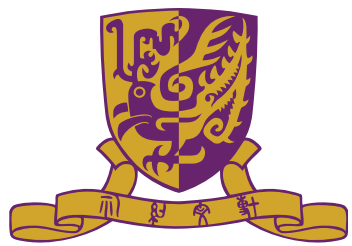
A Unified Convergence Theorem for Stochastic Optimization Methods

Xiao Li, Andre Milzarek

School of Data Science
The Chinese University of Hong Kong, Shenzhen



Acknowledgments

- Full paper is in the proceeding of NeurIPS 2022 and is available at

  https://arxiv.org/abs/2206.03907

- Highly appreciate funding resources that supported this work, such as NSFC, Shenzhen Science and Technology Program, AIRS, etc.

Motivation: Convergence of SGD — I

Suppose we apply SGD to minimize the smooth nonconvex problem:

$$\min_x \ f(x)$$

- SGD uses stochastic gradient approximations $\boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k)$ to generate a stochastic process $\{\boldsymbol{x}^k\}_k$.
- Convergence results for SGD are typically expressed as (non-asymptotic) complexity bounds of the form:

$$\min_{k \leq T} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \quad \mathbb{E}[\|\nabla f(\boldsymbol{x}^{k_\circ})\|^2] \leq \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \quad (1)$$

  where T is the total number of iterations and $k_\circ$ is an index sampled uniformly at random from $\{0, \ldots, T\}$. (Ghadimi and Lan '13, ...).
- Such type of complexity bounds are ubiquitous in stochastic optimization.

Motivation: Convergence of SGD — I

Suppose we apply SGD to minimize the smooth nonconvex problem:

$$\min_x \ f(x)$$

- SGD uses stochastic gradient approximations $\boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k)$ to generate a stochastic process $\{\boldsymbol{x}^k\}_k$.
- Convergence results for SGD are typically expressed as (non-asymptotic) complexity bounds of the form:

$$\min_{k \le T} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right), \quad \mathbb{E}[\|\nabla f(\boldsymbol{x}^{k_\circ})\|^2] \le \mathcal{O}\left(\frac{1}{\sqrt{T}}\right) \quad (1)$$

  where T is the total number of iterations and $k_\circ$ is an index sampled uniformly at random from $\{0, \ldots, T\}$. (Ghadimi and Lan '13, ...).
- Such type of complexity bounds are ubiquitous in stochastic optimization.

Motivation: Convergence of SGD — II

A First Observation:

- Taking the limit $T \to \infty$ in (1) does not allow to directly establish the asymptotic results:

$$\lim_{k\to\infty} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|] = 0 \quad \text{or} \quad \lim_{k\to\infty} \|\nabla f(\boldsymbol{x}^k)\| = 0 \text{ almost surely.}$$

$\rightsquigarrow$ Such results ensure that limit points of $\{\boldsymbol{x}^k\}_k$ are stationary point of f — more specifically:

Accumulation points of $\{\boldsymbol{x}^k\}_k$ are, almost surely, stationary points of f

Remarks:

- Asymptotic convergence results complement complexity bounds and are a first step towards understanding the limit behavior of $\{\boldsymbol{x}^k\}_k$.
- Typically considered as "sanity and soundness check" in deterministic optimization.

Motivation: Convergence of SGD — II

A First Observation:

- Taking the limit $T \to \infty$ in (1) does not allow to directly establish the asymptotic results:

$$\lim_{k\to\infty} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|] = 0 \quad \text{or} \quad \lim_{k\to\infty} \|\nabla f(\boldsymbol{x}^k)\| = 0 \text{ almost surely.}$$

