Revisiting Gradient Clipping: Stochastic bias and tight convergence guarantees

Anastasia Koloskova [*1] **Hadrien Hendrikx** [*2] **Sebastian U. Stich** [3]

Abstract

Gradient clipping is a popular modification to standard (stochastic) gradient descent, at every iteration limiting the gradient norm to a certain value $c > 0$. It is widely used for example for stabilizing the training of deep learning models (Goodfellow et al., 2016), or for enforcing differential privacy (Abadi et al., 2016). Despite popularity and simplicity of the clipping mechanism, its convergence guarantees often require specific values of c and strong noise assumptions.

In this paper, we give convergence guarantees that show precise dependence on arbitrary clipping thresholds c and show that our guarantees are tight with both deterministic and stochastic gradients. In particular, we show that (i) for deterministic gradient descent, the clipping threshold only affects the higher-order terms of convergence, (ii) in the stochastic setting convergence to the true optimum cannot be guaranteed under the standard noise assumption, even under arbitrary small stepsizes. We give matching upper and lower bounds for convergence of the gradient norm when running clipped SGD, and illustrate these results with experiments.

1. Introduction

This paper focuses on solving general minimization problem of the form

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

where f is a possibly non-convex, and possibly stochastic function. This setting covers many applications, e.g. it covers optimizing deterministic functions if $f_\xi \equiv f\ \forall\xi$. It also covers minimizing the empirical loss in machine learning applications, where $\mathcal{D}$ represents the uniform distribution over training datapoints, and $f_\xi(\mathbf{x})$ is the loss of model $\mathbf{x}$ on the datapoint ξ.

We focus on gradient descent methods with *gradient clipping* for solving (1). Given a clipping radius $c > 0$, step-size $\eta > 0$, and starting from a point $\mathbf{x}_0 \in \mathbb{R}^d$ the gradient clipping algorithm performs the following iterations:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\mathbf{g}_t\,, \quad \text{with} \quad \mathbf{g}_t = \mathrm{clip}_c(\nabla f_\xi(\mathbf{x}_t))\,, \tag{2}$$

where $\mathbf{g}_t$ is a clipped stochastic gradient, and the clipping operator is defined as

$$\mathrm{clip}_c(\mathbf{u}) = \min\left(1, \frac{c}{\|\mathbf{u}\|}\right)\mathbf{u}\,, \qquad \text{for } \mathbf{u}\in\mathbb{R}^d. \tag{3}$$

Gradient clipping is widely used to *stabilize* the training of neural networks, by preventing large occasional gradient values from harming it (Goodfellow et al., 2016). This is particularly useful for mitigating outliers in the training data, and training recurrent models (Pascanu et al., 2012; 2013), in which the noise can induce very large gradients.

Gradient clipping is also an essential part of privacy-preserving machine learning. The widely-used Gaussian Mechanism (Dwork & Roth, 2014) adds noise to the individual gradients to add uncertainty about their true value. Yet, it requires the gradients to have bounded norms for the privacy guarantees to hold. In practice, bounded gradients are enforced through clipping (Abadi et al., 2016).

Gradient clipping has already been widely studied, as we detail in the next section. However, many works choose a specific value for the clipping threshold c in order to guarantee convergence. This suggests that c should be carefully tuned in practice, which is highly undesirable, in particular since the clipping threshold might be dictated by other (e.g., privacy) concerns. Besides, most works impose strong assumptions on the stochastic gradient noise, through either large batches (and thus small stochastic noise), angle conditions, or uniform boundedness of the norm, that might not hold in practice.

In this work, we precisely characterize how the clipping threshold c affects the convergence properties of clipped-SGD for *any* clipping threshold c.

[*]Equal contribution [1]EPFL, Switzerland [2]Inria Grenoble, France (work done in part while at EPFL) [3]CISPA Helmholtz Center for Information Security, Germany. Correspondence to: Anastasia Koloskova <anastasia.koloskova@epfl.ch>, Hadrien Hendrikx <hadrien.hendrikx@inria.fr>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

We consider deterministic and stochastic functions separately, as clipping affects these two settings in different ways.

In the *deterministic case*, clipping only changes the magnitude of the applied gradients, but not their direction. This means that clipped gradient descent can reach the critical points of f, however slower. Intuitively, as the algorithm converges, the gradients become small in magnitude and are not clipped eventually. This means that clipping affects only the speed during the first phase when the gradients are large in magnitude. The main challenge is to tightly characterize this overhead.

In the *stochastic case*, the story is different: the individual stochastic gradients can be large even though the expected gradient is small. Even at the critical points of f, where the expected (full) gradient is zero, there is some probability that individual stochastic gradients are clipped. Moreover, as we do not assume any symmetry of the stochastic gradients, the expected clipped gradient might be non-zero even at critical points of f, forcing the algorithm to drift from these critical points. The direct consequence of this is that clipped SGD *does not converge to the critical points of f in general* (Chen et al., 2020), but only to some neighborhood. In this paper we study the bias introduced by clipping and show that it depends on the noise variance σ^2 and the clipping parameter c, that we precisely define in Section 1.1. As we will further detail in Section 1.2, existing works circumvent this difficulty either by using large clipping thresholds or large mini-batches, or by making strong assumptions on the noise such as uniform boundness, restricted angles between stochastic gradients, etc., and usually requiring specific values for c. Instead, we tightly analyze the convergence of clipped SGD and *characterize precisely* the bias introduced by clipping without any additional assumptions.

More specifically, our contributions are the following:

- In the deterministic setting, we analyze the convergence behavior of clipped gradient descent for non-convex, convex and strongly convex functions. Our analysis shows that in all the cases after some transient regime, clipping does not affect the convergence rate. This initial phase does not affect the leading term of convergence in the convex and non-convex cases. However, in the strongly convex case, this unavoidable initial phase does not ensure linear forgetting of the initial conditions, resulting in a substantial slowdown.
- For stochastic gradients, we show that clipped SGD under the 'heavy-tailed' assumption converges to a neighbourhood of size $\min\{\sigma, \sigma^2/c\}$, measured in terms of the gradient norm.
- We show that this neighborhood size is tight: clipped SGD reduces the gradient norm up to $\min\{\sigma, \sigma^2/c\}$ indeed, provided the step-size is small enough.
- We frame our results using the (L_0, L_1)-smoothness assumption (Zhang et al., 2019), a standard relaxation of smoothness that is well suited to analyzing clipped algorithms.

Through these results, we aim at painting a thorough and accurate landscape of the convergence guarantees of clipping *under the same assumptions as standard SGD, and for any clipping threshold c.* Our goal is that these improved bounds will allow to tighten guarantees for all downstream applications, e.g. privacy, that use clipped-SGD convergence results as black box. Indeed, the clipping threshold is often viewed as an external parameter of the problem in these cases, whereas our flexible guarantees allow to optimize the bounds for c and trade-off convergence speed (or precision) and application-specific requirements.

1.1. Main assumptions

Before discussing related work we will first state the assumptions we use in our work.

Assumption on smoothness. The widely used smoothness assumption in the optimization literature (e.g. Nesterov et al., 2018) is the following:

Assumption 1.1 (Smoothness)**.** Function f satisfies

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L \|\mathbf{x} - \mathbf{y}\| , \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d .$$

Despite its widespread use, this assumption can be restrictive, as the constant L must capture the worst-case smoothness. Zhang et al. (2019) experimentally discovered that for various deep learning tasks, the local smoothness constant L decreases during training, and is proportional to the gradient norm. They reported that the local curvature (smoothness) in the final stages of training could be 1000 times smaller than the curvature at the initialization point (for LSTM training on the PTB dataset). (L_0, L_1)-smoothness (Zhang et al., 2019; 2020a) has been proposed as a natural relaxation of the classical smoothness assumption.

Assumption 1.2 ((L_0, L_1)-smoothness)**.** A differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ is said to be (L_0, L_1)-smooth if it verifies for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ with $\|\mathbf{x} - \mathbf{y}\| \leq \frac{1}{L_1}$:

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq (L_0 + \|\nabla f(\mathbf{x})\| L_1) \|\mathbf{x} - \mathbf{y}\| . \quad (4)$$

We use this as the main assumption in our work. This assumption recovers the standard smoothness Assumption 1.1 by taking $L_1 = 0$. However, taking $L_1 > 0$ allows to obtain smooth-like properties for functions that would otherwise not be smooth, such as $\mathbf{x} \mapsto \|\mathbf{x}\|^3$. Moreover, it is possible that L-smooth functions are (L_0, L_1)-smooth with both of the constants L_0, L_1 significantly smaller than L, such as for the exponential function $x \mapsto e^x$.

Table 1. Comparison of key assumptions and illustration of complexity estimates in the non-convex deterministic case (variance $\sigma = 0$).

reference	smoothness	variance bound	clipping threshold	further assumptions	rate (non-convex, $\sigma = 0$)	
Zhang et al. (2019)	2nd-order (L_0, L_1)	uniform bd.	$c = \Theta(\min\{L_0, \frac{L_0}{L_1}\})$		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}c}\right)$	$\eta \leq \min\left\{\frac{1}{10L_0}, \frac{1}{10cL_1}\right\}$
Zhang et al. (2020a)	(L_0, L_1)	uniform bd.	$c = \Theta(\max\{\epsilon, \frac{L_0}{L_1}\})$		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}}\right)$	$\eta \leq \frac{1}{10L_0}$
Chen et al. (2020)	L	expectation	arbitrary	pos. skewness	$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c} + \sqrt{\eta L c^2}\right)$	
Qian et al. (2021)	(L_0, L_1)	expectation	arbitrary	pos. alignment	$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c} + \sqrt{\eta L_0 c^2} + \sqrt{\eta L_1 c^3}\right)$	$\eta \leq \frac{1}{4cL_1}$
ours	(L_0, L_1)	expectation	arbitrary		$\mathcal{O}\left(\frac{1}{\sqrt{\eta T}} + \frac{1}{\eta T c}\right)$	$\eta \leq \frac{1}{9(L_0 + cL_1)}$

Note that the imposed bound $\|\mathbf{x} - \mathbf{y}\| \leq \frac{1}{L_1}$ in (4) is essential, as otherwise the global growth of the gradients would be similarly restricted as for standard smooth functions (thereby excluding functions such as the mentioned $\mathbf{x} \mapsto \|\mathbf{x}\|^3$).

In their work on clipping algorithms, Zhang et al. (2019) used a slightly stronger smoothness condition that required second-order differentiability. For twice-differentiable functions f, they defined (L_0, L_1)-smoothness as

$$\left\|\nabla^2 f(\mathbf{x})\right\| \leq L_0 + L_1 \left\|\nabla f(\mathbf{x})\right\|, \qquad \forall \mathbf{x} \in \mathbb{R}^d. \tag{5}$$

Later, Zhang et al. (2020a) noticed that the weaker Assumption 1.2 is sufficient for the study of clipping algorithms. We adopt their notion in our work.

Assumption on stochastic variance. Many works in clipping literature (Zhang et al., 2019; 2020a; Yang et al., 2022) assume the following

Assumption 1.3 (Uniform boundness)**.** We say that the stochastic noise of f_ξ is uniformly bounded by σ^2 if for all $\mathbf{x} \in \mathbb{R}^d$,

$$\Pr\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 \leq \sigma^2\right] = 1. \tag{6}$$

While this assumption allows to simplify the analysis of clipping algorithms, this is a very strong assumption.

