Adaptive Gradient Normalization and Independent Sampling for (Stochastic) Generalized-Smooth Optimization

Yufeng Yang, Erin E. Tripp, Yifan Sun, Shaofeng Zou, Yi Zhou

Presnted by Yufeng Yang, ICCOPT 2025, Los Angeles

July 24, 2025

L-smooth condition

Consider the optimization problem

$$\min_{w \in \mathbf{R}^d} f(w) \tag{1}$$

where $f : \mathbf{R}^d \to \mathbf{R}$ denotes a nonconvex and differentiable function; w corresponds to the model parameters.
To study first-order algorithm convergence for optimization (1), classical theory assumes L-smooth condition of $\nabla f(w)$.

Definition: L-smooth

A differentiable function $f : \mathbf{R}^d \to \mathbf{R}$ is said to be L-smooth, if for all $w, w' \in \mathbf{R}^d$, we have

$$\|\nabla f(w) - f(w')\| \leq L\|w - w'\|. \tag{2}$$

Geometric Intuition behind L-smooth

From L-smooth definition, we know

1. "descent inequality":

$$f(w) \leq f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{L}{2} \left\| w - w' \right\|^2.$$

2. one can upper bound $f(w)$ by a quadratic function.

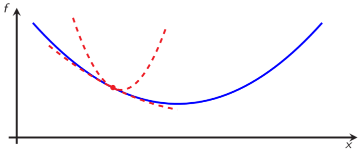


Figure: Visualization of L-smooth & strongly convex function [Taylor et al, (2020)]

Q: Does L-smooth condition hold in real applications?

Motivation Example: Phase Retrieval

Given m intensity measurements $y_r = |a_r^T w|^2 + n_r$ for $r = 1, ..., m$, where a_r is the measurement vector and n_r is the additive noise. Phase retrieval reconstructs underlying object w^* by solving the regression problem,

$$\min_{w \in \mathbf{R}^d} F(w) = \frac{2}{m} \sum_{r=1}^{m} f_\xi(w) = \frac{1}{2m} \sum_{r=1}^{m} \left(y_r - \left|a_r^T w\right|^2\right)^2. \tag{3}$$

Property of $f_\xi(w)$ in (3)

For any $w, w' \in \mathbf{R}^d$, $f_\xi(w) = \frac{1}{4}(y_\xi - |a_\xi^T w|^2)^2$ satisfies

$$\left\|\nabla f_\xi\left(w'\right) - \nabla f_\xi(w)\right\| \leq \left\|w' - w\right\| \cdot \mathcal{O}\left(a_{\max}^{\frac{4}{3}} \left\|\nabla f_\xi\left(w'\right)\right\|^{\frac{2}{3}} + a_{\max}^{\frac{4}{3}} \left\|\nabla f_\xi(w)\right\|^{\frac{2}{3}} + y_{\max} a_{\max}^2\right)$$

Key observation: Additional $\nabla f_\xi(w), \nabla f_\xi(w')$ on the RHS, L-smooth failed.

Motivation Example: DRO

According to [Levy et al. (2020)]; [Jin et al, (2021)], under mild assumptions, ϕ-divergence regularized distributionally robust optimization (DRO) has following dual reformulation

$$\min_{w\in\mathbf{R}^d,\eta\in\mathbf{R}} L(w,\eta) = \lambda\mathbb{E}_{\xi\sim P}\phi^*\left(\frac{\ell_\xi(w)-\eta}{\lambda}\right)+\eta. \tag{4}$$

Property of (4) [Jin et al, (2021)]; [Chen et al, (2023)]

For any $(w,\eta),(w',\eta')\in\mathbf{R}^d\times\mathbf{R}$, under mild assumptions on $\ell_\xi(\cdot)$ and ϕ^*, (4) satisfies

$$\|\nabla L(w,\eta)-\nabla L(w',\eta')\| \le (L+\frac{2M(G+1)^2}{\lambda}+L\|\nabla L(w,\eta)\|) \cdot \|(w,\eta)-(w',\eta').\|$$

Key observation: Additional $\nabla f_\xi(w),\nabla f_\xi(w')$ on the RHS, L-smooth again failed.