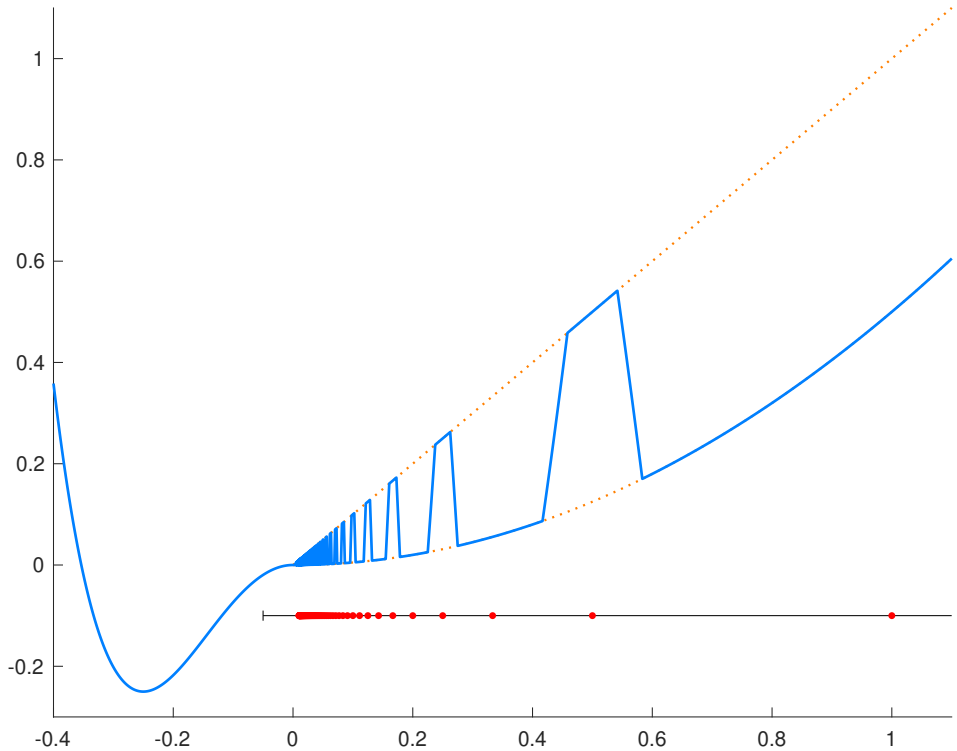
$\rightsquigarrow$ Such results ensure that limit points of $\{\boldsymbol{x}^k\}_k$ are stationary point of f — more specifically:

Accumulation points of $\{\boldsymbol{x}^k\}_k$ are, almost surely, stationary points of f

Remarks:

- Asymptotic convergence results complement complexity bounds and are a first step towards understanding the limit behavior of $\{\boldsymbol{x}^k\}_k$.
- Typically considered as "sanity and soundness check" in deterministic optimization.

Example: Last Iterate Convergence



- Returning the last iterate as output is a common policy in stochastic optimization.
- Complexity results often do not fully support/explain such strategy.

Towards a Unified Convergence Theorem — I

Literature and Discussion:

- (Non-)Asymptotic convergence of SGD is well-studied (Robbins and Monro '51, Ljung '77, '78, Kushner and Clark '78, Polyak and Juditsky '92, Benaïm '96, Bertsekas and Tsitsklis '00, Kushner and Yin '03, Bottou et al. '18, Orabona '20, Mertikopoulus et al. '20, ...).
- As direct transition from complexity results to full asymptotic convergence is not possible, a separate analysis is often required!
- Results are scattered ($\rightsquigarrow$ different assumptions on f and the stochastic oracles).

Our Goal:

- Provide a novel, simple, and unified convergence theorem for deriving expected and almost sure convergence of stochastic methods.

$\rightsquigarrow$ Facilitate the derivation of asymptotic convergence properties.

Towards a Unified Convergence Theorem — I

Literature and Discussion:

- (Non-)Asymptotic convergence of SGD is well-studied (Robbins and Monro '51, Ljung '77, '78, Kushner and Clark '78, Polyak and Juditsky '92, Benaïm '96, Bertsekas and Tsitsklis '00, Kushner and Yin '03, Bottou et al. '18, Orabona '20, Mertikopoulus et al. '20, ...).
- As direct transition from complexity results to full asymptotic convergence is not possible, a separate analysis is often required!
- Results are scattered ($\rightsquigarrow$ different assumptions on f and the stochastic oracles).

