Stability and Convergence of Stochastic Gradient Clipping: Beyond Lipschitz Continuity and Smoothness

Vien V. Mai [1] **Mikael Johansson** [1]

Abstract

Stochastic gradient algorithms are often unstable when applied to functions that do not have Lipschitz-continuous and/or bounded gradients. Gradient clipping is a simple and effective technique to stabilize the training process for problems that are prone to the exploding gradient problem. Despite its widespread popularity, the convergence properties of the gradient clipping heuristic are poorly understood, especially for stochastic problems. This paper establishes both qualitative and quantitative convergence results of the clipped stochastic (sub)gradient method (SGD) for non-smooth convex functions with rapidly growing subgradients. Our analyses show that clipping enhances the stability of SGD and that the clipped SGD algorithm enjoys finite convergence rates in many cases. We also study the convergence of a clipped method with momentum, which includes clipped SGD as a special case, for weakly convex problems under standard assumptions. With a novel Lyapunov analysis, we show that the proposed method achieves the best-known rate for the considered class of problems, demonstrating the effectiveness of clipped methods also in this regime. Numerical results confirm our theoretical developments.

1. Introduction

We study stochastic optimization problems on the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\, f(x) := \mathbb{E}_P[f(x;S)] = \int_{\mathcal{S}} f(x;s)dP(s), \quad (1)$$

where $S \sim P$ is a random variable; $f(x;s)$ is the instantaneous loss parameterized by x on a sample $s \in \mathcal{S}$. Such problems are at the core of many machine-learning applications, and are often solved using stochastic (sub)gradient methods. In spite of their successes, stochastic gradient methods can be sensitive to their parameters (Nemirovski et al., 2009; Asi & Duchi, 2019a) and have severe instability (unboundedness) problems when applied to functions that grow faster than quadratically in the decision vector x (Andradóttir, 1996; Asi & Duchi, 2019a). Consequently, a careful (and sometimes time-consuming) parameter tuning is often required for these methods to perform well in practice. Even so, a good parameter selection is not sufficient to circumvent the instability issue on steep functions.

Gradient clipping and the closely related gradient normalization technique are simple modifications to the underlying algorithm to control the step length that an update can make relative to the current iterate. These techniques enhance the stability of the optimization process, while adding essentially no extra cost to the original update. As a result, gradient clipping has been a common choice in many applied domains of machine learning (Pascanu et al., 2013).

In this work, we consider gradient clipping applied to the classical SGD method. Throughout the paper, we frequently use the following clipping operator

$$\text{clip}_\gamma : \mathbb{R}^n \to \mathbb{R}^n : x \mapsto \min\left\{1, \frac{\gamma}{\|x\|_2}\right\} x,$$

which is nothing else but the orthogonal projection onto the γ-ball. It is important to note that for noiseless gradients, clipping just changes the magnitude and does not effect the search direction. However, in the stochastic setting, the expected value of the clipped stochastic gradient may point in a completely different direction than the true gradient.

Clipped SGD To solve problem (1), we use an iterative procedure that starts from $x_0 \in \mathbb{R}^n$ and $g_0 \in \partial f(x_0, S_0)$ and generates a sequence of points $x_k \in \mathbb{R}^n$ by repeating the following steps for $k = 0, 1, 2, \ldots$:

$$x_{k+1} = x_k - \alpha_k d_k, \quad d_k = \text{clip}_{\gamma_k}(g_k). \quad (2)$$

We refer to α_k as the kth stepsize and γ_k as the kth clipping threshold, while $g_k = f'(x_k, S_k)$ is the kth stochastic subgradient or its mini-batch version, $g_k = \frac{1}{m_k}\sum_{i=1}^{m_k} f'(x_k, S_k^i)$ if multiple samples are used in each iteration.

[1]Division of Decision and Control Systems, EECS, KTH Royal Institute of Technology, Stockholm, Sweden. Correspondence to: V. V. Mai <maivv@kth.se>, M. Johansson <mikaelj@kth.se>.

Proceedings of the 38^{th} International Conference on Machine Learning, PMLR 139, 2021.

1.1. Related work

Our work is closely related to a number of topics which we briefly review below.

Gradient clipping Gradient clipping and normalization were recognized early in the development of subgradient methods as a useful tool to obtain convergence for rapidly growing convex functions (Shor, 1985; Ermoliev, 1988; Alber et al., 1998). For the normalized method, the seminal work (Shor, 1985) establishes a convergence rate for the quantity $\langle g_k / \|g_k\|_2, x_k - x^\star \rangle$ without any assumptions on g_k. By only requiring that subgradients are bounded on bounded sets, which always holds for continuous functions, Alber et al. (1998) prove convergence in the objective value for the clipped subgradient method. The work (Hazan et al., 2015) considers normalized gradient methods for quasi-convex and locally smooth functions. Recently, the authors in (Zhang et al., 2019; 2020) analyze clipped methods for twice-differentiable functions satisfying a more relaxed condition than the traditional L-smoothness. However, much less is known in the stochastic setting. The work (Andradóttir, 1996) proposes a method that uses two independent samples in each iteration and proves its almost sure convergence under the same growth condition used in (Alber et al., 1998). Hazan et al. (2015) establish certain complexity results for a mini-batch stochastic normalized gradient method under some strong assumptions on the closeness of all the generated iterates to the optimal solution and the boundedness of all the individual mini-batch functions. In (Zhang et al., 2019; 2020), stochastic clipped SGD methods are analyzed under the assumption that the noise in the gradient estimate is bounded for every x almost surely. Such assumption does not hold even for Gaussian noise in the gradients. Finally, we refer to (Cutkosky & Mehta, 2020; Curtis et al., 2019; Gorbunov et al., 2020) for recent theoretical developments for clipped and normalized methods on standard L-smooth problems.

Robustness and stability The problems of robustness and stability in stochastic optimization have been emphasized in many studies (see, e.g., (Nemirovski et al., 2009; Andradóttir, 1996; Asi & Duchi, 2019a;b) and references therein). Much recent work on this topic concentrates around model-based algorithms that attempt to construct more accurate models of the objective than the linear one provided by the stochastic subgradient. When such models can be obtained and the resulting update steps can be performed efficiently, these methods often possess good stability properties and can be less sensitive to parameter selection than traditional stochastic subgradient methods. For example, the work (Asi & Duchi, 2019a) establishes almost sure convergence of stochastic (approximate) proximal point methods under the arbitrary growth condition used in (Alber et al., 1998; Andradóttir, 1996) and discussed above. In (Asi & Duchi, 2019b), almost sure convergence of the so-called truncated method is proven for convex functions that can grow polynomially from the the set of solutions.

Weakly convex minimization The class of weakly convex functions is broad, allowing for both non-smooth and non-convex objectives, and has favorable structures for algorithmic foundations and complexity theory. Earlier works on weakly convex minimization (Nurminskii, 1973; Ruszczyński, 1987; Ermol'ev & Norkin, 1998) establish qualitative convergence results for subgradient-based methods. With the recent advances in statistical learning and signal processing, there has been an emerging line of work on this topic (see, e.g., (Duchi & Ruan, 2018; Davis & Grimmer, 2019; Davis & Drusvyatskiy, 2019)). Convergence properties have been analyzed for many popular stochastic algorithms such as: model-based methods (Davis & Drusvyatskiy, 2019; Duchi & Ruan, 2018; Asi & Duchi, 2019b); momentum extensions (Mai & Johansson, 2020); adaptive methods (Alacaoglu et al., 2020); and more.

