High-Probability Convergence for Composite and Distributed Stochastic Minimization and Variational Inequalities with Heavy-Tailed Noise

Eduard Gorbunov MBZUAI

Abdurakhmon Sadiev KAUST

Marina Danilova MIPT

Samuel Horváth MBZUAI

Gauthier Gidel UdeM

Pavel Dvurechensky WIAS

Alexander Gasnikov Innopolis University

Peter Richtárik KAUST

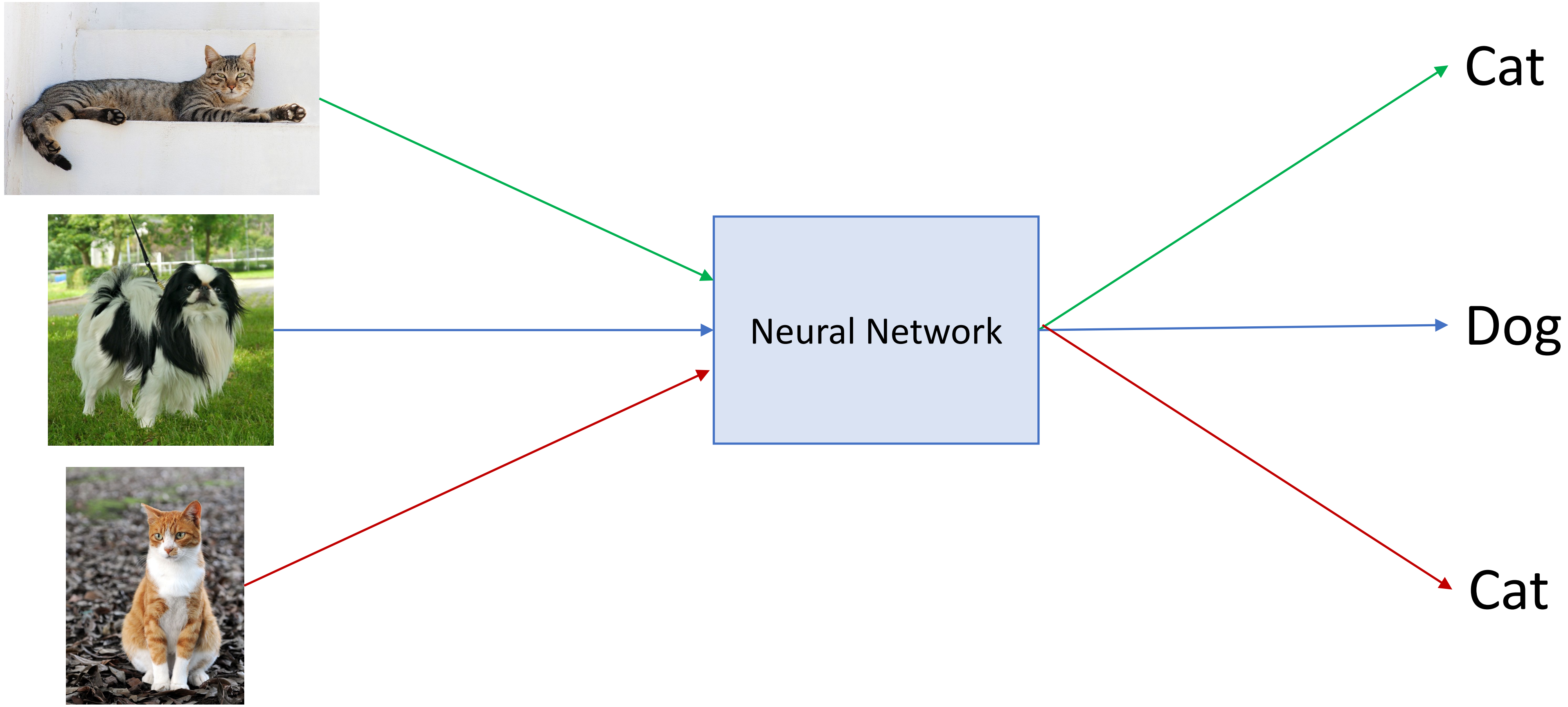
ICML 2024, Vienna

Heavy-Tailed Noise

Typical Machine Learning Problem: Classification

Training data
(n images)

Goal: classify what is on the picture – cat or dog



Pictures sources: https://en.wikipedia.org/wiki/Cat, https://en.wikipedia.org/wiki/Dog

Typical Machine Learning Problem

$$\min_{x \in \mathbb{R}^d} \left\{ f(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x) \right\}$$

Typical Machine Learning Problem

- Dimension of the model: $\boxed{d}$
- Model parameters: $\boxed{x}$

$$\min_{\boxed{x} \in \mathbb{R}^{\boxed{d}}} \left\{ f(\boxed{x}) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(\boxed{x}) \right\}$$

Typical Machine Learning Problem

- Dimension of the model: $\boxed{d}$
- Model parameters: $\boxed{x}$
- Training data: $\boxed{n}$ samples
- Loss on the i-th sample: $\boxed{f_i}(x)$
- Training loss: $\boxed{f}(x)$

$$\min_{\boxed{x} \in \mathbb{R}^{\boxed{d}}} \left\{ \boxed{f}(\boxed{x}) \stackrel{\text{def}}{=} \frac{1}{\boxed{n}} \sum_{i=1}^{\boxed{n}} \boxed{f_i}(\boxed{x}) \right\}$$