Motivation Example: Neural Networks

According to [Zhang et al. (2019)], they empirically observe that the smoothness parameter scale with norm linearly

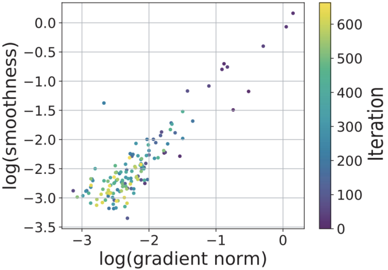


Figure: Gradient norm vs local gradient Lipschitz constant on a log-scale along the training trajectory ([Zhang et al. (2019)]).

Section 2

Generalized Smooth Condition

$\mathcal{L}^*_{\mathbf{asym}}(\alpha)$-generalized smooth condition [Chen et al, (2023)]

- f is differentiable and bounded below.
- There exists constants $L_0, L_1 > 0$ and $\alpha \in [0, 1]$ such that for any w, $w' \in \mathbf{R}^d$, we have

$$\left\|\nabla f(w) - \nabla f(w')\right\| \le \left(L_0 + L_1\left\|\nabla f(w')\right\|^{\alpha}\right)\left\|w - w'\right\|. \quad (5)$$

Under above assumption, we have "descent inequality"

$$f(w) \le f(w') + \langle \nabla f(w'), w - w' \rangle + \frac{1}{2}(L_0 + L_1 \underbrace{\left\|\nabla f(w')\right\|^{\alpha}}_{\text{additional term}})\left\|w - w'\right\|^2. \quad (6)$$

This characterizes a broader class of irregular geometries than those captured by L-smooth condition.

Challenges to GD

Under **generalized-smooth** condition, gradient descent is hard to analyze and performs worse because...

1. it requires an additional assumption that

$$\|\nabla f(w)\| \leq G = \sup\{u | u^2 \lesssim \mathcal{O}(\ell(u) \times \Delta_0)\}, \tag{7}$$

   where ℓ is a sub-quadratic function, according to [Li et al. (2024)].
2. Condition (7) is implicit, hard to find efficient estimation in practice.
3. G is highly dependent on function value gap $\Delta_0 = f(w_0) - f^*$ and initialization distance $\|w_0 - w^*\|$.
4. Convergence is established by requiring learning rate satisfying $\gamma < \mathcal{O}(1/G)$, which can be slow.

Adaptive-Normalized GD

Why Normalization?

Q: Having observed the RHS of "descent inequality" including $(L_0 + L_1\|\nabla f(w)\|^\alpha)\|w - w'\|$, how can we control the term induced by $\|\nabla f(w)\|^\alpha$?
A: Normalized or Clipped gradient descent algorithms

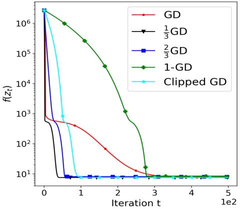
1. In this work, we consider Adaptively Normalized Gradient-Descent [Chen et al, (2023)]. The update rule is

$$\text{(AN-GD)} \quad w_{t+1} = w_t - \gamma \frac{\nabla f(w_t)}{\|\nabla f(w_t)\|^\beta}, \tag{8}$$

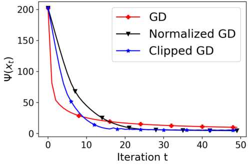
   where $\beta \in [\alpha, 1]$.
2. By allowing $\beta < 1$, when $\|\nabla f(w_t)\|$ is large, β-normalization makes the update more aggressive.
3. when $\|\nabla f(w_t)\|$ is small, β-normalization can stabilize the update against divergence.