By making Assumption 1.3 and using a sufficiently large clipping radius ($c > \sigma$), we can guarantee that stochastic gradients are not clipped at the critical points of f where $\nabla f(\mathbf{x}) = 0$. This ensures that the algorithm can converge to the exact critical points, simplifying the theoretical analysis in (Zhang et al., 2019; 2020a; Yang et al., 2022).

The uniform boundness Assumption 1.3 is a strong assumption and may not always be reflective of the real-world scenarios. For instance, if gradients are perturbed by Gaussian noise, the assumption of uniform boundness does not hold. Additionally, in machine learning applications where $\nabla f_\xi(\mathbf{x})$ represents gradients of a model $\mathbf{x}$ at different datapoints ξ from a dataset $\xi \in \mathcal{D}$, a uniform bound on σ may be large if the dataset $\mathcal{D}$ has even only one outlier point.

In this work, we use the weaker and the more standard variance definition instead (Lan, 2012; Dekel et al., 2012), sometimes called *heavy tailed noise* (Gorbunov et al., 2020).

Assumption 1.4 (Bounded variance)**.** We say that the variance of f_ξ is bounded by σ^2 if for all $\mathbf{x} \in \mathbb{R}^d$

$$\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\|^2\right] \leq \sigma^2. \tag{7}$$

Note that uniform boundedness implies bounded variance (with the same constant), but not the other way round.

1.2. Related work

The literature on gradient clipping is already extensive and still very active. We present the most relevant contributions for our work below and display a selection in Table 1.

Clipping stabilizes learning. Gradient clipping was originally proposed in (Mikolov, 2012) in order to tackle the gradient explosion problem in training of recurrent neural networks. Zhang et al. (2019) proposed to theoretically explain the question why clipped SGD improves the stability of (stochastic) first-order methods, by imposing a relaxed second-order (L_0, L_1)-smoothness assumption (see Equation (5)), and showing the convergence advantages of clipped SGD over unclipped SGD. However they rely on the strong Assumption 1.3 for the stochastic variance and chose the clipping threshold to a specific large enough value. The favorable convergence guarantees were then refined by Zhang et al. (2020a), while still relying on Assumption 1.3 on the stochastic noise and choosing specific value for the clipping threshold. Mai & Johansson (2021) show that this is also the case in the non-smooth setting.

Noise assumptions. Gradient clipping is often analyzed under uniform boundness Assumption 1.3 on the stochastic noise of the gradients in combination with choosing a large enough clipping threshold $c > \sigma$ (Zhang et al., 2019; 2020a; Yang et al., 2022). Choosing large enough values of c simplifies the theoretical analysis. However, in some applications the choice of the clipping threshold c might be dictated by other constraints, such as privacy constraints. Especially because in many practical applications the stochastic noise is heavy-tailed (Zhang et al., 2020b) it would entail large values of σ, and thus c.

To avoid the uniformly bounded noise assumption, some works impose other strong assumptions on the distribution of stochastic gradients. For instance, Qian et al. (2021) re-

strict the angle between stochastic gradients and the true gradient, and Chen et al. (2020) impose a symmetry assumption on the distribution of the stochastic gradients. Gorbunov et al. (2020) analyze clipping under bounded variance (see Assumption 1.4), however, they impose a strong assumption of the size of the minibatches used to scale linearly with T, thus making the effective stochastic variance to be diminishing with the number of iterations T as $\mathcal{O}\left(\frac{\sigma^2}{T}\right)$.

In this paper we take a different route from all these works, and analyse clipped SGD under the much weaker *bounded variance* assumption. Yet, instead of converging to the exact critical points of f, we quantify how large the drift due to clipping is, and thus obtain guarantees for any values of the batch size and c.

Noiseless case. Since the bulk of the assumptions concern the stochastic noise, our setting is the same as the papers mentioned above in the deterministic setting. However, in this case, we give sharper guarantees, essentially proving that clipping does not affect the leading convergence terms (see Table 1).

Clipped federated averaging. Zhang et al. (2022a) study clipping for the FedAvg (McMahan et al., 2016) algorithm, by clipping the model differences sent to the server. However, bounded gradients are needed, and the convergence rate does not recover the rate of FedAvg when the clipping threshold $c \to \infty$. Moreover, clipped FedAvg is biased even when using deterministic gradients. Liu et al. (2022) also study a clipped-FedAvg-like algorithm, and get rid of bias issues through assuming symmetric noise distributions around their means.

Differentially private SGD. Differential privacy has become the gold standard for protecting privacy, thus raising interest from the stochastic optimization community (Chaudhuri et al., 2011; Song et al., 2013; Duchi et al., 2014). However, to ensure differential privacy, boundedness of the stochastic gradients (Wang et al., 2017; Bassily et al., 2019; Das et al., 2022) (or a related condition, such as Lipschitzness of the objective function) has to hold. This is rarely true in practice, but instead enforced via clipping, such as in the DP-SGD algorithm (Abadi et al., 2016). Indeed, Lipschitzness requires the gradients to be bounded, whereas smoothness only requires boundedness of the Hessian. Although smoothness implies Lipschitzness on a bounded domain, this bound is usually very conservative and leads to poor guarantees.

Bagdasaryan et al. (2019) experimentally measure the effect of DP-SGD (clipping and additional noise) on model accuracy. They observe that the gradients do not converge to zero norm, so that the assumptions under which exact convergence is shown are often not verified indeed. Besides, underrepresented classes have higher gradient norm (so DP-SGD affects fairness).

Connection to adaptive methods. It is worth noting that clipped SGD is related to adaptive methods, such as the Adam algorithm (Kingma & Ba, 2015), or normalized SGD (Hazan et al., 2015; Levy, 2016), that also perform a scaling of the gradient. However, these two algorithms are not equivalent to clipped SGD and the convergence results for the Adam algorithm (Reddi et al., 2018; Zhang et al., 2020b; 2022b) and normalized SGD (Zhao et al., 2021) cannot be directly translated to the clipped SGD.

2. Deterministic Setting

In this section we consider gradient clipping algorithm (2) with full (deterministic) gradients, i.e. with

$$\nabla f_\xi(\mathbf{x}) \equiv \nabla f(\mathbf{x}), \qquad \forall \xi \in \mathcal{D}, \forall \mathbf{x} \in \mathbb{R}^d. \tag{8}$$

In this setting, the clipping operator (3) only changes the magnitude of the applied gradients, without changing its direction (as opposed to taking the expectation of clipped stochastic gradients). Thus, we can expect convergence to the exact minima, resp. critical points, of the function f. It still remains unclear how much does such a change in the magnitude of the gradients affect the convergence speed of the algorithm.

In our theoretical results we show that the drastic slow down happens *only* if the function f is *strongly convex*, in which case the initial conditions (distance to optimum) are not forgotten linearly anymore once clipping is applied. However, the leading term in the error ϵ is unaffected. If the function f is either *convex* or *non-convex*, the clipping threshold c does not affect the leading term of convergence, and affects *only the higher-order terms*.

2.1. Non-convex functions

Theorem 2.1 (non-convex)**.** *If f satisfies Assumption 1.2, then clipped gradient descent (2) with deterministic gradients (8) and with stepsize $\eta \leq [9(L_0 + cL_1)]^{-1}$ guarantees an error:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t)\| \leq \mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right), \tag{9}$$

where T is the number of iterations, $F_0 = f(\mathbf{x}_0) - f^\star$.

This theorem is a consequence of Theorem 3.3 for $\sigma = 0$.

Comparison to the unclipped gradient descent. The convergence rate of gradient descent (without clipping) assuming the standard L-smoothness Assumption 1.1 is equal to (Ghadimi & Lan, 2013):

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}\left(\frac{F_0}{\eta T}\right), \tag{10}$$

where the stepsize must be smaller than $\eta \leq \frac{1}{L}$. In the future discussion we will assume that $L_0 + cL_1 \leq L$, as we can always choose L_1 to be zero. In many cases, both L_0 and L_1 are significantly smaller than L (as discussed in Section 1.1). Thus, compared to the unclipped gradient descent, clipped gradient descent (2):

(i) allows for larger stepsizes η (up to the constant 9 in the stepsize constraint). This result is due to the refined (L_0, L_1) smoothness assumption and such an improvement in the stepsize has the same spirit as the discovery made by Zhang et al. (2019) for the (L_0, L_1) second-order smoothness assumption (5), although their bound on the stepsize is different.

(ii) has an additional term $\frac{F_0}{\eta T c}$ that depends on the clipping radius c. This term is of the order $\frac{1}{T}$, while the leading (the slowest decreasing, asymptotically dominating) term is of order $\frac{1}{\sqrt{T}}$.
If c is small, this term will slow down the algorithm significantly. However, when c is chosen larger than the final target accuracy ϵ, clipping affects the convergence speed only by a constant factor. Intuitively, this is because the number of steps when clipping happens is only a constant fraction of the total required number of iterations to converge.[1] As we frequently know the final target accuracy, our result shows that the clipping threshold could be set to avoid the adversarial effect of clipping. However, in practice the clipping threshold might be dictated by other needs.

(iii) has the different convergence measure $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|$ instead of $\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\mathbf{x}_t)\|^2$ that is more commonly used (e.g. in (10) for unclipped gradient descent).

Comparison to the prior work. We summarized differences to the prior works in Table 1. Zhang et al. (2019) and Zhang et al. (2020a) analyzed deterministic gradient clipping however setting the clipping threshold c to some specific, large enough values. Qian et al. (2021) and Chen et al. (2020) analyse clipped SGD under arbitrary choice of the clipping threshold c. In particular, assuming deterministic gradients ($\sigma = 0$), Qian et al. (2021) obtain the convergence rate of $\mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c} + c\sqrt{\eta L_0} + c^{3/2}\sqrt{\eta L_1}\right)$, $\eta < 1/4cL_1$, as the two last terms $c\sqrt{\eta L_0} + c^{3/2}\sqrt{\eta L_1}$ do not decrease to zero under the constant stepsizes η. That is strictly worse than ours in Theorem 2.1. Chen et al. (2020) prove the rate $\mathcal{O}\left(\sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c} + c\sqrt{\eta L}\right)$ without any constraint on the stepsize, however they have to take small stepsizes $\eta = 1/\sqrt{T}$ as the term $c\sqrt{\eta L}$ is not decreasing in T. Notably, this terms prevents the error from converging to 0 under constant step-sizes, which can be obtained in the deterministic setting, as we showed above.

[1] Formally: because the final accuracy $\epsilon = \sqrt{F_0/\eta T} + F_0/\eta T c \geq \sqrt{F_0/\eta T}$, and thus if clipping threshold is larger than that, $c \geq \sqrt{F_0/\eta T}$, then the convergence speed $\sqrt{F_0/\eta T} + F_0/\eta T c \leq 2\sqrt{F_0/\eta T}$ is affected only by a constant.

2.2. Convex functions

We now prove an equivalent theorem when f is convex, i.e. assuming additionally:

Assumption 2.2 (Convexity)**.** Function f satisfies

$$f(\mathbf{x}) - f(\mathbf{y}) \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle\,, \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d\,.$$

We also assume that infimum of f is achieved in $\mathbb{R}^d$.