Typical Machine Learning Problem

- Dimension of the model: $\boxed{d}$
- Model parameters: $\boxed{x}$
- Training data: $\boxed{n}$ samples
- Loss on the i-th sample: $\boxed{f_i}(x)$
- Training loss: $\boxed{f}(x)$

$$\min_{\boxed{x} \in \mathbb{R}^{\boxed{d}}} \left\{ \boxed{f}(\boxed{x}) \stackrel{\text{def}}{=} \frac{1}{\boxed{n}} \sum_{i=1}^{\boxed{n}} \boxed{f_i}(\boxed{x}) \right\}$$

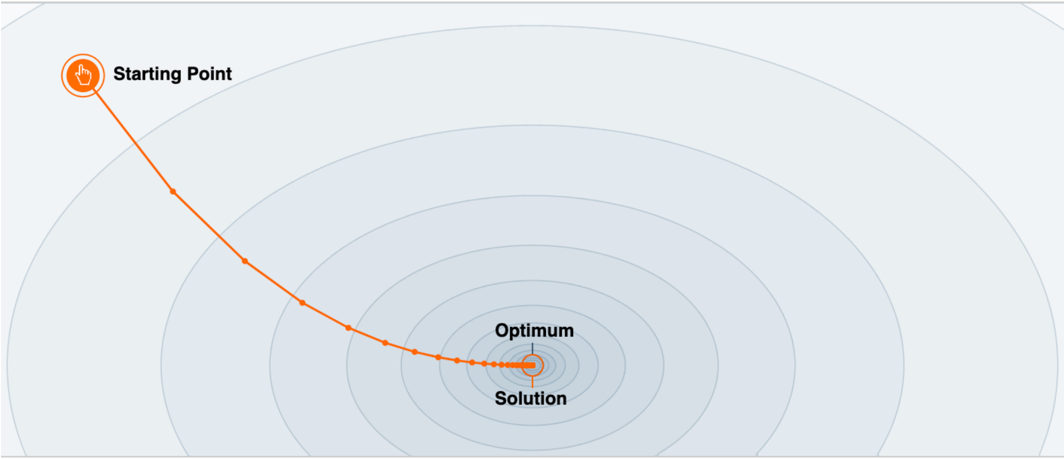
$\boxed{d}$ and $\boxed{n}$ are usually very large…

Computation of $\nabla f(x)$ is very expensive $\Longrightarrow$ stochastic methods are used

Gradient Descent vs Stochastic Gradient Descent

Gradient Descent (GD)

$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

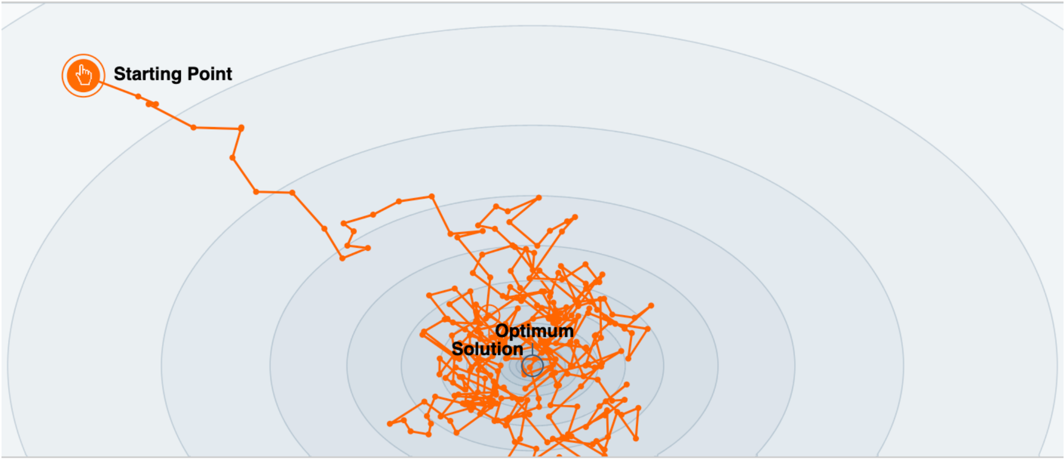


Convergence to the exact optimum asymptotically
High computation cost of one iteration

Stochastic Gradient Descent (SGD)

$$x^{k+1} = x^k - \gamma \nabla f_{\boxed{i_k}}(x^k)$$

Random index from $\{1, 2, \dots, n\}$



Convergence to the neighborhood of the solution
Cheap iterations
Faster convergence (for most of ML problems)

