OPT2024: 16th Annual Workshop on Optimization for Machine Learning

On Convergence of DP-SGD with Adaptive Clipping

Egor Shulgin SHULGIN.YEGOR@GMAIL.COM
Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

Stochastic Gradient Descent (SGD) with gradient clipping has emerged as a powerful technique for stabilizing neural network training and enabling differentially private optimization. While constant clipping has been extensively studied, adaptive methods like quantile clipping have shown empirical success without thorough theoretical understanding. This paper provides the first comprehensive convergence analysis of SGD with gradient quantile clipping (QC-SGD). We demonstrate that QC-SGD suffers from a bias problem similar to constant-threshold clipped SGD, but show this can be mitigated through a carefully designed quantile and step size schedule. Furthermore, the analysis is extended to the differentially private case. We establish theoretical foundations for this widely-used heuristic and identify open problems to guide future research.

Keywords: gradient clipping, differentially private optimization

1. Introduction

It is hard to imagine the success of modern Machine Learning without effective optimization, the cornerstone of which are Stochastic Gradient Descent (SGD) type methods [4, 23]. However, SGD is not perfect, particularly in the context of Deep Learning. Efficient neural network training often requires modifications of SGD to stabilize optimization. For example, exploding gradients issue [20, 21] is often tackled by the use of *gradient clipping* operator, which scales down the input vector's norm if it exceeds a certain threshold. Moreover, gradient clipping plays a vital role in privacy-preserving machine learning [7, 8]. Rigorous differential privacy guarantees are usually established by relying on the Gaussian mechanism [8], which requires bounded per-example sensitivity to control the amount of noise added. The most commonly used in practice Differentially Private SGD (DP-SGD) [1] method enforces such bound by clipping the gradients.

Clipped SGD was shown to be superior to vanilla SGD for minimizing generalized smooth functions [26] and when stochastic gradient noise is heavy-tailed [27]. However, the effectiveness of gradient clipping hinges critically on the choice of the clipping threshold, denoted as τ. This introduces an additional hyperparameter that requires careful tuning, a challenge that is especially pronounced in private optimization settings where performance can be highly sensitive to this threshold [5, 14]. Furthermore, each training run incurs an additional privacy loss, making extensive hyperparameter search prohibitively expensive from a privacy perspective [19].

Adaptive clipping. The problem described above has been addressed [2] in the setting of private Federated Learning [11, 13, 15, 16]. Specifically, Andrew et al. [2] proposed to adaptively select the clipping threshold based on the distribution of gradient norms (or updates) of the participating clients. Their method efficiently estimates a quantile and applies it as the clipping threshold. Crucially, their privacy analysis revealed that this adaptive approach incurs only negligible additional privacy loss. Extensive empirical evaluations have shown [2] that quantile clipping is competitive with, and often outperforms, carefully tuned constant clipping baselines. This adaptive strategy offers the additional benefit of adjusting to the evolving gradient distribution throughout the federated optimization process. This adaptability is particularly valuable given the significant variability observed across different machine learning tasks and datasets (see Figure 1), suggesting that a uniform clipping schedule may be suboptimal. The success of the adaptive clipping technique has led to its widespread adoption for multiple applications [24, 25], even beyond privacy-constrained settings [6], and implementation in Federated Learning libraries [3, 10].

However, despite its practical success and adoption, the theoretical properties of stochastic gradient quantile clipping remain largely unexplored. This research gap motivates our current study. We aim to provide a comprehensive optimization analysis of SGD with quantile clipping. In particular, by building upon recent work [17], we demonstrate that SGD with quantile clipping (QC-SGD in short) suffers from a bias problem, preventing convergence, similar to that observed in constant clipped SGD. We design a quantile and step size schedule that effectively eliminates the identified bias problem. Our analysis reveals the crucial relationship between the chosen quantile value and the step size in QC-SGD, providing insights into how this interplay affects convergence.

2. Problem and Assumptions

We consider a classical stochastic optimization problem

$$\min_{x\in\mathbb{R}^d}\Big[f(x) \coloneqq \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)]\Big], \tag{1}$$

where $f_\xi : \mathbb{R}^d \to \mathbb{R}$ is a loss of machine learning model parametrized by vector $x \in \mathbb{R}^d$ on data point ξ. Thus, f is excess loss over the distribution of data $\xi \sim \mathcal{D}$. To enable optimization analysis, we rely on standard assumptions for non-convex stochastic optimization.

Assumption 1 (Stochastic gradient) *Stochastic gradient estimator is unbiased* $\mathbb{E}\left[\nabla f_\xi(x)\right] = \nabla f(x)$ *and has* ***bounded*** q***-th moment*** *for* $q \in (1, 2]$

$$\left(\mathbb{E}_{\xi\sim\mathcal{D}}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^q\right]\right)^{1/q} \le \sigma_q, \qquad \forall x \in \mathbb{R}^d. \tag{2}$$

Condition (2) is usually referred as heavy tailed noise [27] for $q \in (1, 2)$. For $q = 2$, it recovers classical bounded variance assumption [9, 18].

Assumption 2 (Function) *The function* f *is differentiable and* L*-**smooth**, meaning there exists* $L > 0$ *such that*

$$f(x + h) \le f(x) + \langle\nabla f(x), h\rangle + \frac{L}{2}\|h\|^2, \qquad \forall x, h \in \mathbb{R}^d. \tag{3}$$

Additionally, f *is lower-bounded by* $f^{\inf} \in \mathbb{R}$.