1.2. Contributions

The performance of stochastic (sub)gradient methods depends heavily on how rapidly the underlying function is allowed to grow. Much convergence theory for these methods hinges on the L-smoothness assumption for differentiable functions or uniformly bounded subgradients for non-smooth ones. These conditions restrict the corresponding convergence rates to functions with at most quadratic and linear growth, respectively. Beyond these well-behaved classes, there is abundant evidence that SGD and its relatives may fail to converge. It is our goal in this work to show, both theoretically and empirically, that the addition of the clipping step greatly improves the convergence properties of SGD. To that end, we make the following contributions:

- We establish stability and convergence guarantees for clipped SGD on convex problems with arbitrary growth (exponential, super-exponential, etc.) of the subgradients. We show that clipping coupled with standard mini-batching suffices to guarantee almost sure convergence. Even more, a finite convergence rate can also be obtained in this setting.

- We then turn to convex functions with polynomial growth and show that without the need for mini-batching, clipped SGD can essentially achieve the same optimal convergence rate as for stochastic strongly convex and Lipschitz continuous functions.

- We consider a momentum extension of clipped SGD for weakly convex minimization under standard growth conditions. With a carefully constructed Lyapunov function,

we are able to overcome the bias introduced by the clipping step and preserve the best-known sample complexity for this function class.

Our experiments on phase retrieval, absolute linear regression, and classification with neural networks reaffirm our theoretical findings that gradient clipping can: i) stabilize and guarantee convergence for problems with rapidly growing gradients; ii) retain and sometimes improve the best performance of their unclipped counterparts even on standard problems. We note also that none of the convergence results in this work require hard-to-estimate parameters to set the clipping threshold.

Notation For any $x, y \in \mathbb{R}^n$, we denote by $\langle x, y\rangle$ the Euclidean inner product of x and y. We denote by $\partial f(x)$ the Fréchet subdifferential of f at x; $f'(x)$ denotes any element of $\partial f(x)$. The ℓ_2-norm is denoted by $\|\cdot\|_2$. For a closed and convex set $\mathcal{X}$, the distance and the projection map are given respectively by: $\operatorname{dist}(x, \mathcal{X}) = \min_{z\in\mathcal{X}} \|z - x\|_2$ and $\Pi_{\mathcal{X}}(x) = \operatorname{argmin}_{z\in\mathcal{X}} \|z - x\|_2$. $\mathbf{1}\{E\}$ denotes the indicator function of an event E; i.e., $\mathbf{1}\{E\} = 1$ if E is true and 0 otherwise. The closed ℓ_2-ball centered at x with radius $r > 0$ is denoted $\mathrm{B}(x, r)$. We denote by $\mathcal{F}_k := \sigma(S_0, \ldots, S_{k-1})$ the σ-field formed by the first k random variables $S_0, \ldots, S_{k-1}$, so that $x_k \in \mathcal{F}_k$. Finally, we will impose the following basic assumption throughout the paper.

Assumption A1. *Let S be a sample drawn from P and $f'(x, S) \in \partial f(x, S)$, we have: $\mathbb{E}[f'(x, S)] \in \partial f(x)$.*

2. Stability and its consequences for convex minimization

In this section, we study the stability of the clipped SGD algorithm and its consequence for the minimization of (possibly non-smooth) convex functions. We first specify the assumptions needed for the results in this section starting with the basic quadratic growth condition.

Assumption A2 (Quadratic growth)**.** *There exists a scalar $\mu > 0$ such that*

$$f(x) - f(x^\star) \geq \mu \operatorname{dist}(x, \mathcal{X}^\star)^2, \quad \forall x \in \operatorname{dom}(f).$$

Assumption A2 gives a *lower bound* on the speed at which the objective f grows away from the solution set $\mathcal{X}^\star$. Since we are interested in problems that may exhibit exploding subgradients, this growth condition is a rather natural assumption. Note also that in many machine learning applications, the addition of a quadratic regularization term to improve generalization results in problems which fundamentally have quadratic growth.

Assumption A3 (Finite variance)**.** *There exists a scalar $\sigma > 0$ such that:*

$$\mathbb{E}\left[\|f'(x, S) - f'(x)\|_2^2\right] \leq \sigma^2, \quad \forall x \in \operatorname{dom}(f),$$

where $f'(x) = \mathbb{E}[f'(x, S)] \in \partial f(x)$.

Finally, unless otherwise stated, we assume that the stepsizes α_k are square summable but not summable:

$$\alpha_k \geq 0, \quad \sum_{i=0}^{\infty} \alpha_k = \infty, \quad \text{and} \quad \sum_{i=0}^{\infty} \alpha_k^2 < \infty.$$

Before detailing the stability and convergence analyses of clipped SGD, Example 1 shows that even with stepsizes that are as small as $O(1/k)$, the vanilla SGD method may fail miserably when applied to a function satisfying Assumptions A2–A3. We refer to (Asi & Duchi, 2019a) for more examples of the potential instability of SGD.

Example 1 (Super-Exponential Divergence of SGD)**:** Let $f(x) = x^4/4 + \epsilon x^2/2$ with $\epsilon > 0$ and consider the SGD algorithm applied to f with the stepsizes $\alpha_k = \alpha_1/k$:

$$x_{k+1} = x_k - \frac{\alpha_1}{k}\left(x_k^3 + \epsilon x_k\right).$$

Then, if we let $x_1 \geq \sqrt{3/\alpha_1}$, it holds for any $k \geq 1$ that $|x_k| \geq |x_1|\, k!$. $\diamond$

Despite its simplicity, the example highlights that moving beyond standard (upper) quadratic models, SGD may fail to guarantee any convergence. Our goal in this section is to: (i) show that with a simple clipping step added to SGD, the resulting algorithm becomes much more stable; and (ii) to prove strong convergence guarantees for clipped SGD in new settings. Next, we state the first of these results:

Proposition 1 (Stability)**.** *Let Assumptions A1, A2, and A3 hold. Let $\gamma_k \leq \gamma/\sqrt{\alpha_k}$ for some $\gamma > 0$. Let $C = \sigma^2/(2\mu) + \gamma^2$, then, the iterates generated by the clipped SGD method satisfy*

$$\mathbb{E}\left[\operatorname{dist}(x_k, \mathcal{X}^\star)^2\right] \leq \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + C\sum_{i=0}^{k-1} \alpha_i. \quad (3)$$

Some remarks on Proposition 1 are in order. First, unlike SGD, where the distance to the optimal set may grow super-exponentially, the clipped version will not diverge faster than the sum of the used stepsizes. For example, with the stepsizes $O(1/k)$ in Example 1, the sum is only of order $\log(k)$. Such a guarantee will play a critical role in establishing all the convergence results in the subsequent sections. Second, the proposition is reported for time-varying clipping thresholds to facilitate the proofs of some subsequent results. We note however that the similar estimate holds for

the constant scheme with a slightly different scaling constant. Finally, we mention that the bound (3) is similar to the classical results for the stochastic *proximal point* iteration (Ryu & Boyd, 2014, Theorem 6), but slightly weaker than the best bounds for that algorithm (Asi & Duchi, 2019a, Corollary 3.1).