H. Robbins, S. Monro. A stochastic approximation method (The annals of mathematical statistics 1951).

Pictures source: https://fa.bianp.net/teaching/2018/COMP-652/

Gradient Descent vs Stochastic Gradient Descent

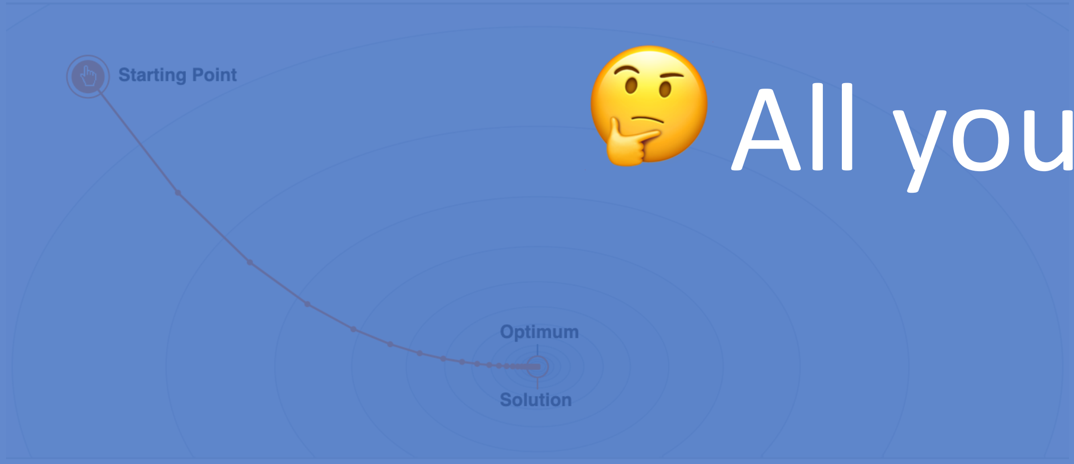
Gradient Descent (GD)

$$x^{k+1} = x^k - \gamma \nabla f(x^k)$$

Stochastic Gradient Descent (SGD)

$$x^{k+1} = x^k - \gamma \nabla f_{\boxed{i_k}}(x^k)$$

Random index from $\{1, 2, \dots, n\}$



🤔 All you need is SGD?

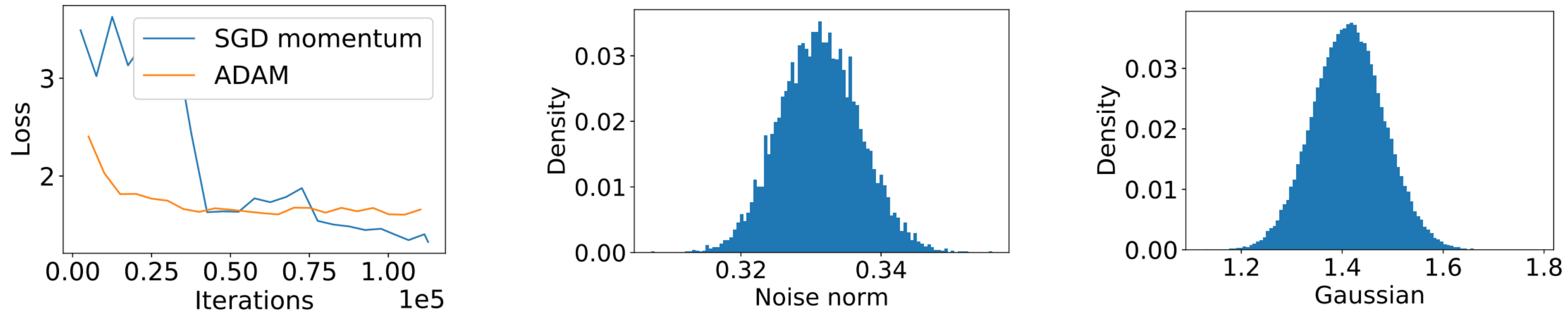
Convergence to the exact optimum asymptotically
High computation cost of one iteration

Convergence to the neighborhood of the solution
Cheap iterations
Faster convergence (for most of ML problems)

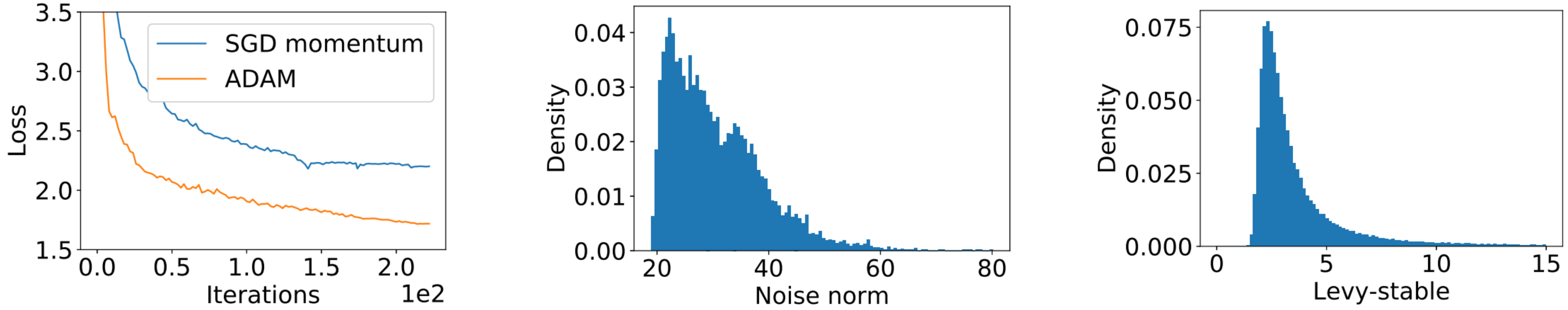
H. Robbins, S. Monro. A stochastic approximation method (The annals of mathematical statistics 1951).