Theorem 2.3 (convex)**.** *If f is L-smooth[2] (Assumption 1.1), (L_0, L_1) smooth (Assumption 1.2) and convex (Assumption 2.2), then clipped gradient descent (2) with deterministic gradients (8) and with stepsize $\eta \leq (L_0 + cL_1)^{-1}$ guarantees an error:*

$$f(\mathbf{x}_T) - f^\star \leq \mathcal{O}\left(\frac{R_0^2}{\eta T} + \frac{R_0^4 L}{\eta^2 T^2 c^2}\right)\,, \tag{11}$$

where $R_0^2 = \|\mathbf{x}_0 - \mathbf{x}_\star\|^2$, $f^\star = f(\mathbf{x}_\star)$, and $\mathbf{x}^\star = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

In comparison, under the same assumptions as in Theorem 2.3, unclipped gradient descent converges at the rate

$$f(\mathbf{x}_T) - f^\star \leq \mathcal{O}\left(\frac{R_0^2}{\eta T}\right)$$

under the condition that the stepsize is smaller than $\eta \leq L^{-1}$ (Nesterov et al., 2018). Similarly to the non-convex case, the convergence rate of the clipped gradient descent is slowed down by the higher-order term $R_0^4 L/\eta^2 T^2 c^2$. Yet, again, it is enough to set the clipping threshold c bigger than the final target accuracy ϵ (multiplied by $\sqrt{L}$ this time) to avoid the slowdown effect of this term, since for high accuracies more time is actually spent using unclipped gradients.

Also, similarly to the non-convex case, clipped GD allows for the larger stepsizes η that would result in the faster convergence.

2.3. Strongly convex functions

In this section we consider strongly-convex functions f.

Assumption 2.4 (strong-convexity)**.** There exists a constant $\mu > 0$ such that function f satisfies for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$,

$$f(\mathbf{x}) - f(\mathbf{y}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|_2^2 \leq \langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{y}\rangle\,.$$

Similarly to the convex and non-convex cases, clipping does not affect the leading term of convergence (as $\epsilon \to 0$) as we show in the following theorem. This is due to the fact that for any fixed $c > 0$, gradients are eventually never clipped.

[2] We can relax this assumption. Equation (11) also holds with L replaced by L_T, defined as $L_T := \max_{t \leq T}\{L_0 + L_1\|\nabla f(\mathbf{x}_t)\|\}$.

Theorem 2.5 (Strongly convex case). *If f is μ-strongly convex (Assumption 2.4), L-smooth[3] (Assumption 1.1) and (L_0, L_1) smooth (Assumption 1.2), then clipped gradient descent (2) with deterministic gradients (8) and with stepsize $\eta \leq (L_0 + cL_1)^{-1}$ needs at most*

$$T = \mathcal{O}\left(\frac{1}{\mu\eta}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{R_0}{c\eta}\min\left(\sqrt{\frac{L}{\mu}}, \frac{LR_0}{c}\right)\right) \quad (12)$$

iterations to reach accuracy $R_T^2 \leq \epsilon$, where $R_t^2 = \|\mathbf{x}_t - \mathbf{x}_\star\|^2$ and $\mathbf{x}^\star = \arg\min_{\mathbf{x}} f(\mathbf{x})$.

Compared to the unclipped case, there is an extra term in the strongly convex case, which does not decrease with ϵ and corresponds to the overhead of clipping. This means that during the initial phase of convergence, when the gradients are clipped, the convergence speed is sublinear, and one would have to set $c = \mathcal{O}\left(1/\log(\frac{1}{\epsilon})\right)$ in order for the clipping do not affect the convergence speed, that is much larger than in the non-convex and convex cases.

Intuitively, since the clipped gradient norm is fixed, the iterates can actually move only up to ηc each step towards the optimum. If the initial distance to the optimum R_0 happened to be large, in the best case scenario, one would need at least $\frac{R_0}{\eta c}$ steps to reach the optimum. The dependency on c for the first term in the $\min$ is tight.[4]

Note that after a constant (independent of ϵ) number of iterations, the algorithm converges linearly, at a rate that depends on $(L_0 + cL_1)$ only, that can be significantly smaller than the dependency on L in the GD convergence rate.

3. Stochastic Functions

In the deterministic setting gradient clipping achieves convergence to the exact minimizer or a stationary point, respectively, and clipping only affects initial convergence speed. Yet, this does not hold in the stochastic setting where clipping introduces *unavoidable bias*. The main reason behind this is that the expectation of the clipped stochastic gradients is different (in both norm and direction) from the clipped true gradient.

While it was known before that the clipped SGD does not converge under the bounded variance Assumption 1.4 (Chen et al., 2020), in this section we will *precisely characterize* lower bounds on the error that clipped SGD can achieve, and then we provide upper bounds that match our lower bounds.

[3]We can relax this assumption by defining instead $L := \max_{t \leq T}\{L_0 + L_1\|\nabla f(\mathbf{x}_t)\|\}$.

[4]Formally, consider the function $x \mapsto \frac{1}{2}x^2$ and initial iterate $x_0 = 1$. Suppose we aim to reach a target accuracy $\epsilon < \frac{1}{4}$ with clipping threshold $c < \frac{1}{2}$. We see that unless $|x| < c$, the gradient $f'(x) = x$ will get clipped to value c, and hence after $\frac{1}{2c}$ iterations, cannot reach a point with norm smaller than $\frac{1}{2}$ and squared norm less than $\frac{1}{4}$ respectively.

3.1. Unavoidable bias introduced by clipping

If the gradient clipping algorithm converges, it has to be towards its fixed points, i.e. to points $\mathbf{x}^\star$ such that $\mathbb{E}\left[\text{clip}_c(\nabla f_\xi(\mathbf{x}^\star))\right] = 0$. This is achieved in the limit of small step-sizes (to counter stochastic noise).

We now formally show a lower bound that states that fixed points of clipped SGD are not necessary optimal or critical points of the objective function f. In fact, there exist stochastic gradient noise distributions under which critical points of clipped SGD are σ far away from critical points of f.

Theorem 3.1 (Small clipping radius). *We fix a class of functions that have variance at most σ^2 (Def 1.4) and smoothness parameters $L_0 = 1, L_1 = 0$ (Assumption 1.2). Then, for any clipping threshold $c \leq 2\sigma$, we can find a function f within this fixed class such that the fixed points of clipped-SGD exist (i.e. points $\mathbf{x}^\star$ which verify $\mathbb{E}\left[\text{clip}_c(\nabla f_\xi(\mathbf{x}^\star))\right] = 0$), and that for all such fixed-points $\mathbf{x}^\star$ of clipped-SGD it holds that $\|\nabla f(\mathbf{x}^\star)\| \geq \sigma/12$.*

Proof sketch. We define the stochastic function

$$f_\xi(x) = \frac{1}{2}\begin{cases}(x+a)^2 & \text{w. p. } p\\ x^2 & \text{w. p. } (1-p)\end{cases}$$

where $a > 0$ and $p < 1/2$. The expected function is thus $f(x) = \frac{1}{2}[p(x+a)^2 + (1-p)x^2]$.

The result is then obtained by choosing $a = 4\sigma$ and $p = (2-\sqrt{3})/4 < 1/4$ is such that $p(1-p) = 1/16$. The statement follows by using the standard algebra, as detailed in Appendix C.5. □

The previous impossibility result holds when the clipping radius is small ($c < 2\sigma$). We further show that by taking a larger clipping radius c, we can reduce the neighborhood size to which clipped-SGD converges from σ to σ^2/c, but cannot completely eliminate it.

Theorem 3.2 (Large clipping radius). *We fix a class of functions that have variance at most σ^2 (Def 1.4) and smoothness parameters $L_0 = 1, L_1 = 0$ (Assumption 1.2). Then, for any clipping threshold $c \leq 2\sigma$, we can find a function f within this fixed class such that the fixed points $\mathbf{x}^\star$ of clipped-SGD exist ($\mathbb{E}\left[\text{clip}_c(\nabla f_\xi(\mathbf{x}^\star))\right] = 0$) and $\|\nabla f(\mathbf{x}^\star)\| \geq \sigma^2/6c$.*

Proof sketch. We use the same function as Theorem 3.1, this time with $a = 2c$ and $p(1-p) = \sigma^2/a^2$. □

Our lower bounds in Theorems 3.1 and 3.2 mean that with only assuming Def. 1.4 and Assumption 1.2, there cannot

exist problem-dependent values for c that would give exact convergence for any function. We note that the fact that clipping might introduce a bias when the noise is only bounded in expectation is not new (Chen et al., 2020). The interesting thing about Theorems 3.1 and 3.2 is that they are matched by the upper bound in Theorem 3.3, meaning that we *precisely capture* the strength of the bias introduced in this case.

Uniformly bounded noise. Note that this lower bound crucially relies on the noise being bounded by σ *in expectation* (Assumption 1.4). Indeed, the results hinge on the fact that stochastic gradients are clipped with probability p, thus introducing a bias. If we keep this Bernoulli noise constant (and therefore will ensure uniformly bounded noise Assumption 1.3) and increase c, then this bias would completely disappear for $c \approx a$, because then the clipping radius would be larger than the uniform bound on variance.

3.2. Convergence results

We now introduce the central result of this paper: the convergence of clipped SGD that match the lower bounds above.

Theorem 3.3. *If f is (L_0, L_1)-smooth (but not necessarily convex) and we run clipped SGD for T steps with step-size $\eta \leq 1/[9(L_0+cL_1)]$, then $\min_{t\in[0,T]} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|$ is upper bounded by*

$$\mathcal{O}\left(\min\left\{\sigma, \frac{\sigma^2}{c}\right\} + \sqrt{\eta(L_0+cL_1)}\sigma + \sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right),$$

where $F_0 = f(\mathbf{x}_0) - f^\star$.

The convergence rate contains four terms: the first term does not decrease with neither the stepsize η nor the number of iterations T and it is due to unavoidable bias, as explained in the previous section. Due to Theorems 3.1, 3.2 this term is tight and cannot be improved. The second term is the stochastic noise term that decreases with the stepsize, and the last two terms are the optimization terms that describe how clipping affects convergence when the stochastic noise is zero ($\sigma = 0$), matching the convergence in Theorem 2.1. Note that we precisely quantify the bias of clipped SGD under the general bounded variance assumption.

Comparison to the unclipped SGD. Under the standard smoothness Assumption 1.1, unclipped SGD requires the stepsize to be smaller than $\eta \leq L^{-1}$ and it converges at the rate (Bottou et al., 2018)

$$E_T \leq \mathcal{O}\left(\sqrt{\eta L}\sigma + \sqrt{\frac{F_0}{\eta T}}\right).$$

where[5] $E_T := \left(\frac{1}{T}\sum_{t=0}^{T}\left\|\nabla f(\mathbf{x}_t)\right\|^2\right)^{\frac{1}{2}}$. In comparison to the unclipped SGD, clipped SGD (2),

- Has an unavoidable bias term $\min\left\{\sigma, \frac{\sigma^2}{c}\right\}$ that we discussed in detail in the previous Section 3.1.
- Has a smaller stochastic noise term, assuming that[6] $L_0 + cL_1 \leq L$.
- Similarly to the deterministic case (Thm. 2.1), has an additional higher-order term $F_0/\eta T c$.

We want to highlight that the bias term $\min\left\{\sigma, \frac{\sigma^2}{c}\right\}$ in our convergence rate is tight. This implies in particular that under general expected bounded noise Assumption 1.4, clipped SGD cannot converge to the exact critical points of f, but the convergence neighborhood size decreases with increasing c.

Similarly to Zhang et al. (2019), clipped SGD improves over unclipped SGD the dependence on the smoothness parameter from L to L_0. In contrast to Zhang et al. (2019) in our work we do not assume specific values of c, and use a weaker expected bounded noise Assumption 1.4.

The complete proof of Theorem 2.5 can be found in Appendix C. We now give an intuitive proof sketch of Theorem 3.3.

Proof sketch (Theorem 3.3). The proof is split into two different cases, depending on how big c is compared to σ.

