Stability and Convergence of Stochastic Gradient Clipping

Vien V. Mai and Mikael Johansson

KTH - Royal Institute of Technology

Stochastic optimization

Stochastic optimization problem:

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

Often solved by gradient-based methods using i.i.d. samples drawn from P

SGD: $x_{k+1} = x_k - \alpha_k g_k, \qquad g_k \in \partial f(x_k, S_k)$

Countless variants: momentum, adaptive schemes, averaging,...

Many known problems

- **sensitive** to algorithm parameters $\rightarrow$ costly parameter-tuning
- **unbounded** iterates when f grows quickly

Instability of SGD and relatives

Convergence proofs rely on Lipschitz continuous or bounded gradients

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

$\rightarrow f$ must grow slower than a quadratic everywhere

Example. Let $f(x) = x^4/4 + \epsilon x^2/2$, consider SGD with $\alpha_k = \alpha_1/k$:

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

Then, if $x_1 \geq \sqrt{3/\alpha_1}$, it holds for any $k \geq 1$ that

$$|x_k| \geq |x_1| \, k!.$$

Super-exponential divergence even in the noiseless setting

Stochastic gradient clipping

Clipping operator

$$\mathrm{clip}_\gamma : x \mapsto \min\left\{1, \frac{\gamma}{\|x\|_2}\right\} x$$

Gradient clipping:

- widely used in training models prone to exploding gradients
- introduces **nontrivial bias**

Contributions: effectiveness of gradient clipping in two regimes

- stability and convergence results for rapidly growing convex functions
- sample complexity for weakly convex functions

Outline

- Background and motivation
- **Stability and convergence for fast growing convex functions**
- Sample complexity for stochastic weakly convex minimization
- Numerical examples
- Summary and conclusions

Problem and algorithm: the convex case

Problem:

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

Clipped SGD:

$$x_{k+1} = x_k - \alpha_k d_k, \quad d_k = \text{clip}_{\gamma_k}\left(\frac{1}{m_k}\sum_{i=1}^{m_k} f'(x_k, S_k^i)\right).$$

- m_k is the mini-batch size
- $f'(x_k, S_k^i)$ is a stochastic (sub)gradient

Q: Is clipped SGD any better than standard SGD?

Stability

Assumption. There exists $\mu > 0$ such that such that for all x:

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

Stability: With gradient variance σ^2 and clipping threshold γ, then

$$\mathbb{E}[\operatorname{dist}(x_k, \mathcal{X}^\star)^2] \leq \operatorname{dist}(x_0, \mathcal{X}^\star)^2 + (\sigma^2/(2\mu) + \gamma^2)\sum_{i=0}^{k-1} \alpha_i$$

$\rightarrow$ will not diverge faster than the sum of used stepsizes

Example: For $\alpha_i = O(1/i)$, we have $\sum_{i=0}^{k-1} \alpha_i = \log(k)$.

$\rightarrow$core building block for all the subsequent convergence guarantees

Arbitrary growth

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

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

- gradients can grow arbitrarily
- only the *proximal point method has known asymptotic convergence*

We show clipped SGD with mini-batching converges in this case.

First convergence results

Key estimate. Let $\varrho_k := \min\{1, \gamma_k / \|g_k\|_2\}$ and $e_k = \operatorname{dist}(x_k, \mathcal{X}^\star)$, then

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

Asymptotic convergence. Suppose $\sum_{k=0}^{\infty} \alpha_k / m_k < \infty$, then

$$\operatorname{dist}(x_k, \mathcal{X}^\star) \xrightarrow{a.s.} 0.$$

Finite convergence. Let $\alpha_k = \alpha_0 K^{-\tau}$ with $\tau \in (1/2, 1)$, and fix $\delta \in (0, 1)$:

$$\operatorname{dist}(x_K, \mathcal{X}^\star)^2 \leq O\left(\frac{1}{\delta K^\tau}\right), \quad \text{w.p. at least } 1 - 3\delta.$$

Polynomial growth

Assumption. There exist $L_0, L_1, \sigma \geq 0$ and $2 \leq p < \infty$ such that

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

Convexity of f implies

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

Example: $f(x) = x^4/4 + \epsilon x^2/2$ has $L_0 = L_1 = 2(1+\epsilon)$ and $p = 4$.