Pictures source: https://fa.bianp.net/teaching/2018/COMP-652/

Choice of the Method is Important



(a) (b) ImageNet training (c) Synthetic Gaussian



(e) (f) Bert pretraining (g) Synthetic Levy-stable

!! If the noise is heavy-tailed, SGD is not a good choice **(not even guaranteed to converge)**

✍ Heavy-tailed noise in the stochastic gradients is typical for training LLMs and GANs

J. Zhang et al. Why are adaptive methods good for attention models? (NeurIPS 2020).

From Empirical Risk To Expected Risk Minimization

Empirical risk minimization (ERM):

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^{n} f_i(x) \right\}$$

Risk minimization (RM):

$$\min_{x\in\mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)] \right\}$$

From Empirical Risk To Expected Risk Minimization

Empirical risk minimization (ERM): $$\min_{x\in\mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{i=1}^n f_i(x) \right\}$$

Risk minimization (RM): $$\min_{x\in\mathbb{R}^d} \left\{ f(x) \stackrel{\text{def}}{=} \mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)] \right\}$$

- The first problem is a special case of the second one
- If n is large enough, then the minimizer of ERM is close to the minimizer of RM

Therefore, let us focus on RM from now on in this talk

Heavy-Tailed Noise

$$\mathbb{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^{\boxed{\alpha}}\right] \leq \sigma^{\boxed{\alpha}}$$

$$\boxed{1 < \alpha \leq 2}$$

🫣 When $\alpha < 2$ variance can be **unbounded**

Heavy-Tailed Noise

$$\mathbb{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^{\boxed{\alpha}}\right] \leq \sigma^{\boxed{\alpha}}$$

$$\boxed{1 < \alpha \leq 2}$$

🫣 When $\alpha < 2$ variance can be **unbounded**

🫣 SGD can diverge:

$$\|x^1 - x^*\|^2 = \|x^0 - x^*\|^2 - 2\gamma_0\langle x^0 - x^*, \nabla f_{\xi^0}(x^0)\rangle + \|\nabla f_{\xi^0}(x^0)\|^2$$

SGD: $x^{k+1} = x^k - \gamma_k \nabla f_{\xi^k}(x^k)$

Heavy-Tailed Noise

$$\mathbb{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^{\boxed{\alpha}}\right] \leq \sigma^{\boxed{\alpha}}$$

$$\boxed{1 < \alpha \leq 2}$$

🫣 When $\alpha < 2$ variance can be **unbounded**

🫣 SGD can diverge:

$$\underbrace{\boxed{\mathbb{E}\|x^1 - x^*\|^2}}_{\text{Unbounded}} = \|x^0 - x^*\|^2 - \mathbb{E}[2\gamma_0\langle x^0 - x^*, \nabla f_{\xi^0}(x^0)\rangle] + \underbrace{\boxed{\mathbb{E}\|\nabla f_{\xi^0}(x^0)\|^2}}_{\text{Unbounded}}$$

SGD: $x^{k+1} = x^k - \gamma_k \nabla f_{\xi^k}(x^k)$

Heavy-Tailed Noise

$$\mathbb{E}\left[\|\nabla f_{\xi}(x) - \nabla f(x)\|^{\boxed{\alpha}}\right] \leq \sigma^{\boxed{\alpha}}$$

$$\boxed{1 < \alpha \leq 2}$$

When $\alpha < 2$ variance can be unbounded

💡 Gradient clipping fixes SGD!

SGD can diverge:

$$\underbrace{\boxed{\mathbb{E}\|x^1 - x^*\|^2}}_{\textbf{Unbounded}} = \|x^0 - x^*\|^2 - \mathbb{E}[2\gamma_0\langle x^0 - x^*, \nabla f_{\xi^0}(x^0)\rangle] + \underbrace{\boxed{\mathbb{E}\|\nabla f_{\xi^0}(x^0)\|^2}}_{\textbf{Unbounded}}$$

SGD: $x^{k+1} = x^k - \gamma_k \nabla f_{\xi^k}(x^k)$

SGD vs Clipped-SGD

SGD:

$$x^{k+1} = x^k - \gamma_k \nabla f_{\xi^k}(x^k)$$

Clipped-SGD:

$$x^{k+1} = x^k - \gamma_k \boxed{\text{clip}_{\lambda_k}\left(\nabla f_{\xi^k}(x^k)\right)}$$

$$\boxed{\text{clip}_\lambda(x) = \min\left\{1, \frac{\lambda}{\|x\|}\right\} x}$$

R. Pascanu et al. On the difficulty of training recurrent neural networks. (ICML 2013).

High-Probability Convergence

In-Expectation vs High-Probability Guarantees

In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$

- 😐 Typically, depend only on some moments of stochastic gradient, e.g., variance

In-Expectation vs High-Probability Guarantees

In-expectation guarantees: $\mathbb{E}[\|x - x^*\|^2] \leq \varepsilon$, $\mathbb{E}[f(x) - f(x^*)] \leq \varepsilon$, $\mathbb{E}[\|\nabla f(x)\|^2] \leq \varepsilon$