3. Stochastic Gradient Descent with Quantile Clipping

We consider the following Stochastic Gradient Descent type method with step size $\gamma_t > 0$

$$x^{t+1} = x^t - \gamma_t g^t, \tag{4}$$

where $g^t = g(x^t)$ has the form of a clipped stochastic gradient estimator $g^t = \alpha_{\xi^t}(x^t)\nabla f_{\xi^t}(x^t)$ for

$$\alpha_\xi(x) = \min\left\{1, \frac{\tau(x)}{\|\nabla f_\xi(x)\|}\right\}, \tag{5}$$

and $\tau(x)$ is p-th quantile (of random variable $\|\nabla f_\xi(x)\|$) clipping threshold, defined as

$$\mathrm{Prob}(\|\nabla f_\xi(x)\| \le \tau(x)) = p. \tag{6}$$

We use the name SGD with Quantile Clipping (QC-SGD) for the described algorithm. Note that Clipped SGD is a special case of QC-SGD for $\tau(x) \equiv \tau$. This algorithm was originally introduced in the seminal work by Merad and Gaïffas [17] in the context of robust optimization with a corrupted oracle. They analyzed it as a Markov chain, while we are interested in optimization properties with a focus on differentially private settings.

3.1. Preliminaries

At first, we present some crucial properties of the described gradient estimator $g(x^t)$ and clipping threshold $\tau(x)$ needed for convergence analysis.

Lemma 1 (Merad and Gaïffas [17]) *Assume that stochastic gradient estimator $\nabla f_\xi(x)$ satisfies Assumption 1, $\alpha_\xi(x)$ is chosen as (5), and p-th quantile clipping threshold $\tau(x)$ satisfies (6). Then for all $x \in \mathbb{R}^d$,*

$$\tau(x) \le \|\nabla f(x)\| + \sigma_q (1-p)^{-1/q}, \tag{7}$$

$$\|\mathbb{E}\left[\alpha_\xi(x)\nabla f_\xi(x)\right] - \overline{\alpha}(x)\nabla f(x)\| \le \sigma_q (1-p)^{1-1/q}, \tag{8}$$

where $\overline{\alpha}(x) := \mathbb{E}\left[\alpha_\xi(x)\right]$.

Note that (8) is different from typical results [12, 27] characterizing the bias of gradient clipping due to the additional multiplier $\overline{\alpha}(x)$ before the gradient $\nabla f(x)$. This has a significant impact on the convergence analysis of QC-SGD and makes it different from Clipped SGD.

3.2. Convergence analysis

Our analysis relies on the following recursion.

Lemma 2 *Suppose f is L-smooth (2) and stochastic gradients satisfy Assumption 1. Then, for $\beta > 0$ the iterates of QC-SGD (4) satisfy*

$$\begin{aligned}\mathbb{E}\left[f(x^{t+1}) \mid x^t\right] \le f(x^t) - \gamma_t\left(\overline{\alpha}(x^t) - \beta/2 - \gamma_t L\right)\left\|\nabla f(x^t)\right\|^2 \\ + \frac{\gamma_t}{2}\beta^{-1}(1-p)^{2-2/q}\sigma_q^2 + \gamma_t^2 L\sigma_q^2(1-p)^{-2/q}\end{aligned} \tag{9}$$

Equipped with Lemma 2, we provide a general convergence result for QC-SGD.

Theorem 3 (General case) *Suppose f is L-smooth (2) and stochastic gradients satisfy Assumption 1. Then, for the step size chosen as*

$$0 < \gamma_t \le \frac{2p - \beta - c}{2L}, \tag{10}$$

where $c, \beta \in (0,1)$ and $\beta + c \leq 2p$ the iterates of QC-SGD (4) *satisfy*

$$\frac{c}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq 2\frac{f(x^0) - \mathbb{E}\left[f(x^T)\right]}{\Gamma_T} + \sigma_q^2 \frac{1}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t h^{-2/q}\left(2L\gamma_t + \beta^{-1}h^2\right), \quad (11)$$

where $\Gamma_T := \sum_{t=0}^{T-1}\gamma_t$ and $h := 1 - p$.

Theorem 3 indicates that QC-SGD can find an approximate stationary point. Importantly, more aggressive clipping (smaller p) requires decreasing the step size γ_t according to (10). However, overall performance heavily depends on h and the sequence γ_t. Next, we discuss how 2 possible choices affect convergence.

Corollary 4 (Constant parameters) *For constant step size $\gamma_t \equiv \gamma \leq (2p - \beta - c)/(2L)$ convergence bound* (11) *reduces to*

$$\frac{c}{T}\sum_{t=0}^{T-1}\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{2\left(f(x^0) - \mathbb{E}\left[f(x^T)\right]\right)}{\gamma T} + 2\gamma L\sigma_q^2 h^{-2/q} + \beta^{-1}\sigma_q^2 h^{2-2/q}. \quad (12)$$

Obtained result shares similarities with classical SGD (with bounded stochastic gradient variance) convergence guarantees. Namely, the first term in upper bound (12) is basically the same and decreases with rate $\mathcal{O}(1/T)$. The second term is larger by factor $h^{-2/q}$ as $h = 1 - p \in [0,1)$ indicating that more aggressive clipping (smaller p) increases the neighborhood. However, the fundamental difference to SGD is that exact convergence to the stationary point (making gradient arbitrary small: $\mathbb{E}\left\|\nabla f(x^t)\right\|^2 \leq \varepsilon^2$) can not be guaranteed for any constant step size γ as the third term in (12) can not be controlled. This means that the method has an irreducible bias, unlike standard SGD, which enjoys convergence rate $\mathcal{O}(1/\sqrt{T})$ by choosing step size γ as $\mathcal{O}(1/\sqrt{T})$.

In addition, our analysis indicates a trade-off between convergence speed and the size of the neighborhood. Specifically, larger β decreases the latter (third term in (12)) but requires choosing a smaller step size[1] $\gamma \leq \mathcal{O}(p - \beta)$ leading to slower convergence due to the first term in (12) inversely proportional to γ.