2.1. Convergence under arbitrary growth

Having studied the stability of clipped SGD, we now turn to its consequences for the actual convergence guarantees. We first remark that on *deterministic* convex problems, the procedure (2) is known to be convergent under the very weak growth condition summarized in Assumption A4 below (Alber et al., 1998). Concretely, the subgradients can grow arbitrarily (exponentially, super-exponentially, etc.) as long as they are bounded on bounded sets. However, the situation is less clear as stochastic noise enters the problem. Under A4, similar convergence results have only been established for the stochastic proximal point method (Asi & Duchi, 2019a) and a scaled stochastic approximation algorithm proposed in (Andradóttir, 1996). Note that the former algorithm relies heavily on the ability to accurately model the objective and efficiently solve the resulting minimization problem in each iteration, while the later one needs two independent search directions to construct its upates. Theorem 1 below demonstrates that gradient clipping coupled with mini-batching can also provide such a strong qualitative guarantee.

Assumption A4. *There exits an increasing function* $G_{\mathrm{big}} : \mathbb{R}_+ \to [0, \infty)$ *such that*

$$\mathbb{E}\left[\|f'(x,S)\|_2^2\right] \leq G_{\mathrm{big}}(\operatorname{dist}(x, \mathcal{X}^\star)), \quad \forall x \in \operatorname{dom}(f).$$

Theorem 1. *Let Assumptions A1, A2, and A3 hold. Let* $\gamma_k = \gamma$ *for all* k. *Consider for each* k *a batch of samples* $S_k^{1:m_k}$ *and let* x_k *be generated by the clipped SGD method with* $g_k = \frac{1}{m_k}\sum_{i=1}^{m_k} f'(x, S_k^i)$. *Define* $\varrho_k = \min\{1, \gamma/\|g_k\|_2\}$ *and* $e_k = \operatorname{dist}(x_k, \mathcal{X}^\star)$, *then*

$$\mathbb{E}\left[e_{k+1}^2 \middle| \mathcal{F}_k\right] \leq \left(1 - \mu\alpha_k \mathbb{E}\left[\varrho_k \middle| \mathcal{F}_k\right]\right) e_k^2 + \frac{\sigma^2 \alpha_k}{\mu m_k} + \alpha_k^2 \gamma^2.$$

Suppose further that $\sum_{k=0}^{\infty} \alpha_k/m_k < \infty$, *then under Assumption A4, we have* $\operatorname{dist}(x_k, \mathcal{X}^\star) \xrightarrow{a.s.} 0$.

Theorem 1 highlights the importance of the clipping step as no amount of samples in a batch can save SGD from divergence in this setting. In particular, it implies that clipped SGD converges for any growth function provided that sufficiently accurate estimates of the subgradients can be obtained. This is in stark contrast to SGD without clipping, where, as Example 1 shows, the iterates may diverge even in the noiseless setting when the objective function grows faster than the quadratic x^2. Since the stepsizes are square summable, taking $m_k = 1/\alpha_k$ suffices to guarantee $\sum_{k=0}^{\infty} \alpha_k/m_k < \infty$.

It turns out that clipping can even provide finite convergence rate in this setting, as stated in the next result.

Theorem 2. *Let Assumptions A1, A2, A3, and A4 hold. Let* $\alpha_k = (k+1)^{-\tau}$ *with* $\tau \in (1/2, 1)$ *and let* x_k *be generated by clipped SGD using batches of* $m_k = 1/\alpha_k$ *samples. Define* $e_k := \operatorname{dist}(x_k, \mathcal{X}^\star)$ *and fix a failure probability* $\delta \in (0,1)$, *then for any* $\epsilon > 0$, *there exists a numerical constant* $c_0 > 0$ *such that*

$$\Pr\left(e_K^2 \leq \epsilon\right) \geq 1 - \delta - \frac{\delta\left(\sigma^2/\mu + \gamma^2\right)\sum_{k=0}^{K-1}\alpha_k^2}{e_0^2} - \frac{c_0\alpha_K}{\epsilon}.$$

Furthermore, if we take $\alpha_k = \alpha = \alpha_0 K^{-\tau}$ *with* $\alpha_0 \leq 1/(\mu\varrho)$, *where* $\varrho = \gamma/(\gamma + G_{\mathrm{big}}^{1/2}(\operatorname{dist}(x_0, \mathcal{X}^\star)/\delta))$ *and* $\eta = \left(\sigma^2/\mu + \gamma^2\right)/\mu\varrho$. *Then, for* $K \in \mathbb{N}_+$ *satisfying* $\mu\varrho\alpha_0 K^{1-\tau} \geq \log\left(e_0^2 K^\tau/\eta\alpha_0\right)$, *we have*

$$\operatorname{dist}(x_K, \mathcal{X}^\star)^2 \leq \frac{2\eta\alpha_0}{\delta K^\tau},$$

with probability at least $1 - 2\delta - \delta \cdot \frac{\left(\sigma^2/\mu + \gamma^2\right)\alpha_0^2}{\operatorname{dist}(x_0, \mathcal{X}^\star)^2 K^{2\tau-1}}$.

The first result in the theorem refines the asymptotic guarantee in Theorem 1 for general time-varying stepsizes and the second one shows the iteration complexity for a constant stepsize and fixed mini-batch size. We have the following remarks: (i) Setting τ close to one in the second claim yields a bound with a similar order-dependence on K and δ as for strongly convex and Lipschitz continuous f (Lan, 2020, eq. (4.2.61)). Note that the last term in the last probability bound is negligible for large K; (ii) The proof of the theorem is motivated by a technique developed in (Davis et al., 2019, Lemma 3.3) to bound the escape probability of their algorithm's iterates.

2.2. Convergence under polynomial growth

For the final set of theoretical results of the section, we consider a more specific function class for which we derive the convergence rate of clipped SGD without the need for mini-batching. In particular, we impose the following conditions on the stochastic subgradients.

Assumption A5. *There exist real numbers* $L_0, L_1, \sigma \geq 0$ *and* $2 \leq p < \infty$ *such that for all* $x \in \operatorname{dom}(f)$:

$$\mathbb{E}\left[\|f'(x,S)\|_2^2\right] \leq L_0 + L_1 \operatorname{dist}(x, \mathcal{X}^\star)^{2(p-1)},$$
$$\mathbb{E}\left[\|f'(x,S) - f'(x)\|_2^{2(p-1)}\right] \leq \sigma^p,$$

where $f'(x) = \mathbb{E}[f'(x,S)] \in \partial f(x)$.