😐 Typically, depend only on some moments of stochastic gradient, e.g., variance

High-probability guarantees: $\mathbb{P}\{\|x - x^*\|^2 \leq \varepsilon\} \geq 1 - \beta$, $\mathbb{P}\{f(x) - f(x^*) \leq \varepsilon\} \geq 1 - \beta$,
$\mathbb{P}\{\|\nabla f(x)\|^2 \leq \varepsilon\} \geq 1 - \beta$

👌 Sensitive to the distribution of the stochastic gradient noise
❗ Harder to obtain with *logarithmic dependence* on $1/\beta$

High-probability results give better understanding of methods behavior

Convergence of SGD: Toy Example

Problem: $$f(x) = \frac{1}{2}\|x\|^2 \quad \text{and} \quad f_\xi(x) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle$$

Convergence of SGD: Toy Example

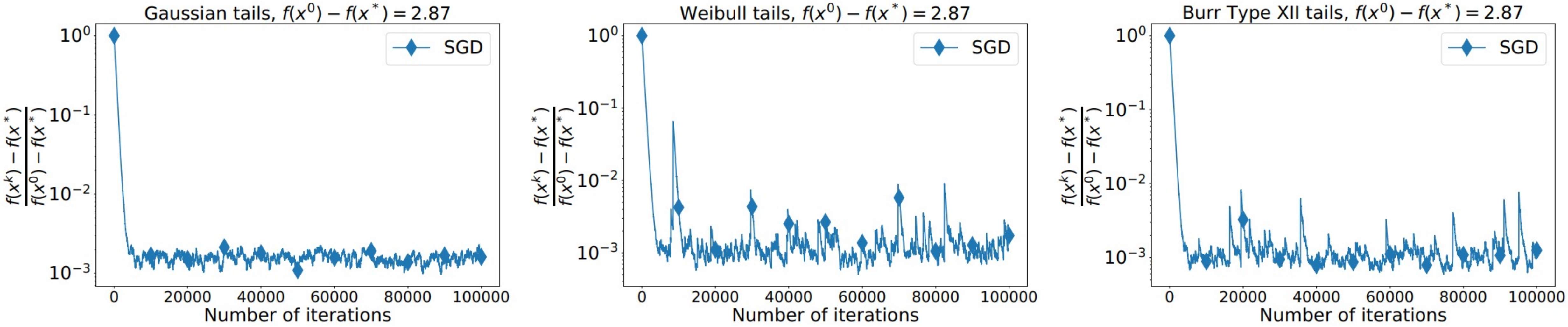
Problem: $f(x) = \frac{1}{2}\|x\|^2$ and $f_\xi(x) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle$

Convergence: $\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k(f(x^0) - f(x^*)) + \frac{\gamma\sigma^2}{2}$

Convergence of SGD: Toy Example

Problem: $$f(x) = \frac{1}{2}\|x\|^2 \quad \text{and} \quad f_\xi(x) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle$$

Convergence: $$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k (f(x^0) - f(x^*)) + \frac{\gamma\sigma^2}{2}$$

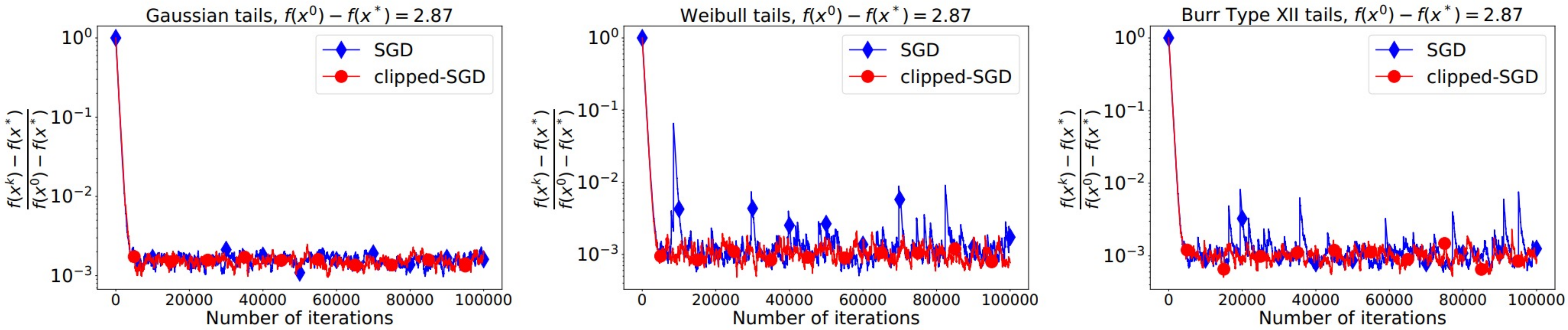


SGD's behavior does depend on the distribution but it is not reflected by in-expectation guarantees!

Convergence of SGD and Clipped-SGD: Toy Example

Problem: $$f(x) = \frac{1}{2}\|x\|^2 \quad \text{and} \quad f_\xi(x) = \frac{1}{2}\|x\|^2 + \langle \xi, x \rangle$$

Convergence: $$\mathbb{E}\left[f(x^k) - f(x^*)\right] \leq (1-\gamma)^k (f(x^0) - f(x^*)) + \frac{\gamma\sigma^2}{2}$$



SGD's behavior does depend on the distribution but it is not reflected by in-expectation guarantees!