Time-varying parameters. By using Theorem 3, we can also jointly design schedules of step size γ_t and quantile values $p_t = 1 - h_t$ to guarantee convergence to the stationary point. Namely, for $\gamma_t = \mathcal{O}(t^{\theta-1})$ and $h_t = \mathcal{O}(t^{\nu})$ we have $\Gamma_T = \mathcal{O}(T^{\theta})$ and the upper bound (11) will be of order[2]

$$\mathcal{O}\left(T^{-\theta} + T^{\theta-1-2\nu/q} + T^{2\nu(1-1/q)}\right), \quad (13)$$

which is minimized for $\theta = (1 - q^{-1})/(2 - q^{-1})$ and $\nu = -q/2$. Thus for standard bounded variance case $q = 2$ the step size has to be decreased as $\gamma_t = \mathcal{O}(t^{-2/3})$ and quantile increased as $p_t = 1 - \mathcal{O}(t^{-1})$ to obtain convergence of order $\mathcal{O}(T^{-1/3})$. This confirms the intuition that the method can converge exactly if clipping bias is eventually eliminated. However, our result does not necessarily require decreasing the clipping threshold as norms of stochastic gradients $\left\|\nabla f_\xi(x^t)\right\|$ may not converge to zero for increasing t.

1. And potentially a larger p to ensure positive step size $\gamma > 0$.
2. We use $\mathcal{O}$ notation to suppress constants other than t, T.

3.3. Comparison to fixed clipping

The latest analysis on SGD with constant clipping ($\tau(x) \equiv \tau$) we are aware of is due to Koloskova et al. [12]. Their (simplified) result indicates that with a proper step size choice, for $q = 2, \sigma_q = \sigma$, and for L-smooth function, the expected squared gradient norm is upper bounded by

$$\mathcal{O}\left(\left(\frac{F^0}{\gamma T \tau}\right)^2 + \frac{F^0}{\gamma T} + \gamma L \sigma^2 + \min\left(\sigma^2, \frac{\sigma^4}{\tau^2}\right)\right), \tag{14}$$

where $F^0 := f(x^0) - \mathbb{E}\left[f(x^T)\right]$. This result is fundamentally similar to quantile clipping (12) as the last term is also irreducible via decreasing step size γ. However, upper bound (14) can be made arbitrary small by choosing step size as $\gamma = \mathcal{O}(T^{-1/2})$ and clipping threshold as $\tau = \mathcal{O}(T^{\lambda}), \lambda \in (0, 1/4)$. While the approach of increasing τ can solve the problem in theory, from a practical perspective, it is not satisfying. As shown by Andrew et al. [2], the evolution of the distribution of the norm of the updates (or pseudo-gradients [22]) may show very different behavior in federated training. Figure 1 shows that norms of the updates may, in fact, increase during optimization. In addition, for differentially private settings bigger τ requires adding larger noise at every iteration resulting in degraded utility of the model [5].

3.4. Bias due to clipping

In the discussion after 4, we mentioned that our result indicates that for any non-trivial fixed quantile $p \in (0, 1)$, exact convergence to the stationary point can not be guaranteed for any step size γ. In order to demonstrate that this effect is not just a result of our (potentially suboptimal) analysis but that the method's estimator is indeed limited, we present the following function (based on [12]).

Example 1 *For $r > 0$ and $\omega \in (1/2, 1)$ define*

$$f_\xi(x) = \frac{1}{2}\begin{cases} (x+r)^2, & \text{with probability } \omega \\ x^2, & \text{with probability } 1-\omega. \end{cases} \tag{15}$$

Then $\nabla f(x) = \mathbb{E}\left[\nabla f_\xi(x)\right] = x + r\omega$, which brings minima for $f(x) = \mathbb{E}\left[f_\xi(x)\right]$ at $x^\star = -r\omega$.

Suppose quantile p is chosen in such a way that half of the stochastic gradients are clipped at every point x (e.g., as the median). Then estimator has the form

$$g(x) := \alpha_\xi(x)\nabla f_\xi(x) = \begin{cases} 1, & \text{with probability } \omega \\ x, & \text{with probability } 1-\omega, \end{cases} \tag{16}$$

which indicates that $x^\dagger = -\omega/(1-\omega)$ is the expected fixed point of QC-SGD as $\mathbb{E}\left[g(x^\dagger)\right] = 0$. Thus, if QC-SGD converges, it must do so towards its fixed points. However, for any $r \neq 1/(1-\omega)$ minimum of f is different from the expected fixed point $x^\star \neq x^\dagger$ and $\|\nabla f(x^\dagger)\| > 0$.

4. Differentially Private case

The most standard way to make clipped SGD (ϵ, δ)-Differentially Private (DP) is by adding isotropic Gaussian noise with variance proportional to the clipping threshold [1] (along with subsampling/mini-batching). This approach applied to QC-SGD results in the following update (DP-QC-SGD for short)

$$x^{t+1} = x^t - \gamma_t \frac{1}{B}\sum_{j=1}^{B}\left(g_j^t + z^t\right), \qquad g_j^t = \min\left\{1, \frac{\tau(x^t)}{\left\|\nabla f_{\xi_j^t}(x^t)\right\|}\right\}\nabla f_{\xi_j^t}(x^t), \tag{17}$$

where $z^t \sim \mathcal{N}\left(0, \left(\tau(x^t)\right)^2 \sigma_{\rm DP}^2 \mathbf{I}\right)$ and $\sigma_{\rm DP} \geq C\sqrt{T\log(1/\delta)}\epsilon^{-1}$ for some universal constant C independent of T, δ, ϵ [1]. For simplicity, we assume that ξ_j^t are uniformly and independently sampled. Next, we present our convergence result for (17).