Theory-Practice Gap of AN-GD

1. [Chen et al, (2023)] proved $\mathcal{O}(\epsilon^{-2})$ convergence for nonconvex and differentiable generalized-smooth function f in order to obtain a ϵ-stationary point.
2. It's unclear why AN-GD performs better than GD for problem like Phase Retrieval, DRO, etc.



Phase Retrieval (Chen et al. (2023))



ϕ-divergence Regularized DRO (Chen et al. (2023))

Generalized PŁ Condition

Generalized Polyak-Łojasiewicz (PŁ) Condition

There exists constants $\mu \in \mathbf{R}_+$ and $0 < \rho \leq 2$ such that $f(\cdot)$ satisfies, for all $w \in \mathbf{R}^d$,

$$\left\|\nabla f(w)\right\|^\rho \geq 2\mu(f(w) - f^*). \tag{9}$$

According to [Zhou et al. (2016)], [Liu et al. (2022)], [Scaman et al. (2022)], phase retrieval, over-parametrized neural-network satisfy this condition under mild assumptions.

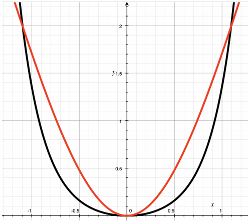


Figure: Red Curve ($\rho = 2$); Black Curve ($\rho = 1$)

Convergence Theory and Its Implications

Convergence Result of AN-GD (Informal)

Let inequalities (5) and (9) hold. Denote $\Delta_t := f(w_t) - f^*$ as the function value gap. define learning rate $\gamma = \mathcal{O}(\frac{(\mu\epsilon)^{\beta/\rho}}{L_0+L_1})$ for some $\beta \in [\alpha, 1]$. Then, to achieve $\Delta_T \leq \epsilon$, the following statements hold.

- When $\rho + \beta < 2$, the total number of iterations must satisfy

$$T \geq \Omega\big((\frac{1}{\epsilon})^{\frac{2-\rho}{\rho}}\big). \tag{10}$$

1. When ρ is very small such that $\rho + \beta < 2$, the effects of β can be marginal.

Convergence Theory and Its Implication, Continued

- If $\rho+\beta=2$, the total number of iterations must satisfy

$$T \geq \Omega\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho}} \log(\frac{\Delta_0}{\epsilon})\big). \tag{11}$$

- If $\rho+\beta>2$, there exists a time T_0 such that the total number of iterations after T_0 must satisfy

$$T \gtrsim \Omega\big(\log\big((\frac{1}{\epsilon})^{\frac{\beta}{\rho+\beta-2}}\big)\big). \tag{12}$$

1. When $\rho=2, \beta=0$, it recovers linear convergence achieved by gradient descent under the standard PŁ and L-smooth condition.
2. Once $\rho+\beta>2$, AN-GD exhibits a two-phase convergence behavior, where the latter phase accelerates the rate from polynomial to local linear convergence.

A Special Example

Moreover, this theorem reveals varying β smaller than 1 do accelerate convergence under certain geometry...

Example

when $\rho = 1$ and consider $\beta_1 = \frac{2}{3}, \beta_2 = 1$, AN-GD achieves the iteration complexities $\mathcal{O}(\epsilon^{-1})$ and $\tilde{\mathcal{O}}(\epsilon^{-1})$ respectively.

Q: Can we generalize AN-GD for solving stochastic optimization problems?

AN-SGD

Through out, we denote $f_\xi(w)$ as the loss function associated with the data sample ξ, and we minimize the expected loss function $F(\cdot)$ satisfies the generalized-smooth condition (inequality (5)).