Our Goal:

- Provide a novel, simple, and unified convergence theorem for deriving expected and almost sure convergence of stochastic methods.

$\rightsquigarrow$ Facilitate the derivation of asymptotic convergence properties.

Towards a Unified Convergence Theorem — II

Key Points:

- The theorem is not tailored to any specific algorithm.
- Several abstract conditions are formulated that suit a vast, general, and modern class of problem structures and algorithms.

$\rightsquigarrow$ Facilitate the derivation of asymptotic convergence properties.

Starting Point:

- Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_k, \mathbb{P})$ be a filtered probability space and let the iterates $\{\boldsymbol{x}^k\}_k$ be adapted to the filtration $\{\mathcal{F}_k\}_k$.
- We assume that there is an associated abstract convergence measure $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ and step sizes $\{\mu_k\}_k$. (Think about $\Phi = \nabla f$)

$\rightsquigarrow$ We now formulate conditions on $\{\boldsymbol{x}^k\}_k$, $\{\mu_k\}_k$, and Φ that ensure convergence of $\mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|]$ and $\|\Phi(\boldsymbol{x}^k)\|$.

Towards a Unified Convergence Theorem — II

Key Points:

- The theorem is not tailored to any specific algorithm.
- Several abstract conditions are formulated that suit a vast, general, and modern class of problem structures and algorithms.

$\rightsquigarrow$ Facilitate the derivation of asymptotic convergence properties.

Starting Point:

- Let $(\Omega, \mathcal{F}, \{\mathcal{F}_k\}_k, \mathbb{P})$ be a filtered probability space and let the iterates $\{\boldsymbol{x}^k\}_k$ be adapted to the filtration $\{\mathcal{F}_k\}_k$.
- We assume that there is an associated abstract convergence measure $\Phi : \mathbb{R}^n \to \mathbb{R}^m$ and step sizes $\{\mu_k\}_k$. (Think about $\Phi = \nabla f$)

$\rightsquigarrow$ We now formulate conditions on $\{\boldsymbol{x}^k\}_k$, $\{\mu_k\}_k$, and Φ that ensure convergence of $\mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|]$ and $\|\Phi(\boldsymbol{x}^k)\|$.

Main Results I

Basic Conditions

(P.1) The function Φ is L_Φ-Lipschitz continuous for some $\mathsf{L}_\Phi > 0$.

(P.2) There is $a > 0$ such that $\sum_{k=0}^{\infty} \mu_k \, \mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|^a] < \infty$.

Conditions for Convergence in Expectation

(P.3) There exist constants $\mathsf{A}, \mathsf{B}, b \geq 0$ and $p_1, p_2, q > 0$ such that

$$\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^q] \leq \mathsf{A}\mu_k^{p_1} \; + \; \mathsf{B}\mu_k^{p_2} \, \mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|^b].$$

(P.4) The sequence $\{\mu_k\}_k$ and the parameters a, b, q, p_1, p_2 satisfy

$$\mu_k \to 0, \;\; \sum_{k=0}^{\infty} \mu_k = \infty, \;\; \text{and} \;\; a, q \geq 1, \;\; a \geq b, \;\; p_1, p_2 \geq q.$$

Theorem: Convergence in Expectation

Let the sequences $\{\boldsymbol{x}^k\}_k$ and $\{\mu_k\}_k$ be given and assume **(P.1)**–**(P.4)**. Then, we have $\mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|] \to 0$ as $k \to \infty$.

Main Results II

Conditions for Almost Sure Convergence

(P.3′) There exist constants $\mathsf{A}, b \geq 0$, $p_1, p_2, q > 0$ and random vectors $\boldsymbol{A}_k, \boldsymbol{B}_k : \Omega \to \mathbb{R}^n$ such that

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \mu_k^{p_1} \boldsymbol{A}_k + \mu_k^{p_2} \boldsymbol{B}_k$$

and we have $\mathbb{E}[\boldsymbol{A}_k \mid \mathcal{F}_k] = 0$ almost surely, $\mathbb{E}[\|\boldsymbol{A}_k\|^q] \leq \mathsf{A}$ (for all k), and $\limsup_{k\to\infty} \|\boldsymbol{B}_k\|^q/(1 + \|\Phi(\boldsymbol{x}^k)\|^b) < \infty$ almost surely.