Case $c \leq 4\sigma$. In this case, according to the lower bound in Theorem 3.1, we can only show convergence of the gradient norm up to $\Theta(\sigma)$. To achieve this, we only need to consider the case when the gradients have $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$, since otherwise the convergence to $\Theta(\sigma)$ is already achieved. We can show that in the case of large gradients (i.e. $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$), the standard convergence results hold under uniformly bounded noise, because the gradient norm is large enough to compensate the (fixed) variance.

Case $c \geq 4\sigma$. In this case, we analyze clipped SGD as some form of biased gradient descent. Note that under Uniform Boundedness (Assumption 1.3), the bias eventually vanishes for such large clipping thresholds. We precisely quantify the remaining bias term B_t instead. After some manipulations, we obtain descent terms such as Equation (32) from Appendix C, and a bias term that writes as

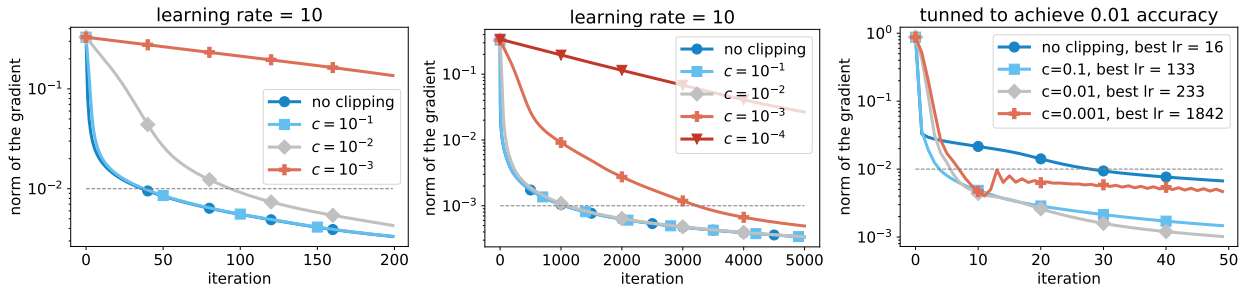
$$B_t = \left\|\mathbb{E}\left[\text{clip}_c(\nabla f_\xi(\mathbf{x}_t))\right] - \text{clip}_c(\nabla f(\mathbf{x}_t))\right\|^2. \quad (13)$$

Using that the clipping operation is a projection on a convex set (on a ball of a radius c), then we can bound this term directly as $B_t \leq \sigma^2$. In particular, we can cancel it with descent terms when $\|\nabla f(\mathbf{x}_t)\|$ is large enough.

When $\|\nabla f(\mathbf{x})\| \leq c/2$, then we can be more precise. In particular we can show that the probability that the stochastic gradient is clipped is smaller than σ^2/c^2. Using this, we

[5] Note that $E_T \geq \min_{t\in[0,T]} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|$.

[6] We can always choose $L_0 = L$ and $L_1 = 0$ to satisfy this equation. Frequently, both L_0 and L_1 are much smaller than L (see discussion in Section 1.1)



(a) Constant stepsize, target $\epsilon = 10^{-2}$ (b) Constant stepsize, target $\epsilon = 10^{-3}$ (c) Tuned stepsize for target $\epsilon = 10^{-2}$

Figure 1. Deterministic clipped gradient descent on the w1a dataset. We investigate the dependence of the convergence rate on the clipping parameter c. In Figures (a) and (b) we see that as soon as the clipping threshold is smaller or equal to the target gradient norm ϵ, the convergence speed is affected only by a constant. In Figure (c), we see that as the clipping threshold c decreases, the best tuned stepsize (tuned to reach $\epsilon = 10^{-2}$ fastest) decreases. These observations in accordance to the theory in Theorem 2.3.

can refine the estimate of the bias as

$$B_t \leq 8\frac{\sigma^4}{c^2} + 32\frac{\sigma^4}{c^4}\left\|\nabla f(\mathbf{x})\right\|^2. \tag{14}$$

The σ^4/c^2 is the bias term that we find in the convergence rate, and the $\|\nabla f(\mathbf{x})\|^2$ term can be canceled with descent terms (that are also proportional to this) provided σ^4/c^4 is small enough. □

Comparison to the prior works. All of the prior works used stronger assumptions allowing for simplifications in their analysis, and allowing to mitigate the bias introduced by the clipping.

For example, (Zhang et al., 2019), (Zhang et al., 2020a), (Yang et al., 2022) considered large clipping thresholds ($c \geq \sigma$) and a stronger assumption of uniform boundness (Assumption 1.3), ensuring that the bias vanishes as we approach the optimum. The other prior work of (Gorbunov et al., 2020) used a specific clipping threshold c and a specific large enough batch size allowing also to consider only one of the cases ($c \geq 4\sigma$). They require the batch sizes to scale linearly with the number of iterations T, thus mitigating stochasticity in their gradients. (Qian et al., 2021) and (Chen et al., 2020) impose some symmetry assumptions on the distribution of the stochastic gradients, which allows them to mitigate the bias introduced by the clipping operator. In the limit case of entirely symmetric distribution, the clipping operator does not change the direction of expected gradient at any point, thus allowing for the similar convergence analysis as with deterministic gradients.

3.3. Extension to differentially private SGD

In differentially private SGD (Abadi et al., 2016) every individual stochastic gradient in the batch is getting clipped individually before averaging the gradients over the batch, i.e. the algorithm is

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \eta\left(\frac{1}{B}\sum_{i\in\mathcal{B}_t}\text{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) + \mathbf{z}_t\right), \tag{15}$$

where $\mathbf{z}_t \sim \mathcal{N}\left(0, \frac{\sigma_{\text{DP}}^2}{d}\mathbf{I}\right)$ is the additional noise due to differential privacy.

As detailed in Appendix C.4 we can extend our analysis to this algorithm in a straightforward way and show that T iterations of (15) allow to obtain gradient norm smaller than:

$$\mathcal{O}\left(\frac{L\eta}{c}\sigma_{\text{DP}}^2 + \sqrt{L\eta\sigma_{\text{DP}}} + \min\left(\sigma^2, \frac{\sigma^4}{c^2}\right) + \eta L\frac{\sigma^2}{B} + \frac{F_0}{\eta T} + \frac{F_0^2}{\eta^2 T^2 c^2}\right).$$

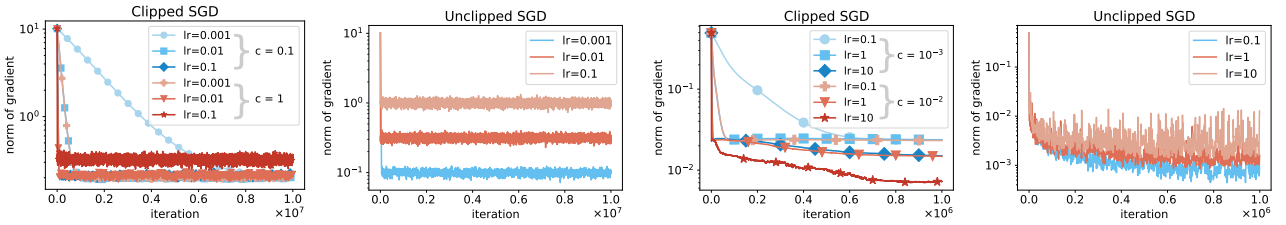
where B is the mini-batch size, and $L = L_0 + \max_t \|\nabla f(\mathbf{x}_t)\| L_1$, corresponding to the smoothness constant according to the standard L-smoothness assumption.

Similarly to the clipped-SGD algorithm considered previously, DP-SGD also suffers from a bias term $\min\left(\sigma^2, \sigma^4/c^2\right)$. Lower bounds in Theorems 3.1, 3.2 apply to DP-SGD, so this bias is also tight and unavoidable.

In comparison to the clipped SGD (2), DP-SGD has additional terms related to the injected privacy noise σ_{DP}, and the stochastic noise (fourth term) is reduced by a factor B due to mini-batching.

In order to have the formal privacy guarantees, one has to set the variance of additional DP noise appropriately, Abadi et al. (2016) prove that for $\sigma_{\text{DP}} \geq \Omega\left(cd\frac{\sqrt{T\log\frac{1}{\delta}}}{\epsilon}\right)$ DP-SGD is (ϵ, δ)-differentially private.

Related to prior work on DP-SGD that incorporate clipping in the convergence analysis (Chen et al., 2020; Yang et al.,



(a) Quadratic function with χ^2 stochastic noise

(b) Logistic regression on w1a dataset (batch sitze = 1).

Figure 2. Stochastic gradient descent on a quadratic function with χ^2 stochastic noise (left), and w1a dataset (right). Without clipping, decreasing the stepsize allows to achieve the smaller gradient norm. However, decreasing the stepsize with clipping does not allow to achieve better performance. This is because of the unavoidable bias term in Theorem 3.3.

2022), our convergence rates are proven only assuming bounded variance in expectation (Def. 1.4) and without extra assumptions on the noise. They showcase the effect of the clipping threshold on the convergence of DP-SGD.

4. Experiments

In this section, we investigate the performance of gradient clipping on logistic regression on the `w1a` dataset (Platt, 1998), and on the artificial quadratic function $f(\mathbf{x}) = \mathbb{E}_{\xi\sim\chi^2(1)}\left[f(\mathbf{x},\xi) := \frac{L}{2}\|\mathbf{x}\|^2 + \langle \mathbf{x}, \xi\rangle\right]$, where $\mathbf{x} \in \mathbb{R}^{100}$, we choose $L = 0.1$, and $\chi^2(1)$ is a (coordinate-wise) chi-squared distribution with 1 degree of freedom. The goal is to highlight our theoretical results.

Deterministic setting. In the deterministic setting, one of our insights was that clipping does not degrade performance too much as long as the clipping threshold is bigger than the final target accuracy. We test this in Figures 1(a), 1(b) for logistic regression on `w1a` dataset, by plotting the clipped-GD with different values of c, for different target accuracies. We see that to reach accuracy 10^{-3}, all values of c (except from $c = 10^{-4}$) perform relatively well. However, choosing $c = 10^{-3}$ is not advisable if we only want to reach an error $\epsilon = 10^{-2}$, as can be seen in Figure 1(b) (note the different scaling of the x-axis in both plots).

In Figure 1(c), we investigate the dependence between the clipping threshold c and the step-size. We tune the stepsize separately for each clipping parameter c over a logarithmic grid between 10^{-1} and 10^4, ensuring that the optimal value is not at the edge of the grid. The best stepsize is selected as the one that reaches the target gradient norm $\epsilon = 10^{-2}$ the fastest. In Figure 1(c) we see that choosing smaller clipping radius allows for larger step-sizes, which speeds-up convergence overall.

In particular, we have verified that in the deterministic setting, clipping does not harm learning as long as the threshold is not too small compared to the target accuracy. Besides, clipping stabilizes learning, thus allowing for larger step-sizes, and thus faster convergence.

Stochastic setting. We investigate clipped-SGD on both quadratic function with $\chi^2(1)$ stochastic noise, and logistic regression on `w1a` dataset. The results are plotted in Figure 2. They also verify the theory, since larger learning rates are always better when using clipping (compared to unclipped). It is also interesting to note that the curves are determined by the product $c\eta$, which is how large the step is when clipping happens. However, we see that in this case, with such small values of c, clipped-SGD does not quite reach the performance of vanilla SGD.

5. Conclusion

In this paper, we have rigorously analyzed gradient clipping, both in the deterministic setting and under standard noise assumptions. While previous works focus on exact convergence under strong assumptions (in particular, often for a fixed clipping threshold), we tightly characterized (with both upper and lower bounds) the bias introduced by clipped-SGD for any clipping threshold.