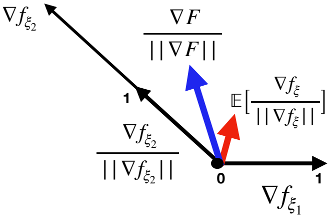
$$\min_{w \in \mathbf{R}^d} F(w) = \mathbb{E}_{\xi \sim \mathbb{P}}\big[f_\xi(w)\big]. \tag{13}$$

The straightforward extension of AN-GD under stochastic setting is to replace $\nabla f(w)$ by $\nabla f_\xi(w)$, resulting

$$\text{(AN-SGD)} \quad w_{t+1} = w_t - \gamma \frac{\nabla f_\xi(w_t)}{\|\nabla f_\xi(w_t)\|^\beta}. \tag{14}$$

The variations of AN-SGD has been studied extensively, for example, Clipped-SGD, Normalized SGD with momentum. They can achieve a sample complexity of $\mathcal{O}(\epsilon^{-4})$ under generalized-smooth and mild noise assumptions.

What's the potential limitation?



1. **Biased** gradient estimator, i.e., $\mathbb{E}[\frac{\nabla f_\xi(w_t)}{\|\nabla f_\xi(w_t)\|^\beta}] \neq \frac{\nabla F(w_t)}{\|\nabla F(w_t)\|^\beta}$. This is due to the dependence between $\nabla f_\xi(w_t)$ and $\|\nabla f_\xi(w_t)\|^\beta$.
2. **Strong** assumption in analysis, i.e.,
   1. *Almost sure bounded approximation error*, i.e., $\|\nabla f_\xi(w) - \nabla F(w)\| \leq \tau_2$ a.s.. ([Zhang et al. (2019)], [Zhang et al. (2020)], [Liu et al. (2022)])
   2. *Large* batch size up to $B \sim \Omega(\epsilon^{-2})$ to control stochastic gradient noise at $\mathcal{O}(\epsilon)$-level. ([Chen et al, (2023)], [Reisizadeh et al. (2023)])

Independent Sampling

We propose the following independently-and-adaptively normalized stochastic gradient (IAN-SG) estimator

$$\text{(IAN-SG estimator)} \quad \frac{\nabla f_{\xi}(w)}{\|\nabla f_{\xi'}(w)\|^{\beta}}, \tag{15}$$

where ξ and ξ' are samples draw *independently* from the underlying data distribution.

Intuition on independent sampling

The independence between ξ and ξ' decorrelates the denominator from the numerator, making update direction unbiased (difference up to a scaling factor), i.e.,

$$\mathbb{E}_{\xi,\xi'}\left[\frac{\nabla f_{\xi}(w)}{\|\nabla f_{\xi'}(w)\|^{\beta}}\right] = \mathbb{E}_{\xi'}\left[\frac{\mathbb{E}_{\xi}\left[\nabla f_{\xi}(w)\right]}{\|\nabla f_{\xi'}(w)\|^{\beta}}\right] \propto \nabla F(w). \tag{16}$$

IAN-SGD Framework

Challenges

Hard to control $\mathbb{E}_{\xi'}[\|\nabla f_{\xi'}(w)\|^{-\beta}]$.

We propose independently-and-adaptively normalized SGD (IAN-SGD) algorithm, where A, Γ,δ are positive constants,

$$\text{(IAN-SGD):} \quad w_{t+1} = w_t - \gamma \frac{\nabla f_\xi(w_t)}{h_t^\beta},$$
$$\text{where} \quad h_t = \max\Big\{1, \Gamma \cdot \Big(A\big\|\nabla f_{\xi'}(w_t)\big\| + \delta\Big)\Big\}. \tag{17}$$

Intuition behind IAN-SGD

1. *Clipping* doesn't slow down convergence too much, as when $\|\nabla F(w)\| \downarrow 0$, generalized-smooth condition reduces to L-smooth condition.
2. *Imposing* constant lower bound, δ, on h_t helps avoid numerical instability in practice. (Similar as Adam, Adagrad etc.)

IAN-SGD Convergence

Noise Assumptions

We adopt the following noise assumptions for analysis.