(P.4′) The sequence $\{\mu_k\}_k$ and the parameters a, b, q, p_1, p_2 satisfy $q \geq 2$, $qa \geq b$, $p_1 > \frac{1}{2}$, $p_2 \geq 1$ and

$$\mu_k \to 0, \quad \sum_{k=0}^{\infty} \mu_k = \infty, \quad \sum_{k=0}^{\infty} \mu_k^{2p_1} < \infty.$$

Theorem: Almost Sure Convergence

Let the sequences $\{\boldsymbol{x}^k\}_k$ and $\{\mu_k\}_k$ be given and assume **(P.1)**–**(P.2)** and **(P.3′)**–**(P.4′)**. Then, it holds that $\|\Phi(\boldsymbol{x}^k)\| \to 0$ almost surely.

Application and Discussion I

The unified theorem can be applied using three main phases.

Phase I — Verify (P.1)–(P.2):

- Condition (**P.1**) is a problem property and is very standard.
- E.g., Φ can be set as ∇f in smooth problems or as the norm of the gradient of the Moreau envelope in weakly convex optimization, ...
- Lipschitz continuity is a minimal assumption in these situations.
- Condition (**P.2**) is linked to complexity results. Its derivation is often based on recursions of the form:

$$\mathbb{E}[\boldsymbol{y}_{k+1} \mid \mathcal{F}_k] \leq (1+\beta_k)\boldsymbol{y}_k - \mu_k \|\Phi(\boldsymbol{x}^k)\|^a + \zeta_k.$$

- Here, $\boldsymbol{y}_k$ is a suitable Lyapunov function measuring (approximate) descent; ζ_k and β_k are error terms with $\sum_{k=0}^{\infty} \max\{\zeta_k, \beta_k\} < \infty$.

$\rightsquigarrow$ Supermartingale convergence theorem: $\sum_{k=0}^{\infty} \mu_k \, \mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|^a] < \infty$ (Robbins and Siegmund '71).

Application and Discussion I

The unified theorem can be applied using three main phases.

Phase I — Verify (P.1)–(P.2):

- Condition (**P.1**) is a problem property and is very standard.
- E.g., Φ can be set as ∇f in smooth problems or as the norm of the gradient of the Moreau envelope in weakly convex optimization, ...
- Lipschitz continuity is a minimal assumption in these situations.

- Condition (**P.2**) is linked to complexity results. Its derivation is often based on recursions of the form:

$$\mathbb{E}[\boldsymbol{y}_{k+1} \mid \mathcal{F}_k] \leq (1+\beta_k)\boldsymbol{y}_k - \mu_k \|\Phi(\boldsymbol{x}^k)\|^a + \zeta_k.$$

- Here, $\boldsymbol{y}_k$ is a suitable Lyapunov function measuring (approximate) descent; ζ_k and β_k are error terms with $\sum_{k=0}^{\infty} \max\{\zeta_k, \beta_k\} < \infty$.

$\rightsquigarrow$ Supermartingale convergence theorem: $\sum_{k=0}^{\infty} \mu_k \, \mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|^a] < \infty$ (Robbins and Siegmund '71).

Application and Discussion I

The unified theorem can be applied using three main phases.

Phase I — Verify (P.1)–(P.2):

- Condition (**P.1**) is a problem property and is very standard.
- E.g., Φ can be set as ∇f in smooth problems or as the norm of the gradient of the Moreau envelope in weakly convex optimization, ...
- Lipschitz continuity is a minimal assumption in these situations.
- Condition (**P.2**) is linked to complexity results. Its derivation is often based on recursions of the form:

$$\mathbb{E}[\boldsymbol{y}_{k+1} \mid \mathcal{F}_k] \leq (1+\beta_k)\boldsymbol{y}_k - \mu_k \|\Phi(\boldsymbol{x}^k)\|^a + \zeta_k.$$

- Here, $\boldsymbol{y}_k$ is a suitable Lyapunov function measuring (approximate) descent; ζ_k and β_k are error terms with $\sum_{k=0}^{\infty} \max\{\zeta_k, \beta_k\} < \infty$.