Theorem 5 (DP-QC-SGD) *Suppose f is L-smooth (2) and stochastic gradients satisfy Assumption 1. Then, for the step size chosen as*

$$\gamma_t \leq \frac{p - \beta/2 - c}{2\mathfrak{S}L}, \tag{18}$$

where $\mathfrak{S} := 1/B + \sigma_{\rm DP}^2$, and $c, \beta \in (0,1)$, and $\beta/2 + c \leq p$ the iterates of DP-QC-SGD (17) *satisfy*

$$\frac{c}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{f(x^0) - \mathbb{E}\left[f(x^T)\right]}{\Gamma_T} + \frac{\sigma_q^2}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t h^{-2/q}\left(2\gamma_t L\mathfrak{S} + \beta^{-1}h^2/2\right), \tag{19}$$

where $\Gamma_T := \sum_{t=0}^{T-1}\gamma_t$ and $h := 1 - p$.

Theorem 5 is similar to non-private result (11) in nature (up to numerical constants) as it shows convergence to a neighborhood of the stationary point. However, there is an important difference expressed in term $\mathfrak{S} = 1/B + \sigma_{\rm DP}^2$ in the denominator of the step size condition (18) and convergence bound (19). Note that $\mathfrak{S}$ can be even smaller than 1 in private federated learning for a big enough cohort size B and a small number of communication rounds T. However, for centralized DP training, $\mathfrak{S}$ is likely to be larger, which results in a smaller step size and larger convergence neighborhood. The latter, though, can be eliminated via a standard SGD step size strategy as the term in (19) involving $\mathfrak{S}$ depends on γ_t^2.

5. Conclusion and Future Directions

We provided the first non-convex convergence results for SGD with (adaptive) quantile clipping, focusing on smooth stochastic optimization under heavy-tailed noise. Our results demonstrate limitations of QC-SGD similar to standard clipped SGD, which can be addressed via a specially designed quantile and step size schedule. Finally, we analyzed a differentially private extension of QC-SGD.

The discovered limitations of the analyzed method raise the question of possible improvements via algorithmic modifications. It is also worth noting that the current analysis is performed for an idealized case when the exact quantile $\tau(x)$ is available. This may not be feasible in certain practical scenarios that only allow access to an approximation. Moreover, despite the great empirical success of adaptive clipping, there are scenarios where it performs suboptimally [25], motivating future research. We hope that our work can serve as a first step towards rigorously understanding this practical technique and eventually will help to improve private learning.

Acknowledgments and Disclosure of Funding

We would like to thank anonymous reviewers for their valuable comments, which helped improve the manuscript, and Abdurakhmon Sadiev for useful technical discussions.

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science.

References

[1] Martin Abadi, Andy Chu, Ian Goodfellow, H Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC conference on computer and communications security*, pages 308–318, 2016. (Cited on page 1, 5, and 6)

[2] Galen Andrew, Om Thakkar, Brendan McMahan, and Swaroop Ramaswamy. Differentially private learning with adaptive clipping. *Advances in Neural Information Processing Systems*, 34:17455–17466, 2021. (Cited on page 1, 2, 5, and 10)

[3] The TensorFlow Federated Authors. TensorFlow Federated, December 2018. URL https://github.com/google-parfait/tensorflow-federated. (Cited on page 2)

[4] Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018. (Cited on page 1)

[5] Zhiqi Bu, Yu-Xiang Wang, Sheng Zha, and George Karypis. Automatic clipping: Differentially private deep learning made easier and stronger. *Advances in Neural Information Processing Systems*, 36, 2024. (Cited on page 1 and 5)

[6] Zachary Charles, Zachary Garrett, Zhouyuan Huo, Sergei Shmulyian, and Virginia Smith. On large-cohort training for federated learning. *Advances in Neural Information Processing Systems*, 34:20461–20475, 2021. (Cited on page 2)

[7] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3*, pages 265–284. Springer, 2006. (Cited on page 1)

[8] Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014. (Cited on page 1)

[9] Saeed Ghadimi and Guanghui Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012. (Cited on page 2)

[10] Filip Granqvist, Congzheng Song, Áine Cahill, Rogier van Dalen, Martin Pelikan, Yi Sheng Chan, Xiaojun Feng, Natarajan Krishnaswami, Vojta Jina, and Mona Chitnis. pfl-research: simulation framework for accelerating research in private federated learning. *arXiv preprint arXiv:2404.06430*, 2024. (Cited on page 2)

[11] Peter Kairouz, H. Brendan McMahan, Brendan Avent, Aurélien Bellet, Mehdi Bennis, Arjun Nitin Bhagoji, Kallista A. Bonawitz, Zachary Charles, Graham Cormode, Rachel Cummings, Rafael G. L. D'Oliveira, Hubert Eichner, Salim El Rouayheb, David Evans, Josh Gardner, Zachary Garrett, Adrià Gascón, Badih Ghazi, Phillip B. Gibbons, Marco Gruteser, Zaïd Harchaoui, Chaoyang He, Lie He, Zhouyuan Huo, Ben Hutchinson, Justin Hsu, Martin Jaggi, Tara Javidi, Gauri Joshi, Mikhail Khodak, Jakub Konečný, Aleksandra Korolova, Farinaz Koushanfar, Sanmi Koyejo, Tancrède Lepoint, Yang Liu, Prateek Mittal, Mehryar Mohri, Richard Nock, Ayfer Özgür, Rasmus Pagh, Hang Qi, Daniel Ramage, Ramesh Raskar, Mariana Raykova, Dawn Song, Weikang Song, Sebastian U. Stich, Ziteng Sun, Ananda Theertha Suresh, Florian Tramèr, Praneeth Vepakomma, Jianyu Wang, Li Xiong, Zheng Xu, Qiang Yang, Felix X. Yu, Han Yu, and Sen Zhao. Advances and open problems in federated learning. *Found. Trends Mach. Learn.*, 14(1-2):1–210, 2021. doi: 10.1561/2200000083. URL `https://doi.org/10.1561/2200000083`. (Cited on page 1)