Clipped-SGD oscillates less around the same value

Some Recent Advances on High-Probability Convergence

- Nazin et al. Algorithms of robust stochastic optimization based on mirror descent method. (Automation and Remote Control, 2019)
- Davis et al. From low probability to high confidence in stochastic convex optimization. (JMLR 2021)
- Gorbunov et al. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. (NeurIPS 2020)
- Cutkosky & Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. (NeurIPS 2021)
- Gorbunov et al. Clipped stochastic methods for variational inequalities with heavy-tailed noise. (NeurIPS 2022)
- Sadiev et al. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. (ICML 2023)
- Nguyen et al. High probability convergence of Clipped-SGD under heavy-tailed noise. (arXiv:2302.05437)
- Liu et al. High probability convergence of stochastic gradient methods. (ICML 2023)
- Nguyen et al. Improved convergence in high probability of clipped gradient methods with heavy tails. (NeurIPS 2023)
- Liu & Zhou. Stochastic Nonsmooth convex optimization with heavy-tailed noises: high-probability bound, in-expectation rate and initial distance adaptation. (arXiv:2303.12277)
- Puchkin et al. Breaking the heavy-tailed noise barrier in stochastic optimization problems. (AISTATS 2024)

Some Recent Advances on High-Probability Convergence

- Nazin et al. Algorithms of robust stochastic optimization based on mirror descent method. (Automation and Remote Control, 2019)
- Davis et al. From low probability to high confidence in stochastic convex optimization. (JMLR 2021)
- Gorbunov et al. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. (NeurIPS 2020)
- Cutkosky & Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. (NeurIPS 2021)
- Gorbunov et al. Clipped stochastic methods for variational inequalities with heavy-tailed noise. (NeurIPS 2022)
- Sadiev et al. High-probability bounds for stochastic optimization and variational inequalities: the case of unbounded variance. (ICML 2023)
- Nguyen et al. High probability convergence of Clipped-SGD under heavy-tailed noise. (arXiv:2302.05437)
- Liu et al. High probability convergence of stochastic gradient methods. (ICML 2023)
- Nguyen et al. Improved convergence in high probability of clipped gradient methods with heavy tails. (NeurIPS 2023)
- Liu & Zhou. Stochastic Nonsmooth convex optimization with heavy-tailed noises: high-probability bound, in-expectation rate and initial distance adaptation. (arXiv:2303.12277)
- Puchkin et al. Breaking the heavy-tailed noise barrier in stochastic optimization problems. (AISTATS 2024)

😐 **Analysis for composite and distributed problems is limited!**

Composite Optimizaton

Stochastic Composite Optimization

$$\min_{x \in \mathbb{R}^d} \left\{ \Phi(x) := f(x) + \Psi(x) \right\}$$

Stochastic Composite Optimization

$$\min_{x \in \mathbb{R}^d} \left\{ \Phi(x) := \boxed{f(x)} + \Psi(x) \right\}$$

Convex and smooth function

Stochastic gradients $\nabla f_\xi(x)$ are available

Stochastic Composite Optimization

$$\min_{x \in \mathbb{R}^d} \left\{ \Phi(x) := \boxed{f(x)} + \boxed{\Psi(x)} \right\}$$

Convex and smooth function
Stochastic gradients $\nabla f_\xi(x)$ are available

"Simple" function (proper, closed, and convex)
Prox-operator (a.k.a. projection) is computable

$$\boxed{\mathrm{prox}_\Psi(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \Psi(y) + \frac{1}{2}\|y - x\|^2 \right\}}$$

Stochastic Composite Optimization: Examples

- Regularized risk minimization

$$\min_{x\in\mathbb{R}^d} \left\{ \Phi(x) := \underbrace{\mathbb{E}_{\xi\sim\mathcal{D}}[f_\xi(x)]}_{f(x)} + \underbrace{\lambda_1\|x\|_1 + \lambda_2\|x\|_2^2}_{\Psi(x)} \right\}$$

$$\min_{x\in\mathbb{R}^d} \{\Phi(x) := f(x) + \Psi(x)\}$$

Stochastic Composite Optimization: Examples

- Regularized risk minimization

$$\min_{x \in \mathbb{R}^d} \left\{ \Phi(x) := \underbrace{\mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)]}_{f(x)} + \underbrace{\lambda_1 \|x\|_1 + \lambda_2 \|x\|_2^2}_{\Psi(x)} \right\}$$

- Constrained risk minimization

closed convex set

$$\min_{x \in \mathbb{R}^d} \left\{ \Phi(x) := \underbrace{\mathbb{E}_{\xi \sim \mathcal{D}}[f_\xi(x)]}_{f(x)} + \Psi(x) \right\}, \quad \Psi(x) := \begin{cases} 0, & \text{if } x \in \boxed{\mathcal{X}} \\ +\infty, & \text{if } x \notin \boxed{\mathcal{X}} \end{cases}$$

$$\min_{x \in \mathbb{R}^d} \{\Phi(x) := f(x) + \Psi(x)\}$$

Stochastic Composite Optimization: Examples

- Distributed optimization

$$\min_{\mathbf{X}=[x_1,\ldots,x_n]\in\mathbb{R}^{d\times n}} \left\{ \Phi(\mathbf{X}) := \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(x_i)}_{f(\mathbf{X})} + \Psi(\mathbf{X}) \right\}$$

$$\Psi(\mathbf{X}) := \begin{cases} 0, & \text{if } x_1 = \ldots = x_n \\ +\infty, & \text{otherwise} \end{cases}$$