Our work paves the way for better understanding clipping when used with other algorithms, such as accelerated or momentum methods or FedAvg. In particular, it can lead to an improved analysis of privacy guarantees in applications that rely on clipped SGD as an underlying black box, on the one hand for existing, but also future applications.

Acknowledgments

The authors would like to thank Ryan McKenna and Martin Jaggi for useful discussions. We also thank anonymous reviewers for their valuable comments. SS acknowledges partial funding from a Meta Privacy Enhancing Technologies Research Award 2022. AK acknowledges funding from a Google PhD Fellowship.

References

Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., and Zhang, L. Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, pp. 308–318, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450341394. doi: 10.1145/2976749.2978318. URL https://doi.org/10.1145/2976749.2978318.

Bagdasaryan, E., Poursaeed, O., and Shmatikov, V. Differential privacy has disparate impact on model accuracy. *Proceedings of the 33rd International Conference on Neural Information Processing Systems*, 2019.

Bassily, R., Feldman, V., Talwar, K., and Guha Thakurta, A. Private stochastic convex optimization with optimal rates. *Advances in neural information processing systems*, 32, 2019.

Bottou, L., Curtis, F. E., and Nocedal, J. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018. doi: 10.1137/16M1080173. URL https://doi.org/10.1137/16M1080173.

Chaudhuri, K., Monteleoni, C., and Sarwate, A. D. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

Chen, X., Wu, S. Z., and Hong, M. Understanding gradient clipping in private sgd: A geometric perspective. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H. (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 13773–13782. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/file/9ecff5455677b38d19f49ce658ef0608-Paper.pdf.

Das, R., Kale, S., Xu, Z., Zhang, T., and Sanghavi, S. Beyond uniform lipschitz condition in differentially private optimization. *arXiv preprint arXiv:2206.10713*, 2022.

Dekel, O., Gilad-Bachrach, R., Shamir, O., and Xiao, L. Optimal distributed online prediction using mini-batches. *J. Mach. Learn. Res.*, 13(null):165–202, jan 2012. ISSN 1532-4435.

Duchi, J. C., Jordan, M. I., and Wainwright, M. J. Privacy aware learning. *Journal of the ACM (JACM)*, 61(6):1–57, 2014.

Dwork, C. and Roth, A. The algorithmic foundations of differential privacy. *Found. Trends Theor. Comput. Sci.*, 9(3–4):211–407, aug 2014. ISSN 1551-305X. doi: 10.1561/0400000042. URL https://doi.org/10.1561/0400000042.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep Learning*. MIT Press, 2016. http://www.deeplearningbook.org.

Gorbunov, E., Danilova, M., and Gasnikov, A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc. ISBN 9781713829546.

Hazan, E., Levy, K. Y., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, pp. 1594–1602, Cambridge, MA, USA, 2015. MIT Press.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In Bengio, Y. and LeCun, Y. (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

Lan, G. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133:1–33, 06 2012. doi: 10.1007/s10107-010-0434-y.

Levy, K. The power of normalization: Faster evasion of saddle points. 11 2016.

Liu, M., Zhuang, Z., Lei, Y., and Liao, C. A communication-efficient distributed gradient clipping algorithm for training deep neural networks. In *Advances in Neural Information Processing Systems*, 2022.

Mai, V. V. and Johansson, M. Stability and convergence of stochastic gradient clipping: Beyond lipschitz continuity and smoothness. In *International Conference on Machine Learning*, 2021.

McMahan, H. B., Moore, E., Ramage, D., Hampson, S., and y Arcas, B. A. Communication-efficient learning of deep networks from decentralized data. In *International Conference on Artificial Intelligence and Statistics*, 2016.

Mikolov, T. *Statistical Language Models based on Neural Networks*. Ph.D. thesis, Brno University of Technology, 2012.

Nesterov, Y. et al. *Lectures on convex optimization*, volume 137. Springer, 2018.

Pascanu, R., Mikolov, T., and Bengio, Y. Understanding the exploding gradient problem. *ArXiv*, abs/1211.5063, 2012.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*, ICML'13, pp. III–1310–III–1318. JMLR.org, 2013.

Platt, J. Fast training of support vector machines using sequential minimal optimization. *Advances in Kernel Methods - Support Vector Learning, MIT Press*, 1998.

Qian, J., Wu, Y., Zhuang, B., Wang, S., and Xiao, J. Understanding gradient clipping in incremental gradient methods. In Banerjee, A. and Fukumizu, K. (eds.), *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130 of *Proceedings of Machine Learning Research*, pp. 1504–1512. PMLR, 13–15 Apr 2021. URL https://proceedings.mlr.press/v130/qian21a.html.

Reddi, S. J., Kale, S., and Kumar, S. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=ryQu7f-RZ.

Song, S., Chaudhuri, K., and Sarwate, A. D. Stochastic gradient descent with differentially private updates. In *2013 IEEE global conference on signal and information processing*, pp. 245–248. IEEE, 2013.

Wang, D., Ye, M., and Xu, J. Differentially private empirical risk minimization revisited: Faster and more general. *Advances in Neural Information Processing Systems*, 30, 2017.

Yang, X., Zhang, H., Chen, W., and Liu, T.-Y. Normalized/clipped sgd with perturbation for differentially private non-convex optimization, 2022. URL https://arxiv.org/abs/2206.13033.

Zhang, B., Jin, J., Fang, C., and Wang, L. Improved analysis of clipping algorithms for non-convex optimization. *Advances in Neural Information Processing Systems*, 33:15511–15521, 2020a.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity, 2019. URL https://arxiv.org/abs/1905.11881.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020b. Curran Associates Inc. ISBN 9781713829546.

Zhang, X., Chen, X., Hong, M., Wu, S., and Yi, J. Understanding clipping for federated learning: Convergence and client-level differential privacy. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato, S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 26048–26067. PMLR, 17–23 Jul 2022a. URL https://proceedings.mlr.press/v162/zhang22b.html.

Zhang, Y., Chen, C., Shi, N., Sun, R., and Luo, Z. Adam can converge without any modification on update rules. *ArXiv*, abs/2208.09632, 2022b.

Zhao, S.-Y., Xie, Y.-P., and Li, W.-J. On the convergence and improvement of stochastic normalized gradient descent. *Science China Information Sciences*, 64, 2021.

A. Implications of (L_0, L_1) smoothness

Lemma A.1. *If Assumption 1.2 holds, then it also holds that*

$$f(\mathbf{y}) - f(\mathbf{x}) \leq \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{(L_0 + \|\nabla f(\mathbf{x})\| L_1)}{2} \|\mathbf{x} - \mathbf{y}\|^2 , \qquad \forall \mathbf{x}, \mathbf{y} \in \mathbb{R}^d \text{ with } \|\mathbf{x} - \mathbf{y}\| \leq \frac{1}{L_1} . \tag{16}$$

For the proof see (Zhang et al., 2020a), Appendix A.1.

Lemma A.2. *If Assumption 1.2 holds, then it also holds that*

$$\|\nabla f(\mathbf{x})\|^2 \leq 2(L_0 + L_1 \|\nabla f(\mathbf{x})\|) (f(\mathbf{x}) - f^\star) \qquad \forall \mathbf{x} \in \mathbb{R}^d ,$$

where $f^\star = \inf_{\mathbf{x}} f(\mathbf{x})$.

Proof of Lemma A.2. We start the proof by applying the previous Lemma A.1 for $\mathbf{y} = \mathbf{x} - \frac{1}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \nabla f(\mathbf{x})$. Note that $\|\mathbf{x} - \mathbf{y}\| = \frac{\|\nabla f(\mathbf{x})\|}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \leq \frac{1}{L_1}$ and we can apply the inequality:

$$f^\star \leq f\left(\mathbf{x} - \frac{1}{L_0 + \|\nabla f(\mathbf{x})\| L_1} \nabla f(\mathbf{x})\right) \overset{(16)}{\leq} f(\mathbf{x}) - \frac{1}{2(L_0 + \|\nabla f(\mathbf{x})\| L_1)} \|\nabla f(\mathbf{x})\|^2 ,$$

and rearranging gives us the desired property. □

B. Deterministic proofs

This section contains the main proofs from the paper. We skip the non-convex proof, since it will be a direct consequence of the stochastic result.

B.1. Convex case (Theorem 2.3)

Defining $\alpha_t = \min\{1, \frac{c}{\|\nabla f(\mathbf{x}_t)\|}\}$ we have:

$$\begin{aligned}\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 &\leq \|\mathbf{x}_t - \mathbf{x}^\star - \eta \alpha_t \nabla f(\mathbf{x}_t)\|^2 = \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \alpha_t^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\alpha_t \eta \langle \nabla f(\mathbf{x}_t), \mathbf{x}_t - \mathbf{x}^\star \rangle \\ &\leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \alpha_t^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\eta \alpha_t (f(\mathbf{x}_t) - f^\star) .\end{aligned}$$

We consider two cases: when clipping happens, and when clipping does not happen.

Case 1: $\alpha_t = 1$, meaning that $\|\nabla f(\mathbf{x}_t)\| \leq c$. Then

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 \|\nabla f(\mathbf{x}_t)\|^2 - 2\eta (f(\mathbf{x}_t) - f^\star) .$$

Using the implication of (L_0, L_1) smoothness and convexity in Lemma A.2,

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2(L_0 + L_1 \|\nabla f(\mathbf{x}_t)\|) (f(\mathbf{x}_t) - f^\star) \leq 2(L_0 + L_1 c) (f(\mathbf{x}_t) - f^\star) .$$

Further,

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + 2(L_0 + L_1 c)\eta^2 (f(\mathbf{x}_t) - f^\star) - 2\eta (f(\mathbf{x}_t) - f^\star) ,$$

and by setting $\eta \leq \frac{1}{2(L_0 + L_1 c)}$ we obtain

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \eta (f(\mathbf{x}_t) - f^\star) .$$

Case 2: $\alpha_t = \frac{c}{\|\nabla f(\mathbf{x}_t)\|}$, meaning that $\|\nabla f(\mathbf{x}_t)\| > c$. Then,

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \eta^2 c^2 - 2\eta \frac{c}{\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star) .$$

If it holds that $\eta^2 c^2 \leq \eta \frac{c}{\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star)$, then we will get

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \eta \frac{c}{\sqrt{2L}} \sqrt{(f(\mathbf{x}_t) - f^\star)}\,. \tag{17}$$

Lets now see under which stepsizes the condition $\eta \leq \frac{1}{c\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star)$ holds by upper bounding the rhs. By (L_0, L_1) smoothness (and Lemma A.2) we know that $(f(\mathbf{x}_t) - f^\star) \geq \frac{\|\nabla f(\mathbf{x}_t)\|^2}{2(L_0 + L_1\|\nabla f(\mathbf{x}_t)\|)}$ and thus

$$\frac{1}{c\,\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star) \geq \frac{1}{2(L_0 \frac{c}{\|\nabla f(\mathbf{x}_t)\|} + L_1 c)} \geq \frac{1}{2(L_0 + L_1 c)}\,,$$

where the last inequality is because $\frac{c}{\|\nabla f(\mathbf{x}_t)\|} \leq 1$ by our assumptions on α_t in this case. This means that using stepsize $\eta \leq \frac{1}{2(L_0 + L_1 c)}$, it will hold that $\eta \leq \frac{1}{c\|\nabla f(\mathbf{x}_t)\|} (f(\mathbf{x}_t) - f^\star)$ and thus (17) will hold.