[12] Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich. Revisiting gradient clipping: Stochastic bias and tight convergence guarantees. In *International Conference on Machine Learning*, pages 17343–17363. PMLR, 2023. (Cited on page 3 and 5)

[13] Jakub Konečný, H. Brendan McMahan, Felix X. Yu, Peter Richtárik, Ananda Theertha Suresh, and Dave Bacon. Federated learning: Strategies for improving communication efficiency. *NIPS Private Multi-Party Machine Learning Workshop*, 2016. (Cited on page 1)

[14] Alexey Kurakin, Shuang Song, Steve Chien, Roxana Geambasu, Andreas Terzis, and Abhradeep Thakurta. Toward training at imagenet scale with differential privacy. *arXiv preprint arXiv:2201.12328*, 2022. (Cited on page 1)

[15] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial Intelligence and Statistics*, pages 1273–1282. PMLR, 2017. (Cited on page 1)

[16] H Brendan McMahan, Daniel Ramage, Kunal Talwar, and Li Zhang. Learning differentially private recurrent language models. In *International Conference on Learning Representations*, 2018. (Cited on page 1)

[17] Ibrahim Merad and Stéphane Gaïffas. Robust stochastic optimization via gradient quantile clipping. *Transactions on Machine Learning Research*, 2024. URL `https://openreview.net/forum?id=HCRkV3kxHW`. (Cited on page 2 and 3)

[18] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4): 1574–1609, 2009. (Cited on page 2)

[19] Nicolas Papernot and Thomas Steinke. Hyperparameter tuning with renyi differential privacy. In *International Conference on Learning Representations*, 2021. (Cited on page 1)

[20] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. Understanding the exploding gradient problem. *arXiv preprint arXiv:1211.5063*, 2012. (Cited on page 1)

[21] Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on Machine Learning*, 2013. (Cited on page 1)

[22] Sashank J. Reddi, Zachary Charles, Manzil Zaheer, Zachary Garrett, Keith Rush, Jakub Konečný, Sanjiv Kumar, and Hugh Brendan McMahan. Adaptive federated optimization. In *International Conference on Learning Representations*, 2021. (Cited on page 5)

[23] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951. (Cited on page 1)

[24] Congzheng Song, Filip Granqvist, and Kunal Talwar. FLAIR: Federated learning annotated image repository. *Advances in Neural Information Processing Systems*, 35:37792–37805, 2022. (Cited on page 2)

[25] Zheng Xu, Yanxiang Zhang, Galen Andrew, Christopher A Choquette-Choo, Peter Kairouz, H Brendan McMahan, Jesse Rosenstock, and Yuanbo Zhang. Federated learning of gboard language models with differential privacy. *arXiv preprint arXiv:2305.18465*, 2023. (Cited on page 2 and 6)

[26] Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019. (Cited on page 1)

[27] Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33:15383–15393, 2020. (Cited on page 1, 2, and 3)

Appendix

Contents

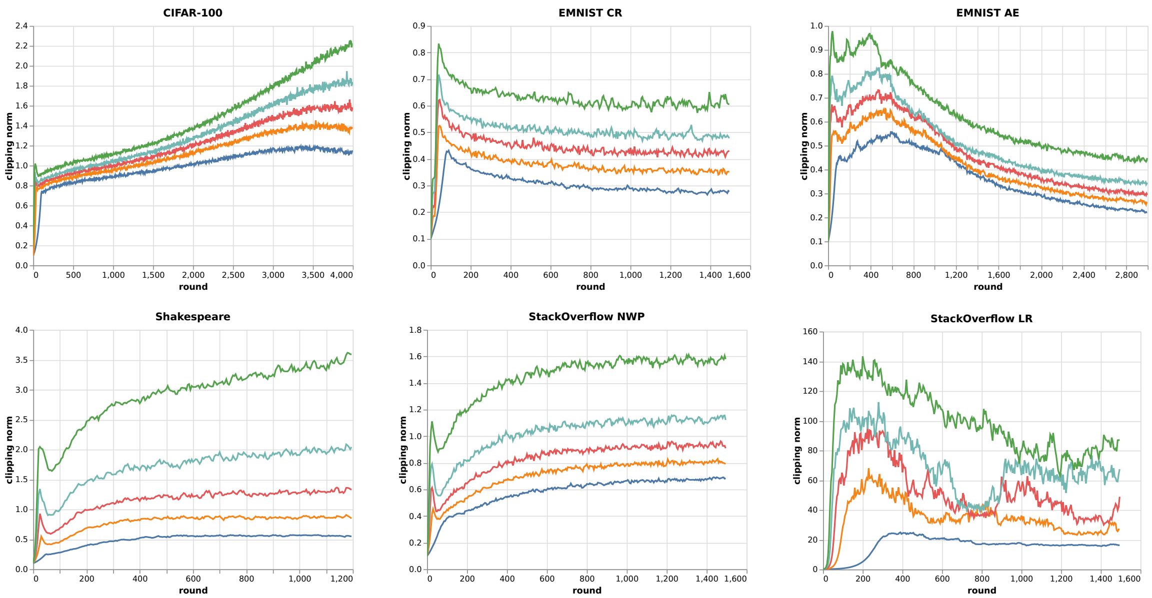

Figure 1: Evolution of the adaptive clipping norm at five different quantiles (0.1, 0.3, 0.5, 0.7, 0.9) on 6 federated learning problems without Differential Privacy noise. Note that each task has a unique shape to its update norm evolution, which further motivates an adaptive approach. Figure taken from [2].