Note that when $p = 2$, we have the standard (upper) quadratic growth model (Polyak & Juditsky, 1992). For

general values of p, Assumption A5 implies that

$$\|f'(x)\|_2^2 \leq \mathbb{E}\Big[\|f'(x,S)\|_2^2\Big] \leq L_0 + L_1 \operatorname{dist}(x,\mathcal{X}^\star)^{2(p-1)}$$

which, since f is assumed to be convex, guarantees that

$$f(x) - f(x^\star) \leq \sqrt{L_0}\operatorname{dist}(x,\mathcal{X}^\star) + \sqrt{L_1}\operatorname{dist}(x,\mathcal{X}^\star)^p.$$

We thus allow the function f to grow polynomially from the set of optimal solutions. For example, $f(x) = x^4/4 + \epsilon x^2/2$ satisfies the assumption with $L_0 = L_1 = 2(1+\epsilon)$ and $p = 4$. The second condition in A5 requires that the $2(p-1)$th central moment is bounded, which amounts to finite variance when $p = 2$. We mention that a closely related assumption has been used in (Asi & Duchi, 2019b, Assumption A3) to analyze a method analogous to the classical Polyak subgradient algorithm. The only difference is in the second condition, where they require bounded variation of a quantity involving the objectives instead of the subgradients. This is because $f(x,S)$ and $f(x^\star)$ are used to construct their updates.

The next lemma explicitly bounds the expected norm of the subgradients and the distance between the iterates and the set of optimal solutions. The proof of this lemma follows the same arguments in (Asi & Duchi, 2019b, Lemma B2) and is reported in Appendix E for completeness.

Lemma 2.1. *Let Assumptions A1, A2, and A5 hold. Let x_k be generated by clipped SGD using $\alpha_k = \alpha_0(k+1)^{-\tau}$ with $\tau \in (1/2,1)$, then there exist positive real constants D_0, D_1, G_0, G_1 (independent of k) such that*

$$\mathbb{E}\left[\|f'(x_k,S)\|_2^2\right] \leq G_0 + G_1 k^{(p-1)(1-\tau)},$$
$$\mathbb{E}\left[\operatorname{dist}(x_k,\mathcal{X}^\star)^{4(p-1)}\right] \leq D_0 + D_1 k^{2(p-1)(1-\tau)}.$$

More specifically, if we define for $q \geq 2$ the quantities:

$$P_0(q) := 2^{\frac{q}{2}}\operatorname{dist}(x_0,\mathcal{X}^\star)^q \quad \textit{and}$$
$$P_1(q) := \left((2\gamma)^q + \mu^{-\frac{q}{2}}\sigma^{\frac{q}{4}+1}\right)\left(\frac{2\alpha_0}{1-\tau}\right)^{\frac{q}{2}},$$

then G_0, G_1, D_0, D_1 are given explicitly by

$$G_0 = L_0 + L_1 P_0(2(p-1)), \quad G_1 = L_1 P_1(2(p-1)),$$
$$D_0 = P_0(4(p-1)), \quad D_1 = P_1(4(p-1)).$$

The lemma reveals an attractive property: the subgradients at the iterates can be made small by setting τ close to one, no matter the value of p. This brings us to a position close to where we would have been if we had assumed Lipschitz continuity of f in the first place. The difference is, however, that the preceding guarantees hold w.r.t the full expectation while in the alternative case, one is given a priori an upper-bound on the quantity $\mathbb{E}[\|f'(x_k,S)\|_2^2 \,|\, \mathcal{F}_k]$. We can now state the main result of this subsection.

Theorem 3. *Let Assumptions A1, A2, and A5 hold. Let x_k be generated by clipped SGD using $\alpha_k = \alpha_0(k+1)^{-\tau}$ with $\tau \in (1/2,1)$ and $\gamma_k = \gamma/\sqrt{\alpha_k}$, then there exists a numerical constant C such that*

$$\mathbb{E}\Big[\operatorname{dist}(x_{k+1},\mathcal{X}^\star)^2\Big] \leq \left(1 - \frac{\mu\alpha_0}{(k+1)^\tau}\right)\mathbb{E}\Big[\operatorname{dist}(x_k,\mathcal{X}^\star)^2\Big] + \frac{C}{(k+1)^{2(1-p(1-\tau))}}.$$

Furthermore, we we take $\tau = 1-\epsilon$ for some $\epsilon > 0$, then

$$\mathbb{E}\Big[\operatorname{dist}(x_k,\mathcal{X}^\star)^2\Big] \leq \frac{C}{\mu\alpha_0}\frac{1}{k^{1+\epsilon(1-2p)}} + o\left(\frac{1}{k^{1+\epsilon(1-2p)}}\right).$$

Some remarks on Theorem 3 are in order:

(i) The numerical constant C can be computed as

$$C = (2\gamma^2/\mu)(L_0^2 + L_1^2(D_0+D_1)) + G_0 + G_1,$$

where D_0, D_1, G_0, and G_1 are given in Lemma 2.1. If f is Lipschitz continuous ($L_1 = 0$), then C reduces to $C = (2\gamma^2/\mu)L_0^2 + L_0$. Thus, by setting $\gamma = O(\sqrt{\mu/L_0})$ so that $C = O(L_0)$, we recover the similar order-dependence on L_0 as in the standard bound for unclipped SGD (Lan, 2020).

(ii) Despite the polynomial growth condition, the first bound in the theorem is quite close to the classical estimate for SGD ($\tau = 1$), when applied to Lipschitz continuous f using the *small* stepsize $O(1/k)$ (Lan, 2020; Beck, 2017; Nemirovski et al., 2009). Note also that our guarantee is valid for a wide range of *large* stepsizes $\alpha_k = O(k^{-\tau})$ with $\tau \in (1/2,1)$, which are more robust than the classical $O(1/k)$ (Nemirovski et al., 2009).

(iii) The convergence result follows from a direct application of Chung's lemma (Chung, 1954, Lemma 4) to the first inequality in Theorem 3. The little o term in the estimate vanishes exponentially fast with the sum of the used stepsizes in the manner of (Chung, 1954, p. 467) (see also (Bach & Moulines, 2011)). Since we are free to pick $\epsilon > 0$, we can, in principle, guarantee a rate that is arbitrarily close to $O(1/k)$. Recall that $O(1/k)$ is also the optimal convergence rate for (stochastic) strongly convex and Lipschitz continuous functions (two contradicting conditions) (Nemirovski et al., 2009, Section 2.1). Hence, gradient clipping is able to essentially preserve the optimal rate of SGD while also supporting a broad class of functions on which SGD and its relatives would diverge super-exponentially.

3. Non-asymptotic convergence for weakly convex functions

The previous section demonstrates that gradient clipping can greatly improve the performance of SGD when applied to convex functions with rapidly growing subgradients. We now turn to non-asymptotic convergence analysis of clipped methods under standard growth conditions, but for a much wider class of weakly convex functions. Our goal is to show that the sample complexity of the clipped methods matches the best-known result for weakly convex problems, emphasizing the effectiveness of gradient clipping for a wide range of problem classes.

Recall that that $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ is called ρ-weakly convex if $f + \frac{\rho}{2} \|\cdot\|_2^2$ is convex. Such functions satisfy the following inequality for any $x, y \in \mathbb{R}^n$ with $g \in \partial f(x)$:

$$f(y) \geq f(x) + \langle g, y - x\rangle - (\rho/2) \|y - x\|_2^2 .$$

Weakly convex optimization problems arise naturally in applications described by compositions of the form $f(x) = h(c(x))$, where $h : \mathbb{R}^m \to \mathbb{R}$ is convex and L_h-Lipschitz and $c : \mathbb{R}^n \to \mathbb{R}^m$ is a smooth map with L_c-Lipschitz Jacobian. Note also that all convex functions and all differentiable functions with Lipschitz continuous gradient are weakly convex. We refer to (Duchi & Ruan, 2018; Davis & Drusvyatskiy, 2019; Asi & Duchi, 2019a; Mai & Johansson, 2020) for practical applications as well as recent theoretical and algorithmic developments for this function class.