Summing the two cases. We define $\mathcal{T}_1$ the set of iterations when clipping does not happen and $\mathcal{T}_2$ as set of iterations when clipping happens. Taking the average over $T+1$ iterations

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_1} (f(\mathbf{x}_t) - f^\star) + \frac{1}{T+1} \sum_{t \in \mathcal{T}_2} \frac{c}{\sqrt{2L}} \sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}\,.$$

This means that both (i)

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_1} (f(\mathbf{x}_t) - f^\star) \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}\,,$$

and (ii)

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_2} \sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \sqrt{2L}}{\eta c(T+1)}\,.$$

For the first inequality (i) using that $x^2 \geq 2\epsilon x - \epsilon^2$ for any $\epsilon, x > 0$, and defining for simplicity $A := \frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}$ we get

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_1} \left(2\epsilon \sqrt{f(\mathbf{x}_t) - f^\star} - \epsilon^2\right) \leq A\,,$$

and thus,

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_1} \sqrt{f(\mathbf{x}_t) - f^\star} \leq \frac{A}{2\epsilon} + \frac{\epsilon}{2}\,.$$

Choosing $\epsilon = \sqrt{A}$, we get

$$\frac{1}{T+1} \sum_{t \in \mathcal{T}_1} \sqrt{f(\mathbf{x}_t) - f^\star} \leq \sqrt{A} \leq \sqrt{\frac{\|\mathbf{x}_0 - \mathbf{x}^\star\|^2}{\eta(T+1)}}\,.$$

This implies that

$$\frac{1}{T+1} \sum_{t=0}^{T} \sqrt{f(\mathbf{x}_t) - f^\star} \leq \sqrt{\frac{R_0^2}{\eta(T+1)}} + \frac{R_0^2 \sqrt{2L}}{\eta c(T+1)}\,.$$

We further use that $f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t)$ and get a last-iterate convergence rate

$$\sqrt{f(\mathbf{x}_T) - f^\star} \leq \sqrt{\frac{R_0^2}{\eta(T+1)}} + \frac{R_0^2 \sqrt{2L}}{\eta c(T+1)}\,.$$

Squaring both of the sides, and using that $(a+b)^2 \leq 2a^2 + 2b^2\ \forall a, b$, we get

$$f(\mathbf{x}_T) - f^\star \leq \frac{2R_0^2}{\eta(T+1)} + \frac{4LR_0^4}{\eta^2 c^2 (T+1)^2}\,.$$

B.2. Strongly convex case (Theorem 2.5)

Recursive argument. First, since the strongly convex function is also convex, we can apply the result of the previous theorem here to get

$$f(\mathbf{x}_T) - f^\star \leq \frac{2R_0^2}{\eta T} + \frac{4LR_0^4}{\eta^2 c^2 T^2}\,.$$

We remind that $R_0 = \|\mathbf{x}_0 - \mathbf{x}^\star\|$. Using strong-convexity, we also know that

$$f(\mathbf{x}_T) - f^\star \geq \frac{\mu}{2} R_t^2\,,$$

Thus,

$$R_t^2 \leq \frac{4R_0^2}{\mu\eta T} + \frac{8LR_0^4}{\mu\eta^2 c^2 T^2}\,.$$

Thus, to get $R_t^2 \leq \frac{R_0^2}{2}$, it is enough to take $t \geq \max\{\frac{16}{\mu\eta}, \frac{6R_0\sqrt{L}}{\eta c\sqrt{\mu}}\}$ (as both terms become less that $R_0^2/4$).

Repeating this argument, we can see the iteration complexity can be bounded by

$$T = \mathcal{O}\left(\frac{1}{\mu\eta}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{R_0\sqrt{L}}{\eta c\sqrt{\mu}}\right)\,.$$

Small gradients. Let us start again from the convex bound (Theorem 2.3). Now, we will instead use that:

$$\|\nabla f(\mathbf{x}_t)\|^2 \leq 2L\left(f(\mathbf{x}_t) - f^\star\right) \leq 2L\frac{R_0^2}{\eta t}\left(1 + \frac{R_0^2 L}{c^2\eta t}\right)\,.$$

Now introduce t_0 which is such that:

$$t_0 = \frac{8LR_0^2}{\eta c^2}, \tag{18}$$

then we have that for all $t \geq t_0$:

$$\|\nabla f(\mathbf{x}_t)\| \leq c. \tag{19}$$

In particular, we know that no clipping happens after t_0, and so we obtain the standard linear convergence rate, so that the final convergence time is:

$$T = O\left(\frac{1}{\eta\mu}\log\left(\frac{R_0^2}{\epsilon}\right) + \frac{LR_0^2}{\eta c^2}\right)\,. \tag{20}$$

Comparing the two rates. Note that no rate is better than the other, and we can use one or the other depending on the relationship between c and $\sqrt{L\mu}R_0$.

C. Stochastic proofs.

We now proceed to the proof of Theorem 3.3. The proof will be in two parts: we will first prove convergence up to σ^2, and then refine this for large values of c.

C.1. Preliminaries

We now state a very simple lemma, which is direct but at the core of our decomposition, and so we highlight it here.

Lemma C.1. *For any $\alpha > 0$ and $\mathbf{u} \in \mathbb{R}^d$, the following holds:*

$$-\nabla f(\mathbf{x})^\top \mathbf{u} = -\frac{\alpha}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2\alpha}\|\mathbf{u}\|^2 + \frac{1}{2\alpha}\|\mathbf{u} - \alpha\nabla f(\mathbf{x})\|^2\,. \tag{21}$$

C.2. First part of the proof: convergence up to σ (small c)

In this section for simplicity we assume that $c < 4\sigma$ and prove that the gradient norm converges up to a level σ. Note that this assumption on c is not restrictive since the case $c > 4\sigma$ is covered by the other part of the proof, in which we show better convergence to $\mathcal{O}(\frac{\sigma^2}{c})$.

Large gradients. Let us start by assuming that $\|\nabla f(\mathbf{x}_t)\| \geq 6\sigma$. Note that numerical constant is (relatively) arbitrary and could be tightened, but we choose it high to keep the proof clean and simple.

We start the analysis by using (L_0, L_1) smoothness property from Lemma A.1. Note that for any stepsize $\eta < \frac{1}{L_0 + cL_1}$ it holds that $\|\mathbf{x}_{t+1} - \mathbf{x}_t\| = \eta\|\mathbf{g}(\mathbf{x}_t)\| \leq \eta c \leq \frac{1}{L_1}$

$$
\begin{aligned}
f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2} \|\mathbf{g}(\mathbf{x}_t)\|^2 \\
&\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2} c^2 \\
&\leq -\eta \nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2 (L_0 + cL_1)}{2} c \|\nabla f(\mathbf{x}_t)\| \,, \qquad (22)
\end{aligned}
$$

where the last inequality is because we assumed that $c \leq 4\sigma \leq \|\nabla f(\mathbf{x}_t)\|$.

Uniformly bounded variance, def. 1.3. In this case, let us first assume that strong variance holds with constant 3, *i.e.*, that $\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| \leq 3\sigma$ with probability one. In this case, we can write, where $\alpha_\xi = \min(1, c/\|\nabla f_\xi(\mathbf{x}_t)\|)$:

$$
\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) &= -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 - \alpha_\xi \nabla f(\mathbf{x}_t)^\top (\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)) \\
&\leq -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 + \alpha_\xi \|\nabla f(\mathbf{x}_t)\| \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| \\
&\leq -\alpha_\xi \|\nabla f(\mathbf{x}_t)\|^2 + 3\alpha_\xi \|\nabla f(\mathbf{x}_t)\| \sigma \\
&\leq -\frac{\alpha_\xi}{2} \|\nabla f(\mathbf{x}_t)\|^2 ,
\end{aligned}
$$

where the last line follows from the fact that $\sigma < \|\nabla f(\mathbf{x}_t)\|/6$. In particular, using the strong variance assumption, we know that $\|\nabla f_\xi(\mathbf{x}_t)\| \leq 2\|\nabla f(\mathbf{x}_t)\|$, so that $\alpha_\xi \geq \min(1, c/(2\|\nabla f(\mathbf{x}_t)\|)) \geq c/(2\|\nabla f(\mathbf{x}_t)\|)$. In particular:

$$
-\nabla f(\mathbf{x}_t)^\top \nabla f_\xi(\mathbf{x}_t) \leq -\frac{c}{4} \|\nabla f(\mathbf{x}_t)\| . \qquad (23)
$$

Then, we can plug this into Equation (22), which leads to:

$$
\mathbb{E}[f(\mathbf{x}_{t+1})] - f(\mathbf{x}_t) \leq -\frac{\eta c}{4} (1 - 2\eta(L_0 + cL_1)) \|\nabla f(\mathbf{x}_t)\| . \qquad (24)
$$

In particular, choosing $\eta \leq (4[L_0 + cL_1])^{-1}$, we obtain:

$$
\frac{\eta c}{8} \|\nabla f(\mathbf{x}_t)\| \leq f(\mathbf{x}_t) - f(\mathbf{x}_{t+1}). \qquad (25)
$$

Bounded variance in expectation, Def 1.4. In this case, we cannot write the same inequalities as before with probability 1. However, we can still guarantee the bound with large enough probability. We define $\delta = \mathbb{1}\{\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\| > 3\sigma\}$. We will use conditional expectations to write

$$
\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x})\right] \leq p(\delta = 0) \underbrace{\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 0\right]}_{:= T_1} + p(\delta = 1) \underbrace{\mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 1\right]}_{:= T_2} .
$$

We bound the first term T_1 the same way as in previous case of uniformly bounded noise. For the second term, by Cauchy-Schwartz inequality, and defining $\alpha = \min(1, c/\|\nabla f(\mathbf{x})\|)$ we write

$$
T_2 = \mathbb{E}\left[-\alpha_\xi \nabla f(\mathbf{x})^\top \nabla f_\xi(\mathbf{x}) | \delta = 1\right] \leq \|\nabla f(\mathbf{x})\| \, \mathbb{E}\left[\|\alpha_\xi \nabla f_\xi(\mathbf{x})\| \, | \delta = 1\right] \leq \alpha \|\nabla f(\mathbf{x})\|^2 ,
$$

where the last inequality is because we assumed that the full gradients are large $\|\nabla f(\mathbf{x})\| > 6\sigma$, but the clipping threshold is small $c \leq 4\sigma$. Thus, the full gradients would always get clipped, and $\|\alpha \nabla f(\mathbf{x})\| = c \geq \|\alpha_\xi \nabla f_\xi(\mathbf{x})\|$. We remind that $\alpha_\xi = \min(1, c/\|\nabla f_\xi(\mathbf{x})\|)$.