1. $\nabla f_\xi(w)$ is unbiased, i.e., $\mathbb{E}_{\xi\sim\mathbb{P}}\big[\nabla f_\xi(w)\big] = \nabla F(w)$.
2. There exists $0 \le \tau_1 < 1, \tau_2 > 0$ such that for any $w \in \mathbf{R}^d$,

$$\big\|\nabla f_\xi(w) - \nabla F(w)\| \le \tau_1 \big\|\nabla F(w)\big\| + \tau_2 \;\; \text{a.s.} \;\; \forall \xi \sim \mathbb{P}. \qquad (18)$$

Above assumption implies

1. $\|\nabla F(w_t)\| \le \frac{1}{1-\tau_1}\|\nabla f_\xi(w_t)\| + \frac{\tau_2}{1-\tau_1}$. Thus, one can choose $A = \frac{1}{1-\tau_1}$, $\delta = \frac{\tau_2}{1-\tau_1}$.
2. When gradient noise is heavy-tailed, i.e., $\tau_1 \uparrow 1$ and τ_2 is large, we should increase A and δ accordingly, ensuring that the normalization term dominates h_t.

IAN-SGD Convergence Continued

Convergence Result(Informal)

For IAN-SGD algorithm, choose learning rate $\gamma = \mathcal{O}(\frac{1}{\sqrt{T}})$, and $A = \frac{1}{1-\tau_1}$ $\delta = \frac{\tau_2}{1-\tau_1}$, $\Gamma = (4L_1\gamma(2\tau_1^2 + 1))^{\frac{1}{\beta}}$.
Denote $\Lambda = F(w_0) - F^* + \frac{1}{2}(L_0 + L_1)(1 + 4\tau_2^2)^2$.
Then, with probability at least $\frac{1}{2}$, IAN-SGD produces a sequence satisfying $\min_{t \leq T} \|\nabla F(w_t)\| \leq \epsilon$ if the total number of iteration T satisfies

$$T \geq \mathcal{O}(\Lambda^2 \epsilon^{-4}). \qquad (19)$$

IAN-SGD Convergence Continued

Above Theorems...

1. recovers similar convergence rate in [Zhang et al. (2019)] when $\tau_1 = 0$.
2. requires sampled ξ, ξ' at $\Omega(1)$-level.
3. establishes $\mathcal{O}(\epsilon^{-4})$ convergence under weaker noise assumption.

Open Problem

However, Our noise assumption (18) is still stronger than expected noise assumption, i.e.,

$$\mathbb{E}_{\xi}\|\nabla f_{\xi}(w) - \nabla F(w)\|^{\kappa} \leq \tau_2^{\kappa}, \kappa \in (1, 2]. \tag{20}$$

[Koloskova et al. (2023)] showed that Clipped-SGD achieves a convergence rate of $\mathcal{O}(\epsilon^{-5})$ when $\kappa = 2$, provided that the sampled ξ is at the $\Omega(1)$ level. Q(Open): Is there a way to modify the algorithm design or refine the analysis so that normalized stochastic gradient methods can achieve $\mathcal{O}(\epsilon^{-4})$ while maintaining an $\Omega(1)$-level batch size under the generalized-smooth and expected noise assumptions?

Section 5

Phase Retrieval and DRO

We compare the objective values of Phase Retrieval (3) and DRO (4) versus sample complexity using IAN-SGD and other baselines in the following figures.

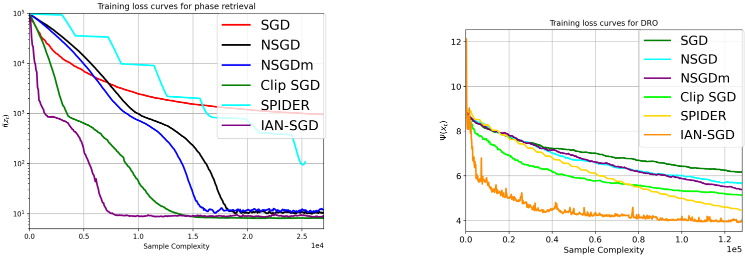


Figure: Loss vs. Sample Plot for Phase Retrieval (Left) and DRO (Right)

Training ResNet

We compare the cross-entropy loss of ResNet on CIFAR-10 versus the number of epochs using IAN-SGD and other baselines in the following figures.