Algorithm We consider the following momentum extension of clipped SGD:

$$x_{k+1} = x_k - \alpha_k d_k \tag{4a}$$
$$d_{k+1} = \mathrm{clip}_\gamma \left((1 - \beta_k)d_k + \beta_k g_{k+1}\right). \tag{4b}$$

Here, $g_{k+1} = f'(x_{k+1}, S_{k+1})$, $\alpha_k > 0$ is the stepsize, $\beta_k \in (0, 1]$ is the momentum parameter, and $\gamma > 0$ is the clipping threshold. The algorithm is initialized from $x_0 \in \mathbb{R}^n$ and $d_0 = \mathrm{clip}_\gamma(g_0)$ with $g_0 \in \partial f(x_0, S_0)$, and generates the sequences $x_k \in \mathbb{R}^n$ (iterates) and $d_k \in R^n$ (search directions). This algorithm goes back to at least (Gupal & Bazhenov, 1972), and in the sequel, we term procedure (4) as clipped stochastic heavy ball (SHB).

Next we state the standing assumption in this section.

Assumption A6. *There exists a positive real constant L such that* $\mathbb{E}\left[\|f'(x, S)\|_2^2\right] \leq L^2, \quad \forall x \in \mathrm{dom}(f).$

This is a very basic assumption for non-smooth optimization (Nemirovski et al., 2009; Davis & Drusvyatskiy, 2019).

As the function f is neither smooth nor convex, even measuring the progress to a stationary point for f is a challenging task. A common practice is then to use the norm of the gradient of the Moreau envelope as a proxy for near-stationarity (Davis & Drusvyatskiy, 2019). This is possible since weakly convex functions admit an implicit smooth approximation through the classical Moreau envelope:

$$f_\lambda(x) = \inf_{y \in \mathbb{R}^n} \left\{ f(y) + 1/(2\lambda) \|x - y\|_2^2 \right\}. \tag{5}$$

For $\lambda < \rho^{-1}$, the point achieving $f_\lambda(x)$ in (5), denoted by $\mathrm{prox}_{\lambda f}(x)$, is unique and given by:

$$\mathrm{prox}_{\lambda f}(x) = \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + 1/(2\lambda) \|x - y\|_2^2 \right\}. \tag{6}$$

With these definitions, for any $x \in \mathbb{R}^n$, the point $\hat{x} = \mathrm{prox}_{\lambda f}(x)$ satisfies:

$$\begin{cases} \|x - \hat{x}\|_2 = \lambda \|\nabla f_\lambda(x)\|_2 , \\ \mathrm{dist}(0, \partial f(\hat{x})) \leq \|\nabla f_\lambda(x)\|_2 . \end{cases} \tag{7}$$

Thus, a small gradient $\|\nabla f_\lambda(x)\|_2$ implies that x is close to a point $\hat{x}$ that is near-stationary for f.

As for most convergence analyses of subgradient-based methods, we aim to establish the following per-iterate estimate (see, e.g., (Nemirovski et al., 2009; Davis & Drusvyatskiy, 2019; Ghadimi & Lan, 2013)):

$$\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] - c_0 \alpha_k \mathbb{E}[e_k] + c_1 \alpha_k^2. \tag{8}$$

Here e_k is some stationarity measure, V_k is a Lyapunov function, α_k is the stepsize, and c_0, c_1 are some real constants. As discussed above, for minimization of weakly convex functions it is natural to consider $e_k = \|\nabla f_\lambda(x_k)\|^2$. It now remains to find an appropriate Lyapunov function V_k. To build up our V_k, we will go through a number of supporting lemmas. We begin with the one that concerns the search direction d_k.

Lemma 3.1. *Let Assumptions A1 and A6 hold. Let $\beta_k = \nu\alpha_k$ for some constant $\nu > 0$ such that $\beta_k \in (0, 1]$. Let x_k be generated by the clipped SHB method, then*

$$f(x_k) + \mathbb{E}\left[\frac{1-\beta_k}{2\nu} \|d_k\|_2^2 \,\middle|\, \mathcal{F}_k\right] \leq f(x_{k-1}) + \frac{1-\beta_{k-1}}{2\nu} \|d_{k-1}\|_2^2 - \alpha_k \mathbb{E}\left[\|d_k\|_2^2 \,|\mathcal{F}_k\right] + \frac{\alpha_{k-1}^2}{2}\left(\frac{\nu L^2}{1-\beta_0} + \rho\gamma^2\right). \tag{9}$$

It is interesting to note that, despite the bias introduced by the clipping operator, the estimate in (9) is equivalent to that of in (Mai & Johansson, 2020, Lemma 3.1). Moreover, our new proof is arguably simpler, more intuitive and applicable to both constant and time-varying parameters.

The next lemma brings the gradient of the Moreau envelope to the stage.

Lemma 3.2. *Let Assumptions A1 and A6 hold. Let $\beta_k = \nu\alpha_k$ for some constant $\nu > 0$ such that $\beta_k \in (0, 1]$. Let x_k be generated by clipped SHB with $\gamma \geq 2L$ and define*

$$W_k = \frac{1}{2\nu} \|d_k - \nabla f_\lambda(x_k)\|_2^2 - \frac{1}{2\nu} \|\nabla f_\lambda(x_k)\|_2^2 + f(x_k).$$

Let $C = \nu L^2 + \rho\gamma^2/2$, then for any $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[W_k | \mathcal{F}_k\right] \leq W_{k-1} - \alpha_{k-1}\mathbb{E}\left[\langle g_k, \nabla f_\lambda(x_k)\rangle \,|\mathcal{F}_k\right] + \alpha_{k-1}\langle d_{k-1}, \nabla f_\lambda(x_{k-1})\rangle + \frac{\alpha_{k-1}}{\lambda\nu}\|d_{k-1}\|_2^2 + C\alpha_{k-1}^2. \tag{10}$$

To motivate the introduction of W_k, we take a step back and consider the problem where f is assumed to be L-smooth and no gradient clipping is applied. In this case, procedure (4) is identical to the algorithm which was analyzed in (Ruszczynski & Syski, 1983) using a Lyapunov function on the form

$$V_k = \nu f(x_k) + \frac{1}{2}\|d_k - \nabla f(x_k)\|_2^2 + \frac{1}{2}\|d_k\|_2^2.$$

The key insight here is to view the sequence of directions d_k as estimates of the true gradients $\nabla f(x_k)$. With reasonable assumptions, the term $\mathbb{E}[\|d_k - \nabla f(x_k)\|_2^2]$ can indeed be driven to zero (Ruszczynski & Syski, 1983, Theorem 1). Since our f is non-smooth, is not immediately applicable to our problem. Nevertheless, we observe that it is useful to view d_k as an estimate of the gradient of the Moreau envelope. This is the reason why $\|d_k - \nabla f_\lambda(x_k)\|_2^2$ appears in W_k, while the other terms arise from the algebraic manipulations to satisfy (10). Finally, due to the presence of the clipping step, some extra care is needed to make the intuition work. In particular, since the d_k's always belong to $\mathrm{B}(0, \gamma)$, we cannot expect that they approximate the $\nabla f_\lambda(x_k)$ unless these also belong to the γ-ball. It turns out that setting $\gamma \geq 2L$ suffices to ensure that $\nabla f_\lambda(x_k) \in \mathrm{B}(0, \gamma)$.