Appendix A. Basic and Auxiliary Facts

For all vectors $a, b \in \mathbb{R}^d$ and $\beta > 0$:

$$\|a+b\| \leq \|a\| + \|b\| \,, \tag{20}$$

$$\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2 \,, \tag{21}$$

$$|\langle a, b\rangle| \leq \|a\| \, \|b\| \,, \tag{22}$$

$$2\langle a, b\rangle \leq \beta \|a\|^2 + \beta^{-1}\|b\|^2 \,. \tag{23}$$

For a set of $n \geq 1$ vectors $a_1, \ldots, a_n \in \mathbb{R}^d$ it holds

$$\left\| \frac{1}{n}\sum_{i=1}^{n} a_i \right\|^2 \leq \frac{1}{n}\sum_{i=1}^{n} \|a_i\|^2 \,. \tag{24}$$

Appendix B. Proofs

B.1. Proof of Lemma 1

We provide the proof here for completeness. Note that

$$\mathbb{I}\left\{\|\nabla f_\xi(x)\| \leq \tau(x)\right\} + \mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\} = 1. \tag{25}$$

1. As a reminder $\overline{\alpha}(x) \coloneqq \mathbb{E}\left[\alpha_\xi(x)\right]$. Then

$$\begin{aligned}
\mathbb{E}\left[\alpha_\xi(x)\nabla f_\xi(x)\right] - \overline{\alpha}(x)\nabla f(x) &= \mathbb{E}\left[\left(\alpha_\xi(x) - \overline{\alpha}(x)\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\right] \\
&= \mathbb{E}\left[\left(\alpha_\xi(x) - \overline{\alpha}(x)\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| \leq \tau(x)\right\}\right] \\
&\quad + \mathbb{E}\left[\left(\alpha_\xi(x) - \overline{\alpha}(x)\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&= \mathbb{E}\left[\left(1 - \overline{\alpha}(x)\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\left(1 - \mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right)\right] \\
&\quad - \overline{\alpha}(x)\mathbb{E}\left[\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&\quad + \mathbb{E}\left[\frac{\tau(x)}{\|\nabla f_\xi(x)\|}\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&= \mathbb{E}\left[\left(1 - \overline{\alpha}(x)\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\right] \\
&\quad + \mathbb{E}\left[\left(\overline{\alpha}(x) - 1\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&\quad - \overline{\alpha}(x)\mathbb{E}\left[\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&\quad + \mathbb{E}\left[\frac{\tau(x)}{\|\nabla f_\xi(x)\|}\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right] \\
&= \mathbb{E}\left[\left(\frac{\tau(x)}{\|\nabla f_\xi(x)\|} - 1\right)\left(\nabla f_\xi(x) - \nabla f(x)\right)\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\right]
\end{aligned}$$

Next, we use the fact that if $\frac{\tau(x)}{\|\nabla f_\xi(x)\|} \in (0,1)$ then $\left|\tau(x)/\left\|\nabla f_\xi(x)\right\| - 1\right| \in (0,1)$. Thus

$$\begin{aligned}
\left\|\mathbb{E}\left[\alpha_\xi(x)\nabla f_\xi(x)\right] - \overline{\alpha}(x)\nabla f(x)\right\| &\leq \mathbb{E}\left[\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\left|\frac{\tau(x)}{\|\nabla f_\xi(x)\|} - 1\right|\left\|\nabla f_\xi(x) - \nabla f(x)\right\|\right] \\
&\leq \mathbb{E}\left[\mathbb{I}\left\{\|\nabla f_\xi(x)\| > \tau(x)\right\}\left\|\nabla f_\xi(x) - \nabla f(x)\right\|\right] \\
&\leq (1-p)^{1-1/q}\,\mathbb{E}\left[\left\|\nabla f_\xi(x) - \nabla f(x)\right\|^q\right]^{1/q} \\
&\overset{(2)}{\leq} (1-p)^{1-1/q}\,\sigma_q.
\end{aligned}$$

2. Denote by $\mathcal{Q}_p\left(\|\nabla f_\xi(x)\|\right) = \tau(x)$ the p-th quantile of $\|\nabla f_\xi(x)\|$ distribution. Then

$$\tau(x) = \mathcal{Q}_p\left(\|\nabla f_\xi(x) - \nabla f(x) + \nabla f(x)\|\right) \overset{(20)}{\leq} \|\nabla f(x)\| + \mathcal{Q}_p\Big(\|\underbrace{\nabla f_\xi(x) - \nabla f(x)}_{\delta_x}\|\Big). \tag{26}$$

By quantile definition and by using Markov's inequality

$$1 - p = \mathrm{Prob}\left\{\|\delta_x\| > \mathcal{Q}_p(\|\delta_x\|)\right\} \leq \left(\frac{\mathbb{E}\left[\|\delta_x\|\right]}{\mathcal{Q}_p(\|\delta_x\|)}\right)^q \leq \left(\frac{\sigma_q}{\mathcal{Q}_p(\|\delta_x\|)}\right)^q, \tag{27}$$

where the last inequality holds for $q > 1$ as $\left(\mathbb{E}\left[\|\delta_x\|\right]\right)^q \leq \mathbb{E}\left[\|\delta_x\|^q\right]$. Therefore $\mathcal{Q}_p(\|\delta_x\|) \leq \frac{\sigma_q}{(1-p)^{1/q}}$