  - n workers/clients are connected with a parameter-server
  - $f_i(x_i)$ - loss on the data available on client i

Stochastic Composite Optimization: Examples

- Distributed optimization

$$\min_{\mathbf{X}=[x_1,\ldots,x_n]\in\mathbb{R}^{d\times n}} \left\{ \Phi(\mathbf{X}) := \underbrace{\frac{1}{n}\sum_{i=1}^{n} f_i(x_i)}_{f(\mathbf{X})} + \Psi(\mathbf{X}) \right\}$$

$$\Psi(\mathbf{X}) := \begin{cases} 0, & \text{if } x_1 = \ldots = x_n \\ +\infty, & \text{otherwise} \end{cases}$$

- n workers/clients are connected with a parameter-server
- $f_i(x_i)$ - loss on the data available on client i

In our work, we consider an explicit form of the distributed problem

Standard Method: Prox-SGD

$$x^{k+1} = \qquad\qquad x^k - \gamma_k \nabla f_{\xi^k}(x^k)$$

Standard Method: Prox-SGD

$$x^{k+1} = \boxed{\mathrm{prox}_{\gamma_k \Psi}}\left(x^k - \gamma_k \nabla f_{\xi^k}(x^k)\right)$$

Standard Method: Prox-SGD

$$x^{k+1} = \boxed{\mathrm{prox}_{\gamma_k \Psi}} \left(x^k - \gamma_k \nabla f_{\xi^k}(x^k) \right)$$

$$\boxed{\mathrm{prox}_{\Psi}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \Psi(y) + \frac{1}{2}\|y - x\|^2 \right\}}$$

Standard Method: Prox-SGD

$$x^{k+1} = \boxed{\mathrm{prox}_{\gamma_k \Psi}} \left(x^k - \gamma_k \nabla f_{\xi^k}(x^k)\boxed{\right)}$$

🤔 Just clip stochastic gradient?

$$\boxed{\mathrm{prox}_{\Psi}(x) := \arg\min_{y \in \mathbb{R}^d} \left\{ \Psi(y) + \frac{1}{2}\|y - x\|^2 \right\}}$$

Failure of the Naïve Approach

Proximal Clipped-SGD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \boxed{\mathrm{clip}_{\lambda_k}(\nabla f_{\xi^k}(x^k))}\right)$$

$$\boxed{\mathrm{clip}_{\lambda}(x) = \min\left\{1, \frac{\lambda}{\|x\|}\right\} x}$$

There is an issue with this method related to the choice of λ_k

Closer Look at the Deterministic Case

Prox-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \nabla f(x^k)\right)$$

Prox-clipped-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^k))\right)$$

Closer Look at the Deterministic Case

Prox-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \nabla f(x^k)\right)$$

Solution is a fixed-point:

$$x^* = \mathrm{prox}_{\gamma_k \Psi}\left(x^* - \gamma_k \nabla f(x^*)\right)$$

No need to decrease stepsizes

Prox-clipped-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^k))\right)$$

Closer Look at the Deterministic Case

Prox-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \nabla f(x^k)\right)$$

Solution is a fixed-point:

$$x^* = \mathrm{prox}_{\gamma_k \Psi}\left(x^* - \gamma_k \nabla f(x^*)\right)$$

No need to decrease stepsizes

Prox-clipped-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^k))\right)$$

Solution is not necessarily a fixed point :

$$\boxed{x^* \neq \mathrm{prox}_{\gamma_k \Psi}\left(x^* - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^*))\right)}$$

This can happen if $\|\nabla f(x^*)\| > \lambda_k$ for all $k \geq k_0$ since

$$\boxed{-\mathrm{clip}_{\lambda_k}(\nabla f(x^*)) \notin \partial \Psi(x^*)}$$

Closer Look at the Deterministic Case

Prox-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \nabla f(x^k)\right)$$

Solution is a fixed-point:

$$x^* = \mathrm{prox}_{\gamma_k \Psi}\left(x^* - \gamma_k \nabla f(x^*)\right)$$

No need to decrease stepsizes

Prox-clipped-GD

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^k))\right)$$

Solution is not necessarily a fixed point :

$$\boxed{x^* \neq \mathrm{prox}_{\gamma_k \Psi}\left(x^* - \gamma_k \mathrm{clip}_{\lambda_k}(\nabla f(x^*))\right)}$$

This can happen if $\|\nabla f(x^*)\| > \lambda_k$ for all $k \geq k_0$ since

$$\boxed{-\mathrm{clip}_{\lambda_k}(\nabla f(x^*)) \notin \partial \Psi(x^*)}$$

In the stochastic case, known results for unconstrained problems require decreasing λ_k for tight convergence in the strongly convex case and acceleration!