We establish near-optimal rate without the need for mini-batching.

Second convergence results

Key observation. Let $\alpha_k = \alpha_0 k^{-\tau}$ with $\tau \in (1/2, 1)$ and $\gamma_k = \frac{\gamma}{\sqrt{\alpha_k}}$, then

$$\mathbb{E}\left[\|f'(x_k, S)\|_2^2\right] \le G_0 + G_1 k^{(p-1)(1-\tau)}.$$

$\rightarrow$ gradient at x_k grows at appropriate rate

Theorem. Let $\tau = 1 - \epsilon$ for some $\epsilon > 0$, then

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

Recall: optimal rate for **Lipschitz continuous** f with $\tau = 1$ is $O\left(\frac{1}{\mu k}\right)$

Summary results for convex functions

Clipping is effective for fast growing convex functions

- much more stable than SGD
- convergence results under arbitrary growth with mini-batching
- near optimal rate for polynomial growth

What if the function grows slowly?

- clipping introduces nontrivial bias $\rightarrow$ might harm convergence
- non-convexity?

Weakly convex minimization and algorithm

Problem:

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

f is ρ-weakly convex, meaning that

$$x \mapsto f(x) + \frac{\rho}{2}\|x\|_2^2 \quad \text{is convex.}$$

Algorithm: Consider a momentum extension

$$\begin{aligned} x_{k+1} &= x_k - \alpha_k d_k \\ d_{k+1} &= \mathrm{clip}_\gamma\left((1-\beta_k)d_k + \beta_k g_{k+1}\right). \end{aligned}$$

Recovers SHB when $\gamma = \infty$; setting $\beta = 1$ gives clipped SGD

Goal: establish sample complexity

Roadmap and challenges

Most complexity results for subgradient-based methods rely on forming:

$$\mathbb{E}[V_{k+1}] \le \mathbb{E}[V_k] - \alpha_k \, \mathbb{E}[e_k] + \alpha_k^2 C^2$$

Immediately yields $O(1/\epsilon^2)$ complexity for $\mathbb{E}[e_k]$

Stationarity measure:

- f convex $\Longrightarrow$ $e_k = f(x_k) - f(x^\star)$
- f smooth $\Longrightarrow$ $e_k = \|\nabla f(x_k)\|_2^2$

Lyapunov analysis (for SGD):

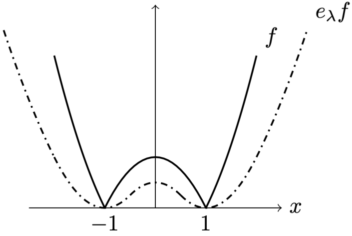
- f convex $\Longrightarrow$ $V_k = \|x_k - x^\star\|_2^2$ [Shor, 1964]
- f smooth $\Longrightarrow$ $V_k = f(x_k)$ [Ghadimi-Lan, 2013]

Not clear what to measure in non-smooth and non-convex case

Convergence to stationarity in weakly convex cases

Moreau envelope

$$F_\lambda(x) = \inf_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\lambda} \|x - y\|_2^2 \right\}$$



Proximal point

$$\hat{x} := \operatorname*{argmin}_{y \in \mathbb{R}^n} \left\{ f(y) + \frac{1}{2\lambda} \|x - y\|_2^2 \right\}.$$

Connection to near-stationarity

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

Small $\|\nabla f_\lambda(x)\|_2 \implies x$ close to a near-stationary point

Lyapunov analysis for clipped SHB

Recall that we wanted

$$\mathbb{E}[V_{k+1}] \leq \mathbb{E}[V_k] - \alpha_k \mathbb{E}[e_k] + \alpha_k^2 C^2,$$

where we chose $e_k = \|\nabla f_\lambda(x_k)\|_2^2$.