B.2. Proof of Lemma 2

By using L-smoothness (2) for iterates of algorithm (4)

$$x^{t+1} = x^t - \gamma_t g^t, \qquad g^t = \alpha_{\xi^t}(x^t)\nabla f_\xi(x^t) = \min\left\{1, \frac{\tau(x^t)}{\|\nabla f_\xi(x^t)\|}\right\}\nabla f_\xi(x^t). \tag{28}$$

we have

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1}) \mid x^t\right] &\leq f(x^t) - \gamma_t \left\langle \nabla f(x^t), \mathbb{E}\left[g^t\right]\right\rangle + \frac{\gamma_t^2 L}{2}\mathbb{E}\left[\left\|g^t\right\|^2\right] \\
&\leq f(x^t) - \gamma_t \left\langle \nabla f(x^t), \mathbb{E}\left[g^t\right] \pm \overline{\alpha}(x^t)\nabla f(x^t)\right\rangle + \frac{\gamma_t^2 L}{2}\mathbb{E}\left[\left\|\alpha_{\xi^t}(x^t)\nabla f_\xi(x^t)\right\|^2\right] \\
&\overset{(5)}{\leq} f(x^t) - \gamma_t\, \overline{\alpha}(x^t)\left\|\nabla f(x^t)\right\|^2 - \gamma_t \left\langle \nabla f(x^t), \mathbb{E}\left[g^t\right] - \overline{\alpha}(x^t)\nabla f(x^t)\right\rangle + \frac{\gamma_t^2 L}{2}\left(\tau(x^t)\right)^2 \\
&\overset{(22)}{\leq} f(x^t) - \gamma_t\, \overline{\alpha}(x^t)\left\|\nabla f(x^t)\right\|^2 + \gamma_t \left\|\nabla f(x^t)\right\| \underbrace{\left\|\mathbb{E}\left[g^t\right] - \overline{\alpha}(x^t)\nabla f(x^t)\right\|}_{G_t} + \frac{\gamma_t^2 L}{2}\left(\tau(x^t)\right)^2 \\
&\overset{(7)}{\leq} f(x^t) - \gamma_t\, \overline{\alpha}(x^t)\left\|\nabla f(x^t)\right\|^2 + \gamma_t \left\|\nabla f(x^t)\right\| G_t + \frac{\gamma_t^2 L}{2}\left(\left\|\nabla f(x^t)\right\| + \sigma_q (1-p)^{-1/q}\right)^2 \\
&\leq f(x^t) - \gamma_t\left(\overline{\alpha}(x^t) - \gamma_t L\right)\left\|\nabla f(x^t)\right\|^2 + \gamma_t \left\|\nabla f(x^t)\right\| G_t + \gamma_t^2 L \sigma_q^2 (1-p)^{-2/q} \\
&\overset{(23)}{\leq} f(x^t) - \gamma_t\left(\overline{\alpha}(x^t) - \gamma_t L\right)\left\|\nabla f(x^t)\right\|^2 + \frac{\gamma_t}{2}\left(\beta\left\|\nabla f(x^t)\right\|^2 + \beta^{-1}G_t^2\right) + \gamma_t^2 L \sigma_q^2 (1-p)^{-2/q},
\end{aligned}$$

where in the last step, we used FenchelYoung inequality for $\beta > 0$. Rearranging the terms yields the desired result.

B.3. Proof of Theorem 3 (QC-SGD)

Denote $h := 1 - p$, then Lemma 2 (with suppressed expectation condition) gives

$$\mathbb{E}\left[f(x^{t+1})\right] - f(x^t) \leq -\gamma_t\left(p - \beta/2 - \gamma_t L\right)\left\|\nabla f(x^t)\right\|^2 + \frac{\gamma_t}{2}\beta^{-1}h^{2-2/q}\sigma_q^2 + \gamma_t^2 L\sigma_q^2 h^{-2/q},$$

where we also used the fact that $\overline{\alpha}(x) \geq p$. Next, we choose the step size as

$$0 \leq \gamma_t \leq \frac{2p - \beta - c}{2L}$$

to enforce condition $p - \beta/2 - \gamma_t L \geq c/2$. This leads to

$$\mathbb{E}\left[f(x^{t+1})\right] - f(x^t) \leq -\frac{c}{2}\gamma_t\left\|\nabla f(x^t)\right\|^2 + \frac{\gamma_t}{2}\beta^{-1}h^{2-2/q}\sigma_q^2 + \gamma_t^2 L\sigma_q^2 h^{-2/q}.$$

After rearranging the terms, we have a recursion

$$c\gamma_t\left\|\nabla f(x^t)\right\|^2 \leq 2\left(f(x^t) - \mathbb{E}\left[f(x^{t+1})\right]\right) + \sigma_q^2\gamma_t h^{-2/q}\left(\beta^{-1}h^2 + 2\gamma_t L\right).$$

Summing over t from 0 to $T-1$ and dividing over $\Gamma_T = \sum_{t=0}^{T-1}\gamma_t$ leads to the final result after unrolling the recursion

$$\frac{c}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t\mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{2\left(f(x^0) - \mathbb{E}\left[f(x^T)\right]\right)}{\Gamma_T} + \frac{\sigma_q^2}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t h^{-2/q}\left(2L\gamma_t + \beta^{-1}h^2\right).$$

B.4. Proof of Theorem 5 (DP-QC-SGD)

As a reminder the original method (4) is changed via mini-batching and adding Gaussian noise

$$x^{t+1} = x^t - \gamma_t \underbrace{\frac{1}{B}\sum_{j=1}^{B}\left(g_j^t + z^t\right)}_{\tilde{g}^t}, \qquad g_j^t = \min\left\{1, \frac{\tau(x^t)}{\left\|\nabla f_{\xi_j^t}(x^t)\right\|}\right\}\nabla f_{\xi_j^t}(x^t), \tag{29}$$

where $z^t \sim \mathcal{N}\left(0, \left(\tau(x^t)\right)^2 \sigma_{\text{DP}}^2 \mathbf{I}\right)$.

Proof By inspecting the proof of Lemma B.2, namely L-smoothness inequality

$$\mathbb{E}\left[f(x^{t+1}) \mid x^t\right] \leq f(x^t) - \gamma_t \left\langle \nabla f(x^t), \mathbb{E}\left[\tilde{g}^t\right] \pm \overline{\alpha}(x^t)\nabla f(x^t)\right\rangle + \frac{\gamma_t^2 L}{2}\mathbb{E}\left[\left\|\tilde{g}^t\right\|^2\right] \tag{30}$$

it is clear that the DP-SGD extension affects the last two terms with $\tilde{g}^t$. Next we show how DP modification affects the second moment and “bias” of the gradient estimator.

Due to the independence of ξ_j^t, the second moment of the stochastic gradient estimator can be upper-bounded as

$$\begin{aligned}
\mathbb{E}\left\|\tilde{g}^t\right\|^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{j=1}^{B}\left(g_j^t + z^t\right)\right\|^2 \overset{(21)}{\leq} 2\mathbb{E}\left\|\frac{1}{B}\sum_{j=1}^{B} g_j^t\right\|^2 + 2\mathbb{E}\left\|z^t\right\|^2 \leq \frac{2}{B^2}\sum_{j=1}^{B}\mathbb{E}\left\|g_j^t\right\|^2 + 2\mathbb{E}\left\|z^t\right\|^2 \\
&\overset{(29)}{\leq} \frac{2}{B^2}\sum_{j=1}^{B}\left(\tau(x^t)\right)^2 + 2\left(\tau(x^t)\right)^2\sigma_{\text{DP}}^2 = 2\left(\tau(x^t)\right)^2\left(1/B + \sigma_{\text{DP}}^2\right).
\end{aligned} \tag{31}$$

Inequality (8) from Lemma 1 is modified in the following way due to $\mathbb{E}\left[z^t\right] = 0$ for every t:

$$\begin{aligned}
\left\|\mathbb{E}\left[\tilde{g}^t\right] - \overline{\alpha}(x^t)\nabla f(x^t)\right\| &= \left\|\frac{1}{B}\sum_{j=1}^{B}\mathbb{E}\left[g_j^t\right] - \overline{\alpha}(x^t)\nabla f(x^t)\right\| \\
&\overset{(20)}{\leq} \frac{1}{B}\sum_{j=1}^{B}\left\|\mathbb{E}\left[g_j^t\right] - \overline{\alpha}(x^t)\nabla f(x^t)\right\| \\
&\overset{(8)}{\leq} (1-p)^{1-1/q}\sigma_q
\end{aligned} \tag{32}$$

Convergence proof based on Section B.3 is changed in the following way

$$\begin{aligned}
\mathbb{E}\left[f(x^{t+1}) \mid x^t\right] &\leq f(x^t) - \gamma_t\left\langle \nabla f(x^t), \mathbb{E}\left[\tilde{g}^t\right]\right\rangle + \frac{\gamma_t^2 L}{2}\mathbb{E}\left[\left\|\tilde{g}^t\right\|^2\right] \\
&\leq f(x^t) - \gamma_t\,\overline{\alpha}(x^t)\left\|\nabla f(x^t)\right\|^2 + \frac{\gamma_t}{2}\left(\beta\left\|\nabla f(x^t)\right\|^2 + \beta^{-1}G_t^2\right) + \gamma_t^2 L\left(1/B + \sigma_{\text{DP}}^2\right)\left(\tau(x^t)\right)^2 \\
&\leq f(x^t) - \gamma_t\left(\overline{\alpha}(x^t) - \beta/2\right)\left\|\nabla f(x^t)\right\|^2 + \frac{\gamma_t}{2}\beta^{-1}\sigma_q^2(1-p)^{2-2/q} \\
&\qquad + \gamma_t^2 L\left(1/B + \sigma_{\text{DP}}^2\right)\left(\left\|\nabla f(x^t)\right\| + \sigma_q(1-p)^{-1/q}\right)^2 \\
&\leq f(x^t) - \gamma_t\left(p - \beta/2 - 2\gamma_t L\left(1/B + \sigma_{\text{DP}}^2\right)\right)\left\|\nabla f(x^t)\right\|^2 \\
&\qquad + \frac{\gamma_t}{2}\beta^{-1}\sigma_q^2(1-p)^{2-2/q} + 2\gamma_t^2 L\left(1/B + \sigma_{\text{DP}}^2\right)\sigma_q^2(1-p)^{-2/q}.
\end{aligned}$$

Denote $\mathfrak{S} \coloneqq 1/B + \sigma_{\mathrm{DP}}^2$, then modified step size condition would be then

$$\gamma_t \leq \frac{p - \beta/2 - c}{2\mathfrak{S}L}. \tag{33}$$

to guarantee that $\overline{\alpha}(x^t) - \beta/2 - 2\gamma_t L\mathfrak{S} \geq c$. This leads to

$$c\gamma_t \left\|\nabla f(x^t)\right\|^2 \leq \mathbb{E}\left[f(x^{t+1}) \mid x^t\right] - f(x^t) + \frac{\gamma_t}{2}\beta^{-1}\sigma_q^2 h^{2-2/q} + 2\gamma_t^2 L\mathfrak{S}\sigma_q^2 h^{-2/q}.$$

Summing over t from 0 to $T-1$ and dividing over $\Gamma_T = \sum_{t=0}^{T-1}\gamma_t$ leads to the final result

$$\frac{c}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t \mathbb{E}\left[\left\|\nabla f(x^t)\right\|^2\right] \leq \frac{f(x^0) - \mathbb{E}\left[f(x^T)\right]}{\Gamma_T} + \frac{\sigma_q^2}{\Gamma_T}\sum_{t=0}^{T-1}\gamma_t h^{-2/q}\left(\beta^{-1}h^2/2 + 2\gamma_t L\mathfrak{S}\right). \tag{34}$$

■