Non-Implementable Fix

New Method: Proximal Clipped-SGD-star

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k(\nabla f(x^*) + \mathrm{clip}_{\lambda_k}(\Delta_k))\right)$$

New Method: Proximal Clipped-SGD-star

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k(\nabla f(x^*) + \mathrm{clip}_{\lambda_k}(\boxed{\Delta_k}))\right)$$

$$\boxed{\Delta_k = \nabla f_{\xi^k}(x^k) - \nabla f(x^*)}$$

Solution is a fixed-point for any choice of λ_k (in the special case of deterministic gradients)

Provable convergence (we have proofs)

New Method: Proximal Clipped-SGD-star

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k\left(\boxed{\nabla f(x^*)} + \mathrm{clip}_{\lambda_k}\left(\boxed{\Delta_k}\right)\right)\right)$$

$$\boxed{\Delta_k = \nabla f_{\xi^k}(x^k) - \boxed{\nabla f(x^*)}}$$

Solution is a fixed-point for any choice of λ_k

Provable convergence (we have proofs)

The method cannot be used: $\nabla f(x^*)$ is unknown in general

Learnable Shifts

New Method: Proximal Clipped-SGD with Shift

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k(\boxed{h^k} + \mathrm{clip}_{\lambda_k}(\Delta_k))\right)$$

learnable shift

$$\Delta_k = \nabla f_{\xi^k}(x^k) - \boxed{h^k}$$

New Method: Proximal Clipped-SGD with Shift

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k\left(\boxed{h^k} + \mathrm{clip}_{\lambda_k}(\Delta_k)\right)\right)$$

learnable shift

$$\Delta_k = \nabla f_{\xi^k}(x^k) - \boxed{h^k}$$

$$\boxed{h^{k+1} = h^k + \nu \cdot \mathrm{clip}_{\lambda_k}(\Delta_k)}$$

New Method: Proximal Clipped-SGD with Shift

$$x^{k+1} = \mathrm{prox}_{\gamma_k \Psi}\left(x^k - \gamma_k\left(\boxed{h^k} + \mathrm{clip}_{\lambda_k}(\Delta_k)\right)\right)$$

learnable shift

$$\Delta_k = \nabla f_{\xi^k}(x^k) - \boxed{h^k}$$

$$\boxed{h^{k+1} = h^k + \nu \cdot \mathrm{clip}_{\lambda_k}(\Delta_k)}$$

Intuition: one step of clipped-SGD applied to

$$\boxed{\min_{h \in \mathbb{R}^d} \frac{1}{2}\|h - \nabla f_{\xi^k}(x^k)\|^2}$$

where $\nabla f_{\xi^k}(x^k)$ can be seen as

a noisy estimate of $\nabla f(x^*)$

h^k approximates $\nabla f(x^*)$

Provable convergence (we have proofs)

Convergence Results: Convex Case

Assumptions

- Convexity $$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$
- Smoothness $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Convergence rate

There exists a choice of stepsizes γ and ν and clipping level λ such that with probability at least $1 - \beta$

$$\Phi(\overline{x}^K) - \Phi(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2A}{K}, \frac{R\zeta_*A}{K}, \frac{\sigma RA^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right)$$

Convergence Results: Convex Case

Assumptions

- Convexity $$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$
- Smoothness $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Convergence rate

There exists a choice of stepsizes γ and ν and clipping level λ such that with probability at least $1 - \beta$

$$\Phi(\overline{x}^K) - \Phi(x^*) = \mathcal{O}\left(\max\left\{\frac{LR^2A}{K}, \frac{R\zeta_* A}{K}, \frac{\sigma R A^{\frac{\alpha-1}{\alpha}}}{K^{\frac{\alpha-1}{\alpha}}}\right\}\right)$$

R – an upper bound on $\|x^0 - x^*\|$, $\zeta_* = \|\nabla f(x^*)\|$, $A = \log\frac{4K}{\beta}$

Logarithmic dependence on β

The rate matches the one for clipped-SGD in the unconstrained case

Convergence Results: Strongly Convex Case

Assumptions

- Strong convexity $$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle + \frac{\mu}{2}\|x - y\|^2$$
- Smoothness $$\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Convergence rate

There exists a choice of stepsizes γ and ν and clipping level λ_k such that with probability at least $1 - \beta$

$$\|x^K - x^*\|^2 = \mathcal{O}\left(\max\left\{R^2 \exp\left(-\frac{\mu K}{LA}\right), R^2 \exp\left(-\frac{\mu R K}{\zeta_* A}\right), \frac{\sigma^2 A^{\frac{2(\alpha-1)}{\alpha}} B}{K^{\frac{2(\alpha-1)}{\alpha}}}\right\}\right)$$

R – an upper bound on $\|x^0 - x^*\|$, $\zeta_* = \|\nabla f(x^*)\|$, $A = \log\frac{K}{\beta}$, B – another logarithmic factor

Logarithmic dependence on β

The rate matches the one for clipped-SGD in the unconstrained case

Extensions and Generalizations

In the paper, we also have

💎 Accelerated rates

💎 Linear speed up for distributed composite problems (even for $\alpha < 2$)

$$\mathbb{E}\left[\|\nabla f_\xi(x) - \nabla f(x)\|^\alpha\right] \le \sigma^\alpha$$

💎 Generalization to the variational inequalities

⚙️ Detailed proofs (with novel Lyapunov function for accelerated method)

Conclusion

Conclusion

Main takeaway:

clip gradient differences for better high-probability convergence
for composite and distributed problems

Come to our poster for more details: Today, 11:30 am (Hall C 4-9 #1014)

Paper:

My website:

(I am on the job market)