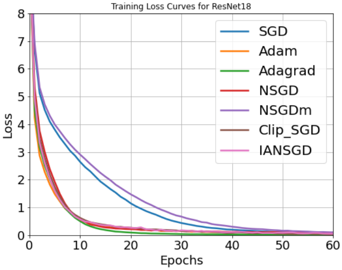


Figure: Loss vs. Epoch Plot for ResNet18 (Left) and ResNet50 (Right)

Paper

Code

Thank You!

References

Adrien Taylor (2020)
Computer-aided analyses in optimization
Machine Learning Research Blog

Jikai Jin, Bohang Zhang, Haiyang Wang, Liwei Wang (2021)
Non-convex distributionally robust optimization: Non-asymptotic analysis.
In Advances in Neural Information Processing Systems, 2021

Daniel Levy, Yair Carmon, John C Duchi , Aaron Sidford (2020).
Large-scale methods for distributionally robust optimization
In Advances in Neural Information Processing Systems, 2020.

Ziyi Chen, Yi Zhou, Yingbin Liang, Zhaosong Lu (2023)
Generalized-smooth nonconvex optimization is as efficient as smooth nonconvex optimization.
In Proceedings of the 40th International Conference on Machine Learning, 2023.

Jingzhao Zhang, Tianxing He, Suvrit Sra, Ali Jadbabaie.
Why gradient clipping accelerates training: A theoretical justification for adaptivity.
In International Conference on Learning Representations, 2019.

Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, Ali Jadbabaie.
Convex and non-convex optimiza- tion under generalized smoothness.
In Advances in Neural Information Processing Systems, 2024.

Yi Zhou, Huishuai Zhang, Yingbin Liang.
Geometrical properties and accelerated gradient solvers of non-convex phase retrieval.
In 2016 54th Annual Allerton Conference on Communication, Control, and Computing (Allerton), pp. 331–335, 2016.

Chaoyue Liu, Libin Zhu, Mikhail Belkin.
Loss landscapes and optimization in over-parameterized non- linear systems and neural networks.
Applied and Computational Harmonic Analysis, 59:85–116, 2022a.

Kevin Scaman, Cedric Malherbe, Ludovic Dos Santos.
Convergence rates of non-convex stochastic gradient descent under a generic lojasiewicz condition and local smoothness.
In Proceedings of the 39th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, 17–23 Jul 2022.

Bohang Zhang, Jikai Jin, Cong Fang, and Liwei Wang.
Improved analysis of clipping algorithms for non- convex optimization.
In Advances in Neural Information Processing Systems, 2020a.

Mingrui Liu, Zhenxun Zhuang, Yunwen Lei, and Chunyang Liao.
A communication-efficient distributed gra- dient clipping algorithm for training deep neural networks.
In Advances in Neural Information Processing Systems, 2022b.

Amirhossein Reisizadeh, Haochuan Li, Subhro Das, and Ali Jadbabaie.
Variance-reduced clipping for non-convex optimization.
ICASSP 2025 - 2025

Anastasia Koloskova, Hadrien Hendrikx, and Sebastian U Stich.
Revisiting gradient clipping: Stochastic bias and tight convergence guarantees.
In International Conference on Machine Learning, pp. 17343–17363. PMLR, 2023.

Acknowledgements

- This work was supported in part by the ***Air Force Research Laboratory Information Directorate*** (AFRL/RI), through AFRL/RI Information Institute, contract number FA8750-20-3-1003.
- The work of Yufeng Yang and Yi Zhou was supported by the ***National Science Foundation*** under grants DMS-2134223, ECCS-2237830.
- The work of Shaofeng Zou was supported in part by ***National Science Foundation*** under Grants CCF-2438429 and ECCS-2438392(CAREER).
- The work of Yifan Sun was supported in part by the ***Office of Naval Research***, grant W911NF-22-1-0292.