$\rightsquigarrow$ Supermartingale convergence theorem: $\sum_{k=0}^{\infty} \mu_k \, \mathbb{E}[\|\Phi(\boldsymbol{x}^k)\|^a] < \infty$ (Robbins and Siegmund '71).

Application and Discussion II

Phase II — Verify (P.3)–(P.4):

- Condition (**P.4**) is a standard diminishing step sizes condition.
- Condition (**P.3**) is often related to certain bounded variance-type assumptions for analyzing stochastic methods.

Phase III — Verify (P.3′)–(P.4′):

- Condition (**P.4′**): Step sizes conditions ✓
- In condition (**P.3′**), the update is decomposed into a martingale term A_k and a bounded error term B_k.
- A helpful trick — decompose x^{k+1} as:

$$x^{k+1} = x^k + \mu_k \cdot \underbrace{\tfrac{1}{\mu_k}(x^{k+1} - x^k - \mathbb{E}[x^{k+1} - x^k \mid \mathcal{F}_k])}_{A_k} + \mu_k \cdot \underbrace{\tfrac{1}{\mu_k}\mathbb{E}[x^{k+1} - x^k \mid \mathcal{F}_k]}_{B_k}$$

  $\rightsquigarrow$ If we have $\mathbb{E}[\|x^{k+1} - x^k\|^q \mid \mathcal{F}_k] = \mathcal{O}(\mu_k^q)$ in an almost sure sense, condition (**P.3′**) holds with $p_1 = p_2 = 1$.

Application and Discussion II

Phase II — Verify (P.3)–(P.4):

- Condition (**P.4**) is a standard diminishing step sizes condition.
- Condition (**P.3**) is often related to certain bounded variance-type assumptions for analyzing stochastic methods.

Phase III — Verify (P.3′)–(P.4′):

- Condition (**P.4′**): Step sizes conditions ✓
- In condition (**P.3′**), the update is decomposed into a martingale term $\boldsymbol{A}_k$ and a bounded error term $\boldsymbol{B}_k$.
- A helpful trick — decompose $\boldsymbol{x}^{k+1}$ as:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k + \mu_k \cdot \underbrace{\tfrac{1}{\mu_k}(\boldsymbol{x}^{k+1} - \boldsymbol{x}^k - \mathbb{E}[\boldsymbol{x}^{k+1} - \boldsymbol{x}^k \mid \mathcal{F}_k])}_{\boldsymbol{A}_k} + \mu_k \cdot \underbrace{\tfrac{1}{\mu_k}\mathbb{E}[\boldsymbol{x}^{k+1} - \boldsymbol{x}^k \mid \mathcal{F}_k]}_{\boldsymbol{B}_k}$$

  - ⇝ If we have $\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^q \mid \mathcal{F}_k] = \mathcal{O}(\mu_k^q)$ in an almost sure sense, condition (**P.3′**) holds with $p_1 = p_2 = 1$.

Application: SGD

We consider SGD for $\min_{x \in \mathbb{R}^n} f(x)$. The iteration of SGD is given by:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k.$$

Here, $\boldsymbol{g}^k$ denotes a stochastic approximation of $\nabla f(\boldsymbol{x}^k)$.