We now have all the ingredients needed to construct the ultimate Lyapunov function:

Lemma 3.3. *Assume the same setting of Lemma 3.2. Let $\lambda > 0$ be such that $\lambda^{-1} \geq 2\rho$ and consider the function:*

$$V_k = f_\lambda(x_k) + W_k + \frac{f(x_k)}{\lambda\nu} + \left(\frac{1-\beta_k}{2\lambda\nu^2} + \frac{\alpha_k}{\lambda\nu}\right)\|d_k\|_2^2.$$

Then, for any $k \in \mathbb{N}_+$,

$$\mathbb{E}\left[V_k | \mathcal{F}_k\right] \leq V_{k-1} - \frac{\alpha_{k-1}}{2}\|\nabla f_\lambda(x_k)\|_2^2 + C\alpha_{k-1}^2, \tag{11}$$

where $C = \lambda^{-1}\gamma^2(1+\rho/(2\nu)) + \nu L^2(1 + 1/(2\lambda\nu(1-\beta_0)))$.

Finally, the following complexity result follows by a standard argument from (11).

Theorem 4. *Let Assumptions A1 and A6 hold. Let k^* be sampled randomly from $\{0, \ldots, K-1\}$ with $\Pr(k^* = k+1) = \alpha_k / \sum_{i=0}^{K-1}\alpha_i$. Let $\Delta = f(x_0) - \inf_x f(x)$ and let C be given in (11). Then, under the same setting of Lemma 3.3, we have*

$$\mathbb{E}\left[\|\nabla F_\lambda(x_{k^*})\|_2^2\right] \leq 2 \cdot \frac{\xi\Delta + 2L^2/\nu + C\sum_{i=0}^{K-1}\alpha_i^2}{\sum_{i=0}^{K-1}\alpha_i},$$

where $\xi = 2 + 1/(\lambda\nu)$. Furthermore, if we set $\alpha = \alpha_0/\sqrt{K}$ and $\nu = 1/\alpha_0$ for some real $\alpha_0 > 0$

$$\mathbb{E}\left[\left\|\nabla F_{1/(2\rho)}(x_{k^*})\right\|_2^2\right] \leq 2 \cdot \frac{\xi\Delta + 2L^2/\nu + C\alpha_0^2}{\alpha_0\sqrt{K}}.$$

Finally, if α_0 is set to $1/\rho$ and $K \geq 2$, we obtain $\mathbb{E}\left[\left\|\nabla f_{1/(2\rho)}(x_{k^*})\right\|_2^2\right] \leq 6 \cdot \frac{\rho\Delta + \gamma^2}{\sqrt{K}}$.

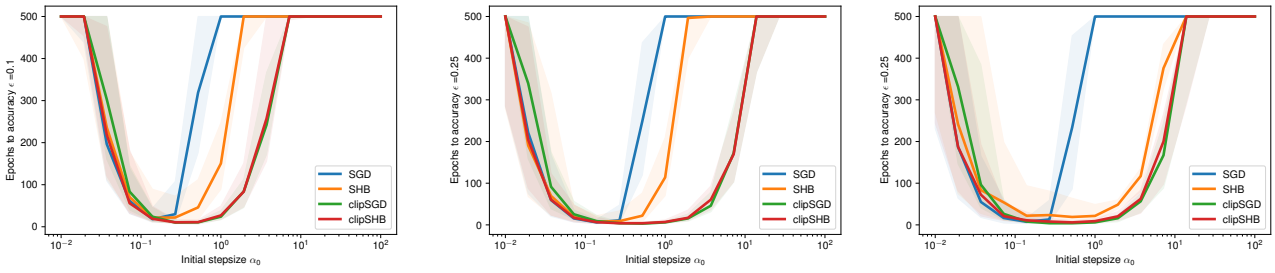
The rate achieved by the clipped SHB is of the same order as the best-known result for weakly convex stochastic problems (Davis & Drusvyatskiy, 2019, Theorem 1). By inspection, all the proofs and convergence results in this section can also be extended (often with significant simplifications) to the case of clipped SGD. The choice $\nu = 1/\alpha_0$ is just for simplicity; we can choose any value of ν as long as $\beta_k = \nu\alpha_k \in (0, 1]$. Since $\beta_k = \nu\alpha_k = O(1/\sqrt{k})$, one can put much more weight on the momentum term than on the fresh subgradient in the search directions d_k. As both α_k and β_k have the same scale, the algorithm can be seen as a single time-scale method (Ghadimi et al., 2020; Ruszczynski & Syski, 1983).

4. Experimental results

For the first two problems, we set up our experiments as follows. We fix $m = 500$, $n = 50$ and generate $A \in \mathbb{R}^{m\times n}$ as $A = QD$, where Q is a matrix with standard normal distributed entries, and D is a diagonal matrix with linearly spaced elements between $1/\kappa$ and 1. Here, $\kappa \geq 1$ represents a condition number which we set to $\kappa = 10$ in all experiments. The algorithms are all randomly initialized at $x_0 \sim \mathcal{N}(0, 1)$ and we use the stepsize $\alpha_k = \alpha_0(k+1)^{-1/2}$, where α_0 is an initial stepsize. We also refer to m stochastic iterations as one epoch (pass over the data). Within each individual run, we set the maximum number of epochs to 500. Each plot reports the results of 30 experiments, visualized as the median of the quantity of interest and the corresponding 90% confidence interval. Finally, the so-called epoch-to-ϵ-accuracy is defined as the smallest number of epochs q needed to reach $f(x_{m\cdot q}) - f(x^\star) \leq \epsilon$.

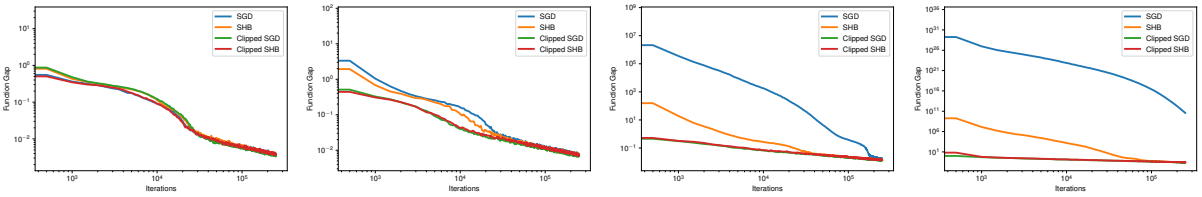
4.1. Phase retrieval

Given m measurements $(a_i, b_i) \in \mathbb{R}^n \times \mathbb{R}$, the (robust) phase retrieval problem seeks a vector $x^\star$ such that



(a) $1-\beta=0.9$, $\epsilon=0.1$ (b) $1-\beta=0.9$, $\epsilon=0.25$ (c) $1-\beta=0.99$, $\epsilon=0.25$

Figure 1. The number of epochs to achieve ϵ-accuracy versus initial stepsize α_0 for phase retrieval with $\gamma=10$.



(a) $\alpha_0=0.139$ (b) $\alpha_0=0.268$ (c) $\alpha_0=0.518$ (d) $\alpha_0=1.0$

Figure 2. The function gap $f(x_k)-f(x^\star)$ versus iteration count for phase retrieval with $\gamma=10$.

$\langle a_i, x^\star\rangle^2 \approx b_i$ for most measurements $i=1,\dots,m$ by solving

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ \frac{1}{m}\sum_{i=1}^{m}\left|\langle a_i,x\rangle^2-b_i\right|.$$

As for the vector b, in each problem instance, we select $x^\star$ uniformly from the unit sphere and construct its elements b_i as $b_i=\langle a_i,x^\star\rangle^2+\delta\zeta_i$, $i=1,\dots,m$, where $\zeta_i\sim\mathcal{N}(0,25)$ models corrupted measurements, and $\delta\in\{0,1\}$ is a binary random variable taking the value 1 with probability $p_{\text{fail}}=0.1$, so that $p_{\text{fail}}\cdot m$ measurements are noisy.

Figure 1 shows the improved robustness provided by gradient clipping for the SGD and SHB algorithms. These clipped methods achieve good accuracies (within the allowed number of epochs) for much wider ranges of initial stepsizes than their unclipped versions. To further elaborate on this, Figure 2 depicts the actual performance for 4 consecutive stepsizes (out of 15) used to produce Figure 1. We can see that these clipped methods always remain stable, while (started from the same initial point) SGD and SHB exhibit the problem of unboundedness when moving beyond their narrow ranges of working parameters.

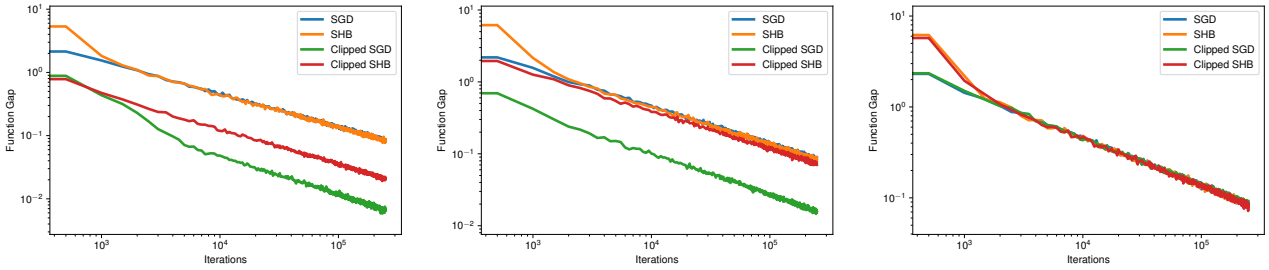
4.2. Absolute linear regression

We consider mean absolute error $f(x)=\frac{1}{m}\|Ax-b\|_1$. For each problem instance, we generate $b=Ax^\star+\sigma w$ for $w\sim\mathcal{N}(0,1)$ and $\sigma=0.01$. The problem is well-behaved as f is both convex and Lipschitz continuous; we did not observe instability of SGD and SHB. Figure 3 shows that gradient clipping does not harm and sometimes can significantly boost the performance of their unclipped counterparts. We can see that although all methods converge with a similar slope, clipped methods may achieve better final accuracies. One possible explanation for this result would be the connection between the used stepsizes and the final error of the (stochastic) subgradient method; we have $\mathbb{E}[f(\bar{x}_k)]-f(x^\star)\le O(\alpha_k)$ for $\alpha_k=O(1/\sqrt{k})$ (Boyd et al., 2003; Duchi, 2018). With clipping, using a smaller γ (if permitted) might have some effect on reducing the effective stepsizes $\alpha_k\min\{1,\gamma/\|g_k\|_2\}$, thereby yielding smaller errors.

4.3. Neural Networks

For our last set of experiments, we consider the image classification task on the CIFAR10 dataset (Krizhevsky et al., 2009) with the ResNet-18 architecture (He et al., 2016). Here, we also compare the previous methods with the Adam algorithm using its default parameters in PyTorch; $\beta_1=0.9$, $\beta_2=0.99$, and $\epsilon=10^{-8}$.[1] Following common practice, we use mini-batch size 128, momentum parameter $\beta=0.9$, and weight-decay coefficient 5×10^{-4} in all experiments. For each algorithm, we conduct 5 experiments (up to 200 epochs) and report the medians of the training loss and test accuracy together with the 90% confidence intervals. For the stepsizes, we use constant values starting with α_0 and reduce them by a factor of 10 every 50 epochs. The initial stepsizes α_0 for Adam are scaled by $1/100$ in actual runs

[1] https://pytorch.org



(a) $\gamma = 0.5$ (b) $\gamma = 1.0$ (c) $\gamma = 10.0$

Figure 3. The function gap $f(x_k) - f(x^\star)$ versus iteration count for absolute linear regression with $\alpha_0 = 5$ and $1 - \beta = 0.9$.

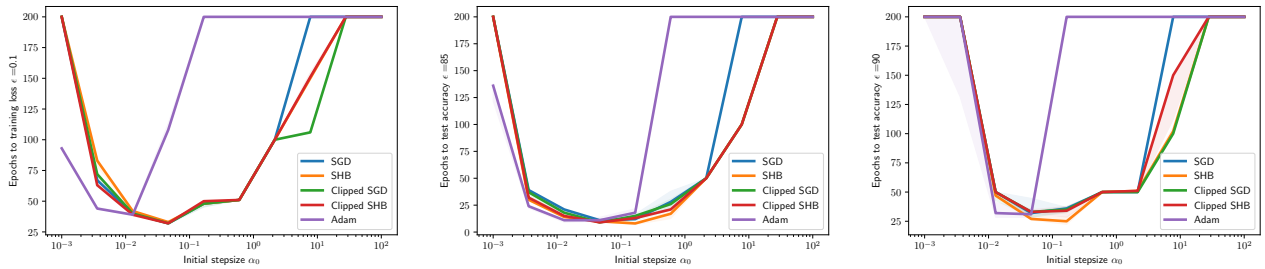


Figure 4. The number of epochs to achieve ϵ training loss and test error versus initial stepsize α_0 for CIFAR10 with $\gamma = 10$.

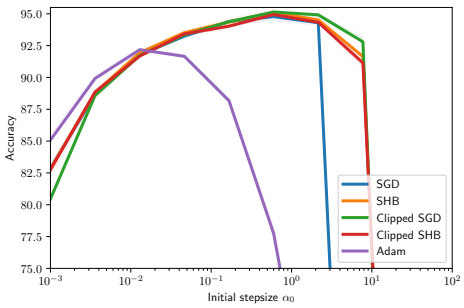


Figure 5. The best achievable accuracy versus initial stepsize α_0 for CIFAR10 with $\gamma = 10$.

(Asi & Duchi, 2019b).

Figure 4 shows the minimum number of epochs required to reach desired values for various performance measures as a function of the initial stepsize. As the classification task on CIFAR10 is a rather well-conditioned problem, the results tell a very similar story to our absolute linear regression experiments. We also observe that Adam is more sensitive to stepsize selection and needs more time to achieve good test performance in this example. To further clarify this, Figure 5 shows that over the tested range of stepsizes, Adam is not able to reach the same best achievable test accuracies that the other methods do.

In summary, the results in this section reinforce our theoretical developments that gradient clipping can: i) stabilize and guarantee convergence for problems with rapidly growing gradients; ii) retain and sometimes improve the best performance of their unclipped counterparts even on standard ("easy") problems.

5. Conclusions

We analyzed clipped subgradient-based methods for solving stochastic convex and non-convex optimization problems. Moving beyond traditional quadratic models, we showed that these methods enjoy strong stability properties and attain classical convergence rates in settings where standard convergence theory does not apply. With a novel Lyapunov analysis, we also proved that the sample complexiy of the methods match the best-known result for weakly convex problems, emphasizing the effectiveness of gradient clipping on a wide range of problem classes.

Acknowledgement

This work was supported in part by the Knut and Alice Wallenberg Foundation, the Swedish Research Council and the Swedish Foundation for Strategic Research. The computations were enabled by resources provided by the Swedish National Infrastructure for Computing (SNIC), partially funded by the Swedish Research Council through grant agreement no. 2018-05973. Finally, we thank the anonymous reviewers for their detailed and valuable feedback.

References

Alacaoglu, A., Malitsky, Y., and Cevher, V. Convergence of adaptive algorithms for weakly convex constrained optimization. *arXiv preprint arXiv:2006.06650*, 2020.

Alber, Y. I., Iusem, A. N., and Solodov, M. V. On the projected subgradient method for nonsmooth convex optimization in a Hilbert space. *Mathematical Programming*, 81(1):23–35, 1998.

Andradóttir, S. A new algorithm for stochastic optimization. In *Proceedings of the 22nd conference on Winter simulation*, pp. 364–366, 1990.

Andradöttir, S. A scaled stochastic approximation algorithm. *Management Science*, 42(4):475–498, 1996.

Asi, H. and Duchi, J. C. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM Journal on Optimization*, 29(3):2257–2290, 2019a.

Asi, H. and Duchi, J. C. The importance of better models in stochastic optimization. *Proceedings of the National Academy of Sciences*, 116(46):22924–22930, 2019b.

Bach, F. and Moulines, E. Non-Asymptotic Analysis of Stochastic Approximation Algorithms for Machine Learning. In *Neural Information Processing Systems (NIPS)*, pp. 451–459, Spain, 2011. URL https://hal.archives-ouvertes.fr/hal-00608041.

Beck, A. *First-order methods in optimization*, volume 25. SIAM, 2017.

Boyd, S., Xiao, L., and Mutapcic, A. Subgradient methods. *Lecture note*, 2003.

Chung, K. L. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pp. 463–483, 1954.

Curtis, F. E., Scheinberg, K., and Shi, R. A stochastic trust region algorithm based on careful step normalization. *Informs Journal on Optimization*, 1(3):200–220, 2019.

Cutkosky, A. and Mehta, H. Momentum improves normalized sgd. In *International Conference on Machine Learning*, pp. 2260–2268. PMLR, 2020.

Davis, D. and Drusvyatskiy, D. Stochastic model-based minimization of weakly convex functions. *SIAM Journal on Optimization*, 29(1):207–239, 2019.

Davis, D. and Grimmer, B. Proximally guided stochastic subgradient method for nonsmooth, nonconvex problems. *SIAM Journal on Optimization*, 29(3):1908–1930, 2019.

Davis, D., Drusvyatskiy, D., and Charisopoulos, V. Stochastic algorithms with geometric step decay converge linearly on sharp functions. *arXiv preprint arXiv:1907.09547*, 2019.

Duchi, J. C. Introductory lectures on stochastic optimization. *The mathematics of data*, 25:99, 2018.

Duchi, J. C. and Ruan, F. Stochastic methods for composite and weakly convex optimization problems. *SIAM Journal on Optimization*, 28(4):3229–3259, 2018.

Ermol'ev, Y. M. and Norkin, V. Stochastic generalized gradient method for nonconvex nonsmooth stochastic optimization. *Cybernetics and Systems Analysis*, 34(2): 196–215, 1998.

Ermoliev, Y. Stochastic quasigradient methods. In *Numerical techniques for stochastic optimization*, pp. 141–185. Springer, 1988.

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

Ghadimi, S., Ruszczyński, A., and Wang, M. A single timescale stochastic approximation method for nested stochastic optimization. *SIAM J. on Optimization*, 2020. Accepted for publication (arXiv preprint arXiv:1812.01094).

Gorbunov, E., Danilova, M., and Gasnikov, A. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *arXiv preprint arXiv:2005.10785*, 2020.

Gupal, A. M. and Bazhenov, L. G. Stochastic analog of the conjugant-gradient method. *Cybernetics and Systems Analysis*, 8(1):138–140, 1972.

Hazan, E., Levy, K., and Shalev-Shwartz, S. Beyond convexity: Stochastic quasi-convex optimization. In *Advances in Neural Information Processing Systems*, pp. 1594–1602, 2015.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Hiriart-Urruty, J.-B. and Lemaréchal, C. *Convex analysis and minimization algorithms*, volume 305. Springer science & business media, 1993.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Lan, G. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer, 2020.

Mai, V. V. and Johansson, M. Convergence of a stochastic gradient method with momentum for non-smooth nonconvex optimization. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 6630–6639, 2020.

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

Nurminskii, E. A. The quasigradient method for the solving of the nonlinear programming problems. *Cybernetics*, 9 (1):145–150, 1973.

Pascanu, R., Mikolov, T., and Bengio, Y. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pp. 1310–1318. PMLR, 2013.

Polyak, B. T. *Introduction to optimization*. Optimization Software, 1987.

Polyak, B. T. and Juditsky, A. B. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

Robbins, H. and Siegmund, D. A convergence theorem for non negative almost supermartingales and some applications. In *Optimizing methods in statistics*, pp. 233–257. Elsevier, 1971.

Ruszczyński, A. A linearization method for nonsmooth stochastic programming problems. *Mathematics of Operations Research*, 12(1):32–49, 1987.

Ruszczynski, A. and Syski, W. Stochastic approximation method with gradient averaging for unconstrained problems. *IEEE Transactions on Automatic Control*, 28(12): 1097–1105, 1983.

Ryu, E. K. and Boyd, S. Stochastic proximal iteration: a non-asymptotic improvement upon stochastic gradient descent. 2014. Author website, early draft.

Shor, N. Z. *Minimization methods for non-differentiable functions*, volume 3. Springer Science & Business Media, 1985.

Zhang, B., Jin, J., Fang, C., and Wang, L. Improved analysis of clipping algorithms for non-convex optimization. *arXiv preprint arXiv:2010.02519*, 2020.

Zhang, J., He, T., Sra, S., and Jadbabaie, A. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In *International Conference on Learning Representations*, 2019.