Key insight. Viewing d_k as an estimate for $\nabla f_\lambda(x_k)$ leads to:

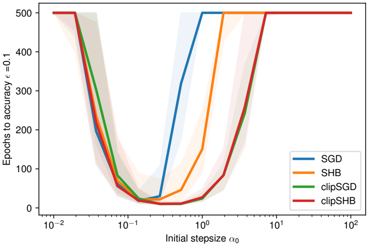
$$W_k = \frac{1}{2\nu} \|d_k - \nabla f_\lambda(x_k)\|_2^2 - \frac{1}{2\nu} \|\nabla f_\lambda(x_k)\|_2^2 + f(x_k).$$

We can then construct the Lyapunov function as:

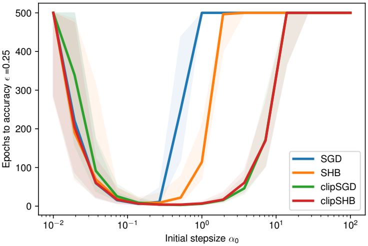
$$\boxed{V_k = f_\lambda(x_k) + W_k + \frac{f(x_k)}{\lambda\nu} + \left(\frac{1-\beta_k}{2\lambda\nu^2} + \frac{\alpha_k}{\lambda\nu}\right) \|d_k\|_2^2 .}$$

$\rightarrow$ immediately yields $\mathbb{E}\left[\, \left\|\nabla f_{1/(2\rho)}(\bar{x}_k)\right\|_2^2 \right] \leq O(1/\sqrt{K})$

Experiments: sensitivity to initial stepsize on phase retrieval



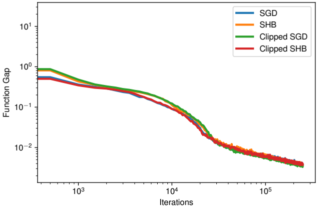
(a) $1-\beta=0.9$, $\epsilon=0.1$



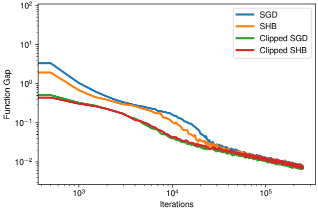
(b) $1-\beta=0.9$, $\epsilon=0.25$

Figure: #epochs to achieve ϵ-accuracy vs. initial stepsize α_0 for phase retrieval.

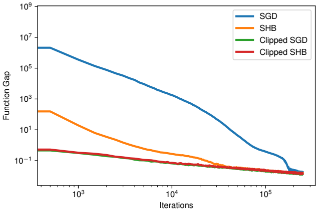
Experiments: convergence behavior on phase retrieval



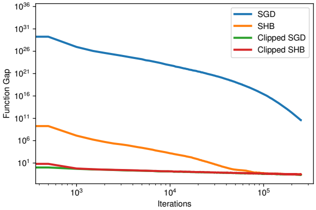
(a) $\alpha_0 = 0.139$



(b) $\alpha_0 = 0.268$



(c) $\alpha_0 = 0.518$



(d) $\alpha_0 = 1.0$

Experiments: neural networks

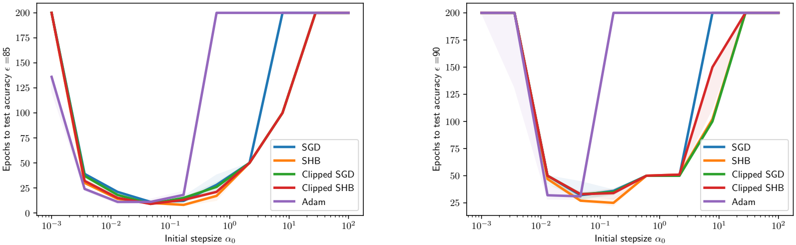


Figure: #epochs to achieve ϵ test accuracy vs. initial stepsize α_0 for CIFAR10

Conclusion

Stochastic gradient clipping

- simple modifications to SGD
- good performance and less sensitive to algorithm parameters

Fast growing convex functions

- various qualitative and quantitative convergence results

Novel Lyapunov analysis

- sample complexity of clipped SHB for weakly convex minimization