Standard Assumptions for SGD

- **(A.1)** $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous on $\mathbb{R}^n$ with modulus $\mathsf{L} > 0$.
- **(A.2)** There is $\bar{f}$ such that $f(x) \geq \bar{f}$ for all $x \in \mathbb{R}^n$.
- **(A.3)** Each $\boldsymbol{g}^k$ defines an unbiased estimator of $\nabla f(\boldsymbol{x}^k)$, i.e., we have $\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k)$ a.s., and there are $\mathsf{C}, \mathsf{D} \geq 0$ with:

$$\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \mathsf{C}[f(\boldsymbol{x}^k) - \bar{f}] + \mathsf{D} \quad \text{a.s.} \quad \forall\, k.$$

- **(A.4)** The step sizes $\{\alpha_k\}_k$ satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

- (**A.3**) is a modern ABC-type condition for the variance (Lei et al. '19, Khaled and Richtárik '20), which is implied by **(A.1)**.

Application: SGD

We consider SGD for $\min_{x\in\mathbb{R}^n} f(x)$. The iteration of SGD is given by:

$$\boldsymbol{x}^{k+1} = \boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k.$$

Here, $\boldsymbol{g}^k$ denotes a stochastic approximation of $\nabla f(\boldsymbol{x}^k)$.

Standard Assumptions for SGD

(A.1) $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous on $\mathbb{R}^n$ with modulus $\mathsf{L} > 0$.

(A.2) There is $\bar{f}$ such that $f(x) \geq \bar{f}$ for all $x \in \mathbb{R}^n$.

(A.3) Each $\boldsymbol{g}^k$ defines an unbiased estimator of $\nabla f(\boldsymbol{x}^k)$, i.e., we have $\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k)$ a.s., and there are $\mathsf{C}, \mathsf{D} \geq 0$ with:

$$\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \mathsf{C}[f(\boldsymbol{x}^k) - \bar{f}] + \mathsf{D} \quad \text{a.s.} \quad \forall\, k.$$

(A.4) The step sizes $\{\alpha_k\}_k$ satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

- (**A.3**) is a modern ABC-type condition for the variance (Lei et al. '19, Khaled and Richtárik '20), which is implied by **(A.1)**.

Derivation for SGD I

We set $\Phi \equiv \nabla f$ and $\mu_k \equiv \alpha_k$.

Phase I — Verify (P.1)–(P.2):

- (**P.1**) follows from (**A.1**) with $\mathsf{L}_\Phi = \mathsf{L}$.
- Using (**A.1**)–(**A.3**) and the descent lemma, we can show

$$\mathbb{E}[f(\boldsymbol{x}^{k+1}) - \bar{f} \mid \mathcal{F}_k] \leq \left(1 + \frac{\mathsf{L}\mathsf{C}\alpha_k^2}{2}\right)[f(\boldsymbol{x}^k) - \bar{f}] - \alpha_k\left(1 - \frac{\mathsf{L}\alpha_k}{2}\right)\|\nabla f(\boldsymbol{x}^k)\|^2 + \frac{\mathsf{L}\mathsf{D}\alpha_k^2}{2}.$$

- Due to (**A.4**) and the supermartingale convergence theorem, we can infer $\sum_{k=0}^{\infty} \alpha_k \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] < \infty$ and $\{\mathbb{E}[f(\boldsymbol{x}^k)]\}_k$ converges.

$\rightsquigarrow$ (**P.2**) holds with $a = 2$.

Derivation for SGD II

Phase II — Verify (P.3)–(P.4):

- By the SGD-update and (**A.3**), we have:

$$\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2] \leq \alpha_k^2 \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \mathsf{C}\alpha_k^2 \mathbb{E}[f(\boldsymbol{x}^k) - \bar{f}] + \mathsf{D}\alpha_k^2.$$

- As $\{\mathbb{E}[f(\boldsymbol{x}^k)]\}_k$ converges, there is F with $\mathbb{E}[f(\boldsymbol{x}^k) - \bar{f}] \leq \mathsf{F}$ for all k.

$\rightsquigarrow$ (**P.3**) is satisfied with $q = 2$, $\mathsf{A} = \mathsf{CF} + \mathsf{D}$, $p_1 = p_2 = 2$, $\mathsf{B} = 1$, and $b = 2$.