Now, it just remains to bound $p(\delta = 1)$. Using Markov inequality, we have that:

$$p(\delta = 1) = p(\|\nabla f_\xi(\mathbf{x}) - \nabla f(\mathbf{x})\|^2 > 9\sigma^2) \leq 1/9. \tag{26}$$

Similarly, $p(\delta = 0) = 1 - p(\delta = 1) \geq 8/9$. In the end, we obtain that:

$$-\mathbb{E}\left[\nabla f(\mathbf{x})^\top \mathbf{g}(\mathbf{x})\right] \leq -c\left(\frac{1}{4} \times \frac{8}{9} - \frac{1}{9}\right)\|\nabla f(\mathbf{x})\| = -\frac{c}{9}\|\nabla f(\mathbf{x})\| . \tag{27}$$

We further plug the result into (22), and obtain

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_t) \leq -\frac{\eta c}{9}\left(1 - \frac{9\eta}{2}(L_0 + cL_1)\right)\|\nabla f(\mathbf{x}_t)\|, \tag{28}$$

and so with $\eta \leq (9[L_0 + cL_1])^{-1}$, we obtain:

$$\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_t) \leq -\frac{\eta c}{18}\|\nabla f(\mathbf{x}_t)\|. \tag{29}$$

Final convergence. If for at least one iteration t it happens that the gradient norm is small $\|\nabla f(\mathbf{x}_t)\| \leq 6\sigma$, then it simply holds that

$$\min_{t\in[1,T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq O\left(\sigma^2\right).$$

Otherwise, for all t iterations the gradient norm is large $\|\nabla f(\mathbf{x}_t)\| > 6\sigma$ and thus (29) holds for all the iterations. Averaging over $1 \leq t \leq T+1$, we obtain

$$\frac{1}{T+1}\sum_{t=0}^{T}\|\nabla f(\mathbf{x}_t)\| \leq \mathcal{O}\left(\frac{f(\mathbf{x}_0) - f^\star}{\eta c T}\right), \tag{30}$$

Combining these two cases we conclude that

$$\min_{t\in[1,T]} \mathbb{E}\|\nabla f(\mathbf{x}_t)\|^2 \leq \mathcal{O}\left(\sigma^2 + \frac{f(\mathbf{x}_0) - f^\star}{\eta c T}\right),$$

C.3. Second part of the proof: convergence up to σ^2/c (large c).

In this second part we assume that the clipping radius is large, $c \geq 4\sigma$. Although, the algorithm (2) clips the stochastic gradients $\nabla f_\xi(\mathbf{x}_t)$, for the proof we will consider the two cases based on the full gradient $\nabla f(\mathbf{x}_t)$: when the full gradient $\nabla f(\mathbf{x}_t)$ is clipped and when it is not clipped.

Similarly to previous case, we start by using (L_0, L_1) smoothness

$$f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\eta\nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2}\|\mathbf{g}(\mathbf{x}_t)\|^2 . \tag{31}$$

First case, full gradient is clipped $\|\nabla f(\mathbf{x}_t)\| > c$.

In this case, we use (21) with $\alpha = \frac{c}{\|\nabla f(\mathbf{x}_t)\|}$ and $\mathbf{u} = \mathbf{g}(\mathbf{x}_t)$. Since $\alpha\nabla f(\mathbf{x}_t) = \mathrm{clip}_c(\nabla f(\mathbf{x}_t))$, this leads to

$$-\nabla f(\mathbf{x}_t)^\top \mathbf{g}(\mathbf{x}_t) = -\frac{c}{2}\|\nabla f(\mathbf{x}_t)\| - \frac{1}{2\alpha}\|\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{1}{2\alpha}\|\mathbf{g}(\mathbf{x}_t) - \mathrm{clip}_c(\nabla f(\mathbf{x}_t))\|^2 . \tag{32}$$

We now use that $\mathbf{g}(\mathbf{x}_t) = \mathrm{clip}_c(\nabla f_\xi(\mathbf{x}_t))$, and use that clipping is a projection on onto a convex set (ball of radius c), and thus is Lipshitz operator with Lipshitz constant 1, we write

$$\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top \mathbb{E}\mathbf{g}(\mathbf{x}_t) &\leq -\frac{c}{2}\|\nabla f(\mathbf{x}_t)\| - \frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right] + \frac{1}{2\alpha}\mathbb{E}\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2 \\
&\leq -\frac{c}{2}\|\nabla f(\mathbf{x}_t)\| - \frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right] + \frac{\sigma^2}{2c}\|\nabla f(\mathbf{x}_t)\| \\
&= -\frac{1}{2\alpha}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right] - \frac{c}{2}\|\nabla f(\mathbf{x}_t)\|\left(1 - \frac{\sigma^2}{c^2}\right) \\
&\leq -\frac{\|\nabla f(\mathbf{x}_t)\|}{2c}\mathbb{E}\left[\|\mathbf{g}(\mathbf{x}_t)\|^2\right] - \frac{c}{4}\|\nabla f(\mathbf{x}_t)\|,
\end{aligned}$$

where in the last line we used that $\sigma^2/c^2 \leq 1/2$ and $\alpha \leq 1$. Plugging this into (31) we get

$$\begin{aligned}
\mathbb{E}\left[f(\mathbf{x}_{t+1})\right] - f(\mathbf{x}_t) &\leq -\frac{\eta \left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] - \frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\eta^2(L_0 + \left\|\nabla f(\mathbf{x}_t)\right\| L_1)}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\
&= -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta \left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta c L_1) + \frac{\eta^2 L_0}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\
&\leq -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta c L_1) + \frac{\eta^2 L_0}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\
&= -\frac{\eta c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right](1 - \eta[L_0 + c L_1])\,.
\end{aligned}$$

In particular, choosing $\eta \leq (L_0 + cL_1)^{-1}$, we obtain:

$$\frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{f(\mathbf{x}_t) - \mathbb{E} f(\mathbf{x}_{t+1})}{\eta}\,. \tag{33}$$

Note that we do not obtain variance terms, but similarly to the previous section it is because we have assumed that the norm of the gradient is larger than σ, then the noise term can be hidden in the gradient norm term.

Second case, $c > \left\|\nabla f(\mathbf{x}_t)\right\| > c/2$**.** The proof follows very closely the previous case with the difference that the full gradient $\nabla f(\mathbf{x}_t)$ is not clipped. We use Equation (21) with $\alpha = 1$. This leads to

$$\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top \mathbb{E}\mathbf{g}(\mathbf{x}_t) &= -\frac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\|^2 \\
&\leq -\frac{1}{2}\left\|\nabla f(\mathbf{x}_t)\right\|^2 - \frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{\sigma^2}{2}\,,
\end{aligned}$$

where on the last line we used that clipping is Lipshitz operator with constant 1, as it is a projection on a convex set. We now use that $-\left\|\nabla f(\mathbf{x}_t)\right\| \leq -c/2$ for the first term and $1 \leq \left\|\nabla f(\mathbf{x}_t)\right\|/c$ for the last term:

$$\begin{aligned}
-\nabla f(\mathbf{x}_t)^\top \mathbb{E}\mathbf{g}(\mathbf{x}_t) &\leq -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\sigma^2}{2c}\left\|\nabla f(\mathbf{x}_t)\right\| \\
&\leq -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{4}\left\|\nabla f(\mathbf{x}_t)\right\|\left(1 - 2\frac{\sigma^2}{c^2}\right) \\
&\leq -\frac{1}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 - \frac{c}{8}\left\|\nabla f(\mathbf{x}_t)\right\|,
\end{aligned}$$

where in the last line we used that $\sigma^2/c^2 \leq 1/4$. Similarly to the previous case, we plug it into (31) and use that $-1 \leq -\frac{\left\|\nabla f(\mathbf{x}_t)\right\|}{c}$ we have that

$$\begin{aligned}
\mathbb{E} f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\frac{\eta \left\|\nabla f(\mathbf{x}_t)\right\|}{2c}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] - \frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| + \frac{\eta^2(L_0 + \left\|\nabla f(\mathbf{x}_t)\right\| L_1)}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right] \\
&\leq -\frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right]\left(\frac{\left\|\nabla f(\mathbf{x}_t)\right\|}{c}(1 - \eta c L_1) - \eta L_0\right) \\
&\leq -\frac{c\eta}{8}\left\|\nabla f(\mathbf{x}_t)\right\| - \frac{\eta}{2}\mathbb{E}\left[\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2\right]\left(\frac{1}{2} - \eta[L_0 + cL_1]\right)
\end{aligned}$$

where on the last line we used that $\left\|\nabla f(\mathbf{x}_t)\right\| > c/2$. Using that $\eta \leq \frac{1}{2}(L_0 + cL_1)^{-1}$

$$\frac{c}{8}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{f(\mathbf{x}_t) - \mathbb{E} f(\mathbf{x}_{t+1})}{\eta}\,. \tag{34}$$

Third case, $\left\|\nabla f(\mathbf{x}_t)\right\| < c/2$**.** In this case, we do not have convergence to the exact optimum.

We start by defining $\delta_t = \mathbb{1}\{\left\|\nabla f_\xi(\mathbf{x}_t)\right\| > c\}$ is the indicator function that at time step t the stochastic gradient is getting clipped. We will start by showing that $\mathbb{E}\delta_t \leq \frac{4\sigma^2}{c^2}$.

$$\mathbb{E}\delta_t = \Pr[\delta_t = 1] = \Pr\left[\left\|\nabla f_\xi(\mathbf{x}_t)\right\| > c\right] \leq \Pr\left[\left\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\right\| > \frac{c}{2}\right] \leq \frac{4\sigma^2}{c^2}\,,$$

where the last inequality is due to Markov's inequality. The first inequality is because $\|\nabla f_\xi(\mathbf{x}_t)\| \leq \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| + \|\nabla f(\mathbf{x}_t)\| \leq \|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\| + \frac{c}{2}$.

Now that we have $\mathbb{E}\delta_t \leq \frac{4\sigma^2}{c^2}$ we can use it to bound the difference $\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2$. In particular, since δ_t takes values 0 and 1, we have that $\mathbb{E}[\delta_t] = p(\delta_t = 1)$ and so $\mathbb{E}[\delta_t X] = \mathbb{E}[\delta_t]\,\mathbb{E}[X|\delta_t]$ for any random variable X.

$$\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 = \left\|\mathbb{E}\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)\nabla f_\xi(\mathbf{x}_t)\delta_t\right\|^2 = \mathbb{E}[\delta_t]^2\left\|\mathbb{E}\left[\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)\nabla f_\xi(\mathbf{x}_t)|\delta_t = 1\right]\right\|^2.$$

At this point, we use Jensen inequality on the conditional expectation (since all terms are positive and the squared norm is a convex function) and get that:

$$\begin{aligned}
\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 &\leq \mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)^2\|\nabla f_\xi(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\leq \mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\leq 2\mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2|\delta_t = 1\right] + 2\mathbb{E}[\delta_t]^2\,\mathbb{E}\left[\|\nabla f(\mathbf{x}_t)\|^2|\delta_t = 1\right] \\
&\leq 2\mathbb{E}[\delta_t]\,\mathbb{E}\left[\|\nabla f_\xi(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2\right] + 2\mathbb{E}[\delta_t]^2\|\nabla f(\mathbf{x}_t)\|^2 \\
&\leq \frac{8\sigma^4}{c^2} + \frac{32\sigma^4}{c^4}\|\nabla f(\mathbf{x}_t)\|^2, \qquad (35)
\end{aligned}$$

where on the second line we used that $\left(1 - \frac{c}{\|\nabla f_\xi(\mathbf{x}_t)\|}\right)^2 \leq 1$ when $\delta_t = 1$, and on the last line that $\mathbb{E}[\delta_t] \leq 4\sigma^2/c^2$. We further use (21) with $\alpha = 1$ and $\mathbf{u} = \mathbb{E}\mathbf{g}(\mathbf{x}_t)$, we get

$$\begin{aligned}
-\nabla f(\mathbf{x})^\top\mathbb{E}\mathbf{g}(\mathbf{x}_t) &= -\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x})\|^2 \\
&\leq -\frac{1}{2}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\sigma^4}{c^2} + \frac{4\sigma^2}{c^2}\|\nabla f(\mathbf{x}_t)\|^2 \\
&\overset{\sigma\leq\frac{c}{4}}{\leq} -\frac{1}{4}\|\nabla f(\mathbf{x})\|^2 - \frac{1}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\sigma^4}{c^2}.
\end{aligned}$$

Plugging this into (31), for $\eta \leq \frac{1}{8(L_0+cL_1)}$, we get by dropping the $\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2$ term and using that $\|\nabla f(\mathbf{x}_t)\| \leq c$ that:

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 - \frac{\eta}{2}\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\|L_1)}{2}\mathbb{E}\|\mathbf{g}(\mathbf{x}_t)\|^2 \qquad (36) \\
&\leq -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0 + cL_1)}{2}\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)\|^2 \\
&\leq -\frac{\eta}{4}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2 + \eta^2(L_0 + cL_1)\|\nabla f(\mathbf{x}_t)\|^2 \\
&\overset{\eta\leq\frac{1}{8(L_0+cL_1)}}{\leq} -\frac{\eta}{8}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\mathbb{E}\|\mathbf{g}(\mathbf{x}_t) - \nabla f(\mathbf{x}_t)\|^2.
\end{aligned}$$

Note that clipping is the orthogonal projection onto the ball of radius c, which we denote proj_c and $\|\nabla f(x_t)\| \leq c$, so it is not affected by the projection. In particular:

$$\mathbb{E}\|g(x_t) - \nabla f(x_t)\|^2 = \mathbb{E}\|\mathrm{proj}_c(\nabla f_\xi(x_t)) - \mathrm{proj}_c(\nabla f(x_t))\|^2 \leq \mathbb{E}\|\nabla f_\xi(x_t) - \nabla f(x_t)\|^2 \leq \sigma^2, \qquad (37)$$

and we thus get

$$\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{\eta}{8}\|\nabla f(\mathbf{x})\|^2 + \frac{4\eta\sigma^4}{c^2} + \eta^2(L_0 + cL_1)\sigma^2, \qquad (38)$$

and so:

$$\frac{1}{8}\|\nabla f(\mathbf{x}_t)\|^2 \leq \frac{f(\mathbf{x}_t) - \mathbb{E}f(\mathbf{x}_{t+1})}{\eta} + \eta(L_0 + cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}. \qquad (39)$$

In particular, we have:

- One variance term that fades with the step-size.
- One bias term that remains even for very small step-sizes.

Wrapping up. We now combine the three cases above. Defining $\mathcal{T}_1$ is the set of indices with $\|\nabla f(\mathbf{x}_t)\| \geq \frac{c}{2}$ (we note that this covers the first and the second cases from above, but both of them leads to the same final inequality (34)), and $\mathcal{T}_2$ is the set of indices with $\|\nabla f(\mathbf{x}_t)\| < \frac{c}{2}$, this inequality (39) holds. Summing up over all the indices $1 \leq t \leq T+1$, we get

$$\frac{1}{8(T+1)}\left(\sum_{t\in\mathcal{T}_1} c\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| + \sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2\right) \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,.$$

This means that both (i)

$$\frac{1}{8(T+1)}\sum_{t\in\mathcal{T}_1} c\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,,$$

and (ii)

$$\frac{1}{8(T+1)}\sum_{t\in\mathcal{T}_2} \mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\|^2 \leq \frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)} + \eta(L_0+cL_1)\sigma^2 + \frac{4\sigma^4}{c^2}\,,$$

for the last inequality using that $x^2 \geq 2\epsilon x - \epsilon^2$ for any $\epsilon, x > 0$, and defining for simplicity $A := 8\frac{f(\mathbf{x}_0)-f^\star}{\eta T} + 8\eta(L_0 + cL_1)\sigma^2 + \frac{32\sigma^4}{c^2}$ we get

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\left(2\epsilon\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| - \epsilon^2\right) \leq A\,,$$

and thus,

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \frac{A}{2\epsilon} + \frac{\epsilon}{2}\,.$$

Choosing $\epsilon = \sqrt{A}$, we get

$$\frac{1}{T+1}\sum_{t\in\mathcal{T}_2}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \sqrt{A} \leq \sqrt{8\frac{f(\mathbf{x}_0)-f^\star}{\eta(T+1)}} + \sqrt{8\eta(L_0+cL_1)\sigma^2} + \sqrt{\frac{32\sigma^4}{c^2}}\,.$$

Summing up the two cases again, and using that $\frac{\sigma}{c} \leq \frac{1}{4}$ we get

$$\frac{1}{T+1}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \mathcal{O}\left(\sqrt{\frac{f(\mathbf{x}_0)-f^\star}{\eta T}} + \frac{f(\mathbf{x}_0)-f^\star}{\eta c T} + \sqrt{\eta(L_0+cL_1)}\sigma + \frac{\sigma^2}{c}\right)\,.$$

C.4. Differentially Private SGD

C.4.1. Modification to the proof to include mini-batches

Using $\mathbf{g}(\mathbf{x}_t) = \frac{1}{B}\sum_{\xi\in\mathcal{B}_t}\mathrm{clip}_c(\nabla f_\xi(\mathbf{x}_t))$, the proof is exactly the same as in the previous case, with the only difference in the case where $c \geq 4\sigma$ and small gradients (third case) $\|\nabla f(\mathbf{x}_t)\| < \frac{c}{2}$. Starting with equation (36), we obtain:

$$\begin{aligned}
\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) &\leq -\frac{\eta}{4}\left\|\nabla f(\mathbf{x})\right\|^2 - \frac{\eta}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0+\|\nabla f(\mathbf{x}_t)\|L_1)}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\leq -\frac{\eta}{4}\left\|\nabla f(\mathbf{x})\right\|^2 - \frac{\eta}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{4\eta\sigma^4}{c^2} + \frac{\eta^2(L_0+cL_1)}{2}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 \\
&\quad + \frac{\eta^2(L_0+cL_1)}{2}\left\|\mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2\,.
\end{aligned}$$

We now estimate the term variance term $\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2$ more tightly in order to get the variance reduction due to the batch size B.

$$\begin{aligned}\mathbb{E}\left\|\mathbf{g}(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 &= \mathbb{E}\left\|\frac{1}{B}\sum_{i\in\mathcal{B}_t}\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2 = \frac{1}{B^2}\sum_{i\in\mathcal{B}_t}\mathbb{E}\left\|\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2\\ &\leq \frac{1}{B^2}\sum_{i\in\mathcal{B}_t} 2\mathbb{E}\left\|\mathrm{clip}_c(\nabla f_{\xi_i}(\mathbf{x}_t)) - \nabla f(\mathbf{x}_t)\right\|^2 + \frac{2}{B}\left\|\nabla f(\mathbf{x}_t) - \mathbb{E}\mathbf{g}(\mathbf{x}_t)\right\|^2\\ &\overset{(35)}{\leq} \frac{2\sigma^2}{B} + \frac{2}{B}\left[\frac{8\sigma^4}{c^2} + \frac{8\sigma^2}{c^2}\left\|\nabla f(\mathbf{x}_t)\right\|^2\right]\\ &\leq \frac{2\sigma^2}{B} + \frac{2}{B}\left[\frac{\sigma^2}{2} + 2\sigma^2\right] \leq 6\frac{\sigma^2}{B},\end{aligned}$$

where we used that $\|\nabla f(\mathbf{x}_t)\| \leq \frac{c}{2}$ and that $\sigma \leq \frac{c}{4}$. The rest of the proof is exactly the same as before, by substituting now the σ^2 term with $\frac{\sigma^2}{B}$, we would arrive at the convergence rate of

$$\frac{1}{T}\sum_{t=0}^{T}\mathbb{E}\left\|\nabla f(\mathbf{x}_t)\right\| \leq \mathcal{O}\left(\sqrt{\frac{f(\mathbf{x}_0) - f^\star}{\eta T}} + \frac{f(\mathbf{x}_0) - f^\star}{\eta c T} + \sqrt{\eta(L_0 + cL_1)}\frac{\sigma}{\sqrt{B}} + \frac{\sigma^2}{c}\right).$$

C.4.2. Modification to the proof to include stochastic noises

The gradients applied in DP-SGD (15) have the form $\mathbf{g}(\mathbf{x}_t) + \mathbf{z}_t$, where $\mathbf{z}_t$ is a Gaussian noise with variance σ_{DP}. In order to add this additional Gaussian noise, we would need to modify the first step of the proof, that is using (L_0, L_1) smoothness

$$\mathbb{E}f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\eta\nabla f(\mathbf{x}_t)^\top\mathbf{g}(\mathbf{x}_t) + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2}\left\|\mathbf{g}(\mathbf{x}_t)\right\|^2 + \frac{\eta^2(L_0 + \|\nabla f(\mathbf{x}_t)\| L_1)}{2}\sigma_{\text{DP}}^2.$$

The rest of the proof remains the same, with having an additional σ_{DP}^2 term in the convergence. We thus would arrive to the following convergence rate where for simplicity we define $L = L_0 + \max_t \|\nabla f(\mathbf{x}_t)\| L_1$

$$O\left(\frac{L\eta}{c}\sigma_{\text{DP}}^2 + \sqrt{L\eta\sigma_{\text{DP}}} + \min\left(\sigma, \frac{\sigma^2}{c}\right) + \sqrt{\eta L}\frac{\sigma}{\sqrt{B}} + \sqrt{\frac{F_0}{\eta T}} + \frac{F_0}{\eta T c}\right).$$

C.5. Lower bound

We now prove the lower bound.

Proofs of Theorems 3.1 and 3.2. Let us consider the simple noise $a\mathcal{B}(p)$, where $a > 0$ and $\mathcal{B}(p)$ is a Bernoulli random variable with mean $p \leq 1/2$. Consider a function such that the stochastic gradients are of the form:

$$\nabla f_\xi(x) = x + a\mathcal{B}(p). \tag{40}$$

Now consider $x = -pc/(1-p)$. The stochastic gradient at x when the Bernoulli is 0 is not clipped, since $|x| = pc/(1-p) \leq c$. Yet, the stochastic gradient for positive values of the Bernoulli random variable is:

$$\nabla f_a(x) = -pc/(1-p) + a \geq a - c \geq c. \tag{41}$$

In particular, we have that:

$$\mathbb{E}\left[\mathrm{clip}_c(\nabla f_\xi(x))\right] = (1-p)x + pc = (1-p)\times(-pc)/(1-p) + pc = 0. \tag{42}$$

Let us now evaluate $\nabla f(x)$. We have:

$$\nabla f(x) = x + pa = p\left(a - \frac{c}{1-p}\right). \tag{43}$$

Small c. Now fix a clipping radius c, such that $c \leq 2\sigma$, and take $a = 4\sigma$. We choose $p(1-p) = 1/16$, so that $p = (2-\sqrt{3})/4 \leq 1/4$. In this case,

$$\nabla f(x) = p\left(a - \frac{c}{1-p}\right) \geq p\left(4\sigma - 2\sigma \times \frac{4}{3}\right) \geq \frac{(2-\sqrt{3})\sigma}{3} \geq \frac{\sigma}{12}. \tag{44}$$

Large c. Now fix a clipping radius c, such that $c \geq \sigma$ and $c \leq a/2$. To ensure that the noise has variance σ^2, p has to be such that:

$$p(1-p) = \sigma^2/a^2 \leq 1/16\,. \tag{45}$$

Thus, we have that $p \leq 1/4$ (since we chose $p < 1/2$). In particular, also using that $c \leq a/2$:

$$\nabla f(x) = \frac{\sigma^2}{a^2(1-p)}\left(a - \frac{c}{1-p}\right) \geq \frac{\sigma^2}{3a(1-p)} \geq \frac{\sigma^2}{3a}\,. \tag{46}$$

It now remains to choose $a = 2c$ (which satisfies all previous conditions), and we obtain:

$$\nabla f(x) \geq \frac{\sigma^2}{6c}\,. \tag{47}$$

□