nabla f(x^{k+1})\| = \mathbb{O}(1/\alpha_{k+1})$ выполняется с высокой вероятностью и в стохастическом случае.

Сходимость clipped-SSTM с большой вероятностью

Теорема 4.2. *Пусть* f *выпуклая и* L*-гладкая на* $B_{3R_0}(x^*) = \{x \in \mathbb{R}^n \,|\, \|x - x^*\| \leqslant 3R_0\}$ *и* (8) *выполняется на* $B_{3R_0}(x^*)$. *Тогда для любых* $\beta \in (0,1)$, $\varepsilon \geqslant 0$ *таких, что* $\ln(\sqrt{L}R_0/\sqrt{\varepsilon\beta}) \geqslant 2$ *существует такой выбор* a, *что* clipped-SSTM *с уровнем клиппинга* $\lambda \sim 1/\alpha_{k+1}$ *достигает* y^N, *удовлетворяющую* $f(y^N) -$

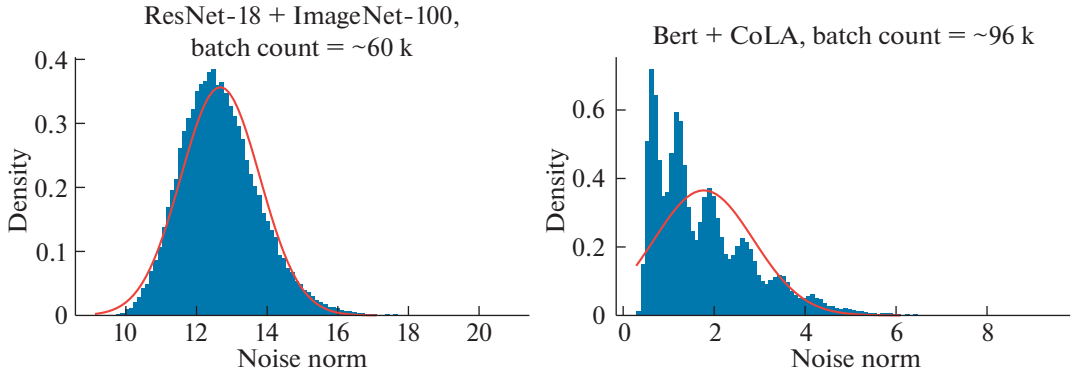

Рис. 2. Распределение шума в стохастических градиентах для `ResNet-18` на датасете `ImageNet-100` и донастройки `BERT` на датасете `CoLA` перед обучением. Красные линии представляют функции плотности вероятности нормальных распределений со средними значениями и дисперсиями, оцененными эмпирически на основе выборок. Batch count – это общее число выборок, использованных для построения гистограммы.

$- f(x^*) \leqslant \varepsilon$ *с вероятностью не менее* $1-\beta$, *используя*

$$\mathcal{O}\left(\max\left\{\sqrt{\frac{LR_0^2}{\varepsilon}}\ln\frac{LR_0^2}{\varepsilon\beta}, \left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\ln\left(\frac{1}{\beta}\left(\frac{\sigma R_0}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right)\right\}\right)$$

итераций/вызовов оракула.

Полученный результат соответствует (с точностью до логарифмических множителей) результатам для `AC-SA` в случае легких хвостов и лучше чем результаты для `clipped-SGD`. Так же в работах [Gorbunov et al., 2020, 2021, 2022, Sadiev et al., 2023] получены результаты для невыпуклых, сильно выпуклых функций и для функций с непрерывными по Гёльдеру градиентами. Как и в случае `clipped-SGD`, все предположения достаточно требовать на шаре вокруг решения.

Численные эксперименты

На практике работоспособность методов была проверена на следующих задачах[1]:

BERT ($\approx 0.6M$ параметров) донастройка на датасете `CoLA`. Мы используем предварительно обученную модель `BERT` и замораживаем все слои, кроме двух последних линейных. В этом наборе данных содержится 8551 предложений, и задача бинарной классификации заключается в определении, является ли предложение грамматически верным.

ResNet-18 ($\approx 11.7M$ параметров) обучение на датасете `ImageNet-100` (первые 100 классов из `ImageNet`). В нем содержится 134395 изображений.

Один из экспериментов был проведен с целью анализа распределения шума в стохастических градиентах. [Gorbunov et al., 2021]. В начальной точке бралось достаточно большое количество пробатченных стохастических градиентов $\nabla f(x^0,\xi_1),\ldots,\nabla f(x^0,\xi_K)$ с размером батча 32 и строились гистограммы для $\|\nabla f(x^0,\xi_1)-\nabla f(x^0)\|_2$, ..., $\|\nabla f(x^0,\xi_K)-\nabla f(x^0)\|_2$, см. рис. 2. Как видно, распределение шума для `BERT + CoLA` значительно отличается от субгауссовского, в то время как распределение для `ResNet-18 + Imagenet-100` почти гауссовское. Наблюдается аналогичное распределения шума в разных точках на траектории методов (`ADAM`, `SGD`, `clipped-SGD`, `clipped-SSTM`) для предложенных задач.

Далее было проведено сравнительное исследование различных методов для решения данных задач: `ADAM`, `SGD`, `clipped-SGD` и `clipped-SSTM` [Gorbunov et al., 2021]. Иллюстрация результатов приведена на рис. 3, 4.

5. ОСНОВНОЙ РЕЗУЛЬТАТ ДЛЯ ВАРИАЦИОННЫХ НЕРАВЕНСТВ

В данной серии исследований также освещается анализ методов для работы с безусловными вариационными неравенствами (VIP), представляющими собой нелинейные уравнения [Harker and Pang, 1990, Ryu and Yin, 2022]. Эти уравнения, в свою очередь, возникают в контексте игровых формулировок задач машинного обучения [Goodfellow et al., 2014, Gidel et al., 2019].

$$\text{Ищем } x^* \in \mathbb{R}^n \text{ такой, что } F(x^*) = 0, \qquad (17)$$

где $F(x) = \mathbb{E}_{\xi\sim\mathcal{D}}[F_\xi(x)]$. Аналогично с задачами минимизации сохраняется предположение об ограниченности центрального α-го момента операто-

[1] Код доступен по ссылке https://github.com/ClippedStochasticMethods/clipped-SSTM

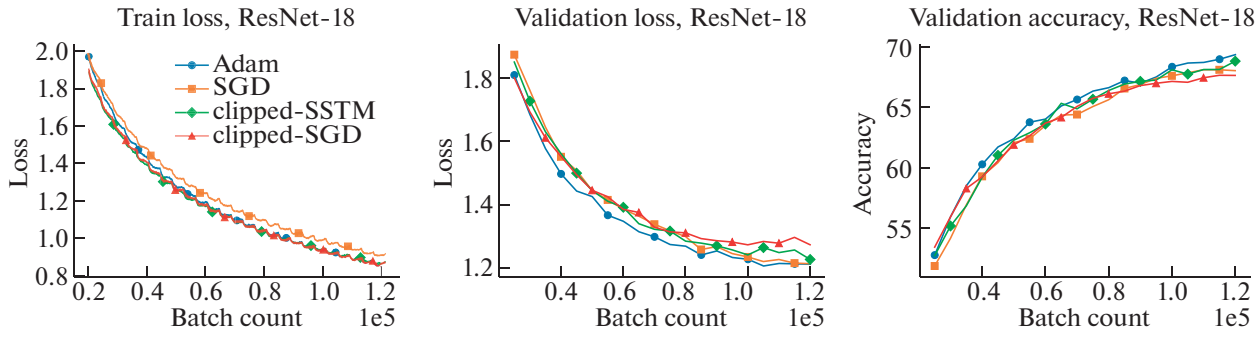

Рис. 3. Показатели потерь (loss) и точности (accuracy) на обучающей и валидационной выборках для различных методов в задаче `ResNet-18 + ImageNet-100`. Здесь "batch count" обозначает общее количество использованных стохастических градиентов. Распределение шума практически гауссовское, поэтому даже при использовании обычного стохастического градиентного спуска (`vanilla SGD`) модель хорошо обучается.

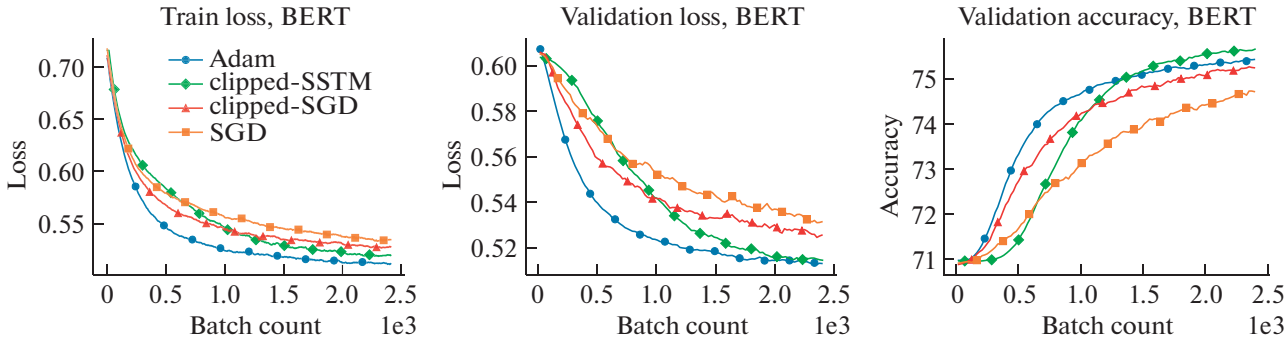

Рис. 4. Показатели потерь (loss) и точности (accuracy) на обучающей и валидационной выборках для различных оптимизаторов в задаче `BERT + CoLA`. Распределение шума имеет тяжелые хвосты, и методы с клиппингом значительно превосходят обычный стохастический градиентный спуск `SGD` по результатам.

ра F. Предполагается, что существует некоторое множество $Q \subseteq \mathbb{R}^n$ и значения $\sigma \geqslant 0$, $\alpha \in (1,2]$, такие что для всех $x \in Q$

$$\mathbb{E}_{\xi\sim\mathcal{D}}[\|F_\xi(x) - F(x)\|^\alpha] \leqslant \sigma^\alpha. \tag{18}$$

`clipped-SEG` — клиппированный стохастический экстраградиентный метод

$$\tilde{x}^k = x^k - \gamma \cdot \mathtt{clip}(F_{\xi_1^k}(x^k), \lambda_k), \tag{19}$$

$$x^{k+1} = x^k - \gamma \cdot \mathtt{clip}(F_{\xi_2^k}(\tilde{x}^k), \lambda_k), \tag{20}$$

где ξ_1^k, ξ_2^k независимо выбираются из $\mathcal{D}_k$ на каждом шаге. Данный метод рассматривается для решения вариационных неравенств (VIP) в предположении о монотонности и липшицевости оператора F. Предполагается, что существует некоторое множество $Q \subseteq \mathbb{R}^n$, для которого оператор F монотонен на Q, т.е. для всех $x, y \in Q$

$$\langle F(x) - F(y), x - y\rangle \geqslant 0, \tag{21}$$

и существует константа $L > 0$, такая что для всех $x, y \in Q$

$$\|F(x) - F(y)\| \leqslant L\|x - y\|. \tag{22}$$

Теорема 5.1 (Сходимость `clipped-SEG`). *Пусть предположения* (18), (21), (22) *выполняются для* $Q = B_{4R}(x^*)$ *и* $0 < \gamma = \mathbb{O}(\min\{1/LA, R/K^{1/\alpha}\sigma A^{(\alpha-1)/\alpha}\})$, $\lambda_k = \lambda = \Theta(R/\gamma A)$, *где* $A = \ln\dfrac{6(K+1)}{\beta} \geqslant 1$, $\beta \in (0,1]$. *Чтобы гарантировать* $Gap_R(\tilde{x}_{\text{avg}}^K) \leqslant \varepsilon$ *с вероятностью* $\geqslant 1-\beta$ `clipped-SEG`, *требуется*

$$\tilde{\mathbb{O}}\left(\max\left\{\frac{LR^2}{\varepsilon}, \left(\frac{\sigma R}{\varepsilon}\right)^{\frac{\alpha}{\alpha-1}}\right\}\right) \tag{23}$$

вызовов оракула.

Основная идея доказательства вышеупомянутой теоремы аналогична принципу в минимизации. Детальное описание результатов для разнообразных случаев можно найти в статьях [Gorbunov et al., 2022, Sadiev et al., 2023]. Также следует

обратить внимание на статью [Gorbunov et al., 2022], где демонстрируется, что градиентный шум во многих генеративно-состязательных сетях (GAN) имеет тяжелые хвосты, и что использование клиппинга положительно сказывается на эффективности SEG.

6. ЗАКЛЮЧЕНИЕ

В данном обзоре демонстрируется и объясняется роль клиппинга в современных state-of-the-art результатах о сходимости с большой вероятностью. Мы рассмотрели важные аспекты задач, связанных с присутствием шума с тяжелыми хвостами в различных областях, включая обработку естественного языка (NLP) и генеративно-состязательные сети (GAN). Клиппинг выделяется в качестве эффективного и простого инструмента для борьбы с шумом с тяжелыми хвостами. Этот подход подтвердил свою эффективность в улучшении оценок вероятностной сходимости методов, в сравнении с гарантиями сходимости альтернативных методов, не использующих клиппинг. Также интересными являются предположения о частичных объяснениях успеха адаптивных методов, включая, например, метод ADAM, в задачах GAN и NLP. Это может предоставить дополнительные идеи для разработки алгоритмов в этих областях. Исходя из результатов и выводов, представленных в данной работе, можно заключить, что использование методов с клиппингом представляет собой перспективный подход для улучшения сходимости в условиях шума с тяжелыми хвостами в различных задачах, и дальнейшие исследования в этом направлении могут привести к еще более значимым результатам и новым практическим применениям.

ИСТОЧНИК ФИНАНСИРОВАНИЯ

Работа выполнена при поддержке Аналитического центра при Правительстве Российской Федерации в соответствии с договором о субсидии (идентификатор договора 000000D730321P5Q0002; грант і 70-2021-00138).

СПИСОК ЛИТЕРАТУРЫ

1. *Bennett G.* Probability inequalities for the sum of independent random variables. Journal of the American Statistical Association. 1962. V. 57 (297). P. 33–45.
2. *Cutkosky A., Mehta H.* High-probability bounds for non-convex stochastic optimization with heavy tails. Advances in Neural Information Processing Systems. 2021. V. 34.
3. *Davis D., Drusvyatskiy D., Xiao L., Zhang J.* From low probability to high confidence in stochastic convex optimization. Journal of Machine Learning Research. 2021. V. 22 (49). P. 1–38.
4. *Devolder O. et al.* Stochastic first order methods in smooth convex optimization. Technical report, CORE, 2011.
5. *Dzhaparidze K., Van Zanten J.* On bernstein-type inequalities for martingales. Stochastic processes and their applications. 2001. V. 93 (1). P. 109–117.
6. *Freedman D.A. et al.* On tail probabilities for martingales. the Annals of Probability. 1975. V. 3 (1). P. 100–118.
7. *Gasnikov A., Nesterov Y.* Universal fast gradient method for stochastic composit optimization problems. arXiv preprint arXiv:1604.05275, 2016.
8. *Ghadimi S., Lan G.* Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. SIAM Journal on Optimization. 2012. V. 22 (2). P. 1469–1492.
9. *Ghadimi S., Lan G.* Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization. 2013. V. 23 (2). P. 2341–2368.
10. *Gidel G., Berard H., Vignoud G., Vincent P., Lacoste-Julien S.* A variational inequality perspective on generative adversarial networks. International Conference on Learning Representations, 2019.
11. *Goodfellow I., Pouget-Abadie J., Mirza M., Xu B., Warde-Farley D., Ozair S., Courville A., Bengio Y.* Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, K.Q. Weinberger, editors, Advances in Neural Information Processing Systems, volume 27. Curran Associates, Inc., 2014.
12. *Gorbunov E., Danilova M., Gasnikov A.* Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. Advances in Neural Information Processing Systems. 2020. V. 33. P. 15042–15053.
13. *Gorbunov E., Danilova M., Shibaev I., Dvurechensky P., Gasnikov A.* Near-optimal high probability complexity bounds for non-smooth stochastic optimization with heavy-tailed noise. arXiv preprint arXiv:2106.05958, 2021.
14. *Gorbunov E., Danilova M., Dobre D., Dvurechenskii P., Gasnikov A., Gidel G.* Clipped stochastic methods for variational inequalities with heavy-tailed noise. Advances in Neural Information Processing Systems. 2022. V. 35. P. 31319–31332.
15. *Harker P.T., Pang J.-S.* Finite-dimensional variational inequality and nonlinear complementarity problems: a survey of theory, algorithms and applications. Mathematical programming. 1990. V. 48 (1-3). P. 161–220.
16. Kingma D.P., Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.
17. *Liu Z., Zhou Z.* Stochastic nonsmooth convex optimization with heavy-tailed noises. arXiv preprint arXiv:2303.12277, 2023.
18. *Liu Z., Nguyen T.D., Nguyen T.H., Ene A., Nguyen H.* High probability convergence of stochastic gradient methods. In International Conference on Machine Learning, PMLR, 2023. P. 21884–21914.
19. *Merity S., Keskar N.S., Socher R.* Regularizing and optimizing lstm language models. In International Conference on Learning Representations, 2018.

20. *Mosbach M., Andriushchenko M., Klakow D.* On the stability of fine-tuning bert: Misconceptions, explanations, and strong baselines. In International Conference on Learning Representations, 2020.
21. *Nazin V., Nemirovsky A., Tsybakov A.B., Juditsky A.* Algorithms of robust stochastic optimization based on mirror descent method. Automation and Remote Control. 2019. V. 80 (7). P. 1607–1627.
22. *Nguyen T.D., Nguyen T.H., Ene A., Nguyen H.L.* High probability convergence of clipped-sgd under heavy-tailed noise. arXiv preprint arXiv:2302.05437, 2023.
23. *Pascanu R., Mikolov T., Bengio Y.* On the difficulty of training recurrent neural networks. In International conference on machine learning, Pmlr, 2013. P. 1310–1318.
24. *Peters M.E., Ammar W., Bhagavatula C., Power R.* Semi-supervised sequence tagging with bidirectional language models. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017. P. 1756–1765.
25. *Ryu E.K., Yin W.* Large-scale convex optimization: algorithms & analyses via monotone operators. Cambridge University Press, 2022.
26. *Sadiev A., Danilova M., Gorbunov E., Horv'ath S., Gidel G., Dvurechensky P., Gasnikov A., Richt'arik P.* High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. In Proceedings of the 40th International Conference on Machine Learning. PMLR, 2023. P. 29563–29648.
27. *Zhang J., Karimireddy S.P., Veit A., Kim S., Reddi S., Kumar S., Sra S.* Why are adaptive methods good for attention models? Advances in Neural Information Processing Systems. 2020. V. 33. P. 15383–15393.

ALGORITHMS WITH GRADIENT CLIPPING FOR STOCHASTIC OPTIMIZATION WITH HEAVY-TAILED NOISE

M. Danilova[a]

[a] *Moscow Institute of Physics and Technology, Moscow, Russia*

Presented by Academician of the RAS A.A. Shananin

This article provides a review of the results of several research studies, in which open questions related to the high-probability convergence analysis of stochastic first-order optimization methods under mild assumptions on the noise were gradually addressed. In the beginning, we introduce the concept of gradient clipping, which plays a pivotal role in the development of stochastic methods for successful operation in the case of heavy-tailed distributions. Next, we examine the importance of obtaining the highprobability convergence guarantees and their connection with in-expectation convergence guarantees. The concluding sections of the article are dedicated to presenting the primary findings related to minimization problems and the results of numerical experiments.

Keywords: Convex optimization, stochastic optimization, first-order methods