- Condition (**P.3**): ✓

Phase III — Verify (P.3′)–(P.4′):

- We can set $\boldsymbol{A}_k := \nabla f(\boldsymbol{x}^k) - \boldsymbol{g}^k$, $\boldsymbol{B}_k := -\nabla f(\boldsymbol{x}^k)$, and $p_1 = p_2 = 1$.
- For $q = b = 2$, we have $\mathbb{E}[\boldsymbol{A}_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\boldsymbol{A}_k\|^2] \leq \mathsf{CF} + \mathsf{D}$.

$\rightsquigarrow$ Condition (**P.3′**) and (**P.4′**) hold.

Derivation for SGD II

Phase II — Verify (P.3)–(P.4):

- By the SGD-update and (**A.3**), we have:

$$\mathbb{E}[\|\boldsymbol{x}^{k+1} - \boldsymbol{x}^k\|^2] \leq \alpha_k^2 \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|^2] + \mathsf{C}\alpha_k^2 \mathbb{E}[f(\boldsymbol{x}^k) - \bar{f}] + \mathsf{D}\alpha_k^2.$$

- As $\{\mathbb{E}[f(\boldsymbol{x}^k)]\}_k$ converges, there is F with $\mathbb{E}[f(\boldsymbol{x}^k) - \bar{f}] \leq \mathsf{F}$ for all k.

$\rightsquigarrow$ (**P.3**) is satisfied with $q = 2$, $\mathsf{A} = \mathsf{CF} + \mathsf{D}$, $p_1 = p_2 = 2$, $\mathsf{B} = 1$, and $b = 2$.

- Condition (**P.3**): ✓

Phase III — Verify (P.3′)–(P.4′):

- We can set $\boldsymbol{A}_k := \nabla f(\boldsymbol{x}^k) - \boldsymbol{g}^k$, $\boldsymbol{B}_k := -\nabla f(\boldsymbol{x}^k)$, and $p_1 = p_2 = 1$.
- For $q = b = 2$, we have $\mathbb{E}[\boldsymbol{A}_k \mid \mathcal{F}_k] = 0$ and $\mathbb{E}[\|\boldsymbol{A}_k\|^2] \leq \mathsf{CF} + \mathsf{D}$.

$\rightsquigarrow$ Condition (**P.3′**) and (**P.4′**) hold.

Convergence of SGD

Corollary: Convergence of SGD

Let the stochastic process $\{\boldsymbol{x}^k\}_k$ be generated by SGD and let the conditions (**A.1**)–(**A.4**) hold. Then, we have $\lim_{k\to\infty} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|] = 0$ and $\lim_{k\to\infty} \|\nabla f(\boldsymbol{x}^k)\| = 0$ almost surely.

Remarks:

- Our unified theorem allows a plugin-type derivation.
- Largely based on existing complexity estimates (**P.2**) + additional bounds for the iterate differences "$\boldsymbol{x}^{k+1} - \boldsymbol{x}^k$".
- Similar asymptotic guarantees can be shown for random reshuffling.

Convergence of SGD

Corollary: Convergence of SGD

Let the stochastic process $\{\boldsymbol{x}^k\}_k$ be generated by SGD and let the conditions (**A.1**)–(**A.4**) hold. Then, we have $\lim_{k\to\infty} \mathbb{E}[\|\nabla f(\boldsymbol{x}^k)\|] = 0$ and $\lim_{k\to\infty} \|\nabla f(\boldsymbol{x}^k)\| = 0$ almost surely.

Remarks:

- Our unified theorem allows a plugin-type derivation.
- Largely based on existing complexity estimates (**P.2**) + additional bounds for the iterate differences "$\boldsymbol{x}^{k+1} - \boldsymbol{x}^k$".
- Similar asymptotic guarantees can be shown for random reshuffling.

A Second Application: prox-SGD

Next, we consider the initial composite-type model

$$\boxed{\min_x \quad f(x) + \varphi(x) =: \psi(x)}$$

and the well-known prox-SGD method

$$\boldsymbol{x}^{k+1} = \mathrm{prox}_{\alpha_k \varphi}(\boldsymbol{x}^k - \alpha_k \boldsymbol{g}^k),$$

where $\boldsymbol{g}^k \approx \nabla f(\boldsymbol{x}^k)$ is a stochastic approximation of $\nabla f(\boldsymbol{x}^k)$, $\{\alpha_k\}_k$ is a suitable step size sequence, and $\mathrm{prox}_{\alpha_k \varphi} : \mathbb{R}^n \to \mathbb{R}^n$,

$$\mathrm{prox}_{\alpha_k \varphi}(x) := \underset{\boldsymbol{y} \in \mathbb{R}^n}{\mathrm{argmin}} \ \ \varphi(y) + \frac{1}{2\alpha_k}\|x - y\|^2$$

is the proximity operator of φ. Our analysis is largely based on the Moreau envelope $\mathrm{env}_{\theta\psi}(x) := \min_{y \in \mathbb{R}^n} \ \ \psi(y) + \frac{1}{2\theta}\|x - y\|^2$ of ψ.

Related Works

- Results for general nonconvex f and under bounded variance conditions are surprisingly rare.

- (Davis and Drusvyatskiy '19) established one of the first complexity results for prox-SGD using the Moreau envelope.
- However, this complexity bound again cannot be easily extended to asymptotic convergence results.
- Earlier studies rely on variance reduction, see, e.g. (Ghadimi et al. '16, Reddi et al. '16).
- All existing asymptotic results ("accumulation points of $\{\boldsymbol{x}^k\}_k$ are stationary points of ψ") rely on differential inclusion approaches (Majewski et al. '18, Davis et al. '20, Geiersbach and Scarinci '21).

  $\rightsquigarrow$ This requires stringent (a.s.) boundedness assumptions on $\{\boldsymbol{x}^k\}_k$, Sard-type conditions, ...

- The convex case is much easier and has been covered, e.g., in (Ghadimi et al. '16, Atchadé et al. '17, Rosasco et al. '20).

Related Works

- Results for general nonconvex f and under bounded variance conditions are surprisingly rare.
- (Davis and Drusvyatskiy '19) established one of the first complexity results for prox-SGD using the Moreau envelope.
- However, this complexity bound again cannot be easily extended to asymptotic convergence results.
- Earlier studies rely on variance reduction, see, e.g. (Ghadimi et al. '16, Reddi et al. '16).
- All existing asymptotic results ("accumulation points of $\{x^k\}_k$ are stationary points of ψ") rely on differential inclusion approaches (Majewski et al. '18, Davis et al. '20, Geiersbach and Scarinci '21).

  $\rightsquigarrow$ This requires stringent (a.s.) boundedness assumptions on $\{x^k\}_k$, Sard-type conditions, ...
- The convex case is much easier and has been covered, e.g., in (Ghadimi et al. '16, Atchadé et al. '17, Rosasco et al. '20).

Related Works

- Results for general nonconvex f and under bounded variance conditions are surprisingly rare.
- (Davis and Drusvyatskiy '19) established one of the first complexity results for prox-SGD using the Moreau envelope.
- However, this complexity bound again cannot be easily extended to asymptotic convergence results.
- Earlier studies rely on variance reduction, see, e.g. (Ghadimi et al. '16, Reddi et al. '16).
- All existing asymptotic results ("accumulation points of $\{x^k\}_k$ are stationary points of ψ") rely on differential inclusion approaches (Majewski et al. '18, Davis et al. '20, Geiersbach and Scarinci '21).

  $\rightsquigarrow$ This requires stringent (a.s.) boundedness assumptions on $\{x^k\}_k$, Sard-type conditions, ...
- The convex case is much easier and has been covered, e.g., in (Ghadimi et al. '16, Atchadé et al. '17, Rosasco et al. '20).

Related Works

- Results for general nonconvex f and under bounded variance conditions are surprisingly rare.
- (Davis and Drusvyatskiy '19) established one of the first complexity results for prox-SGD using the Moreau envelope.
- However, this complexity bound again cannot be easily extended to asymptotic convergence results.
- Earlier studies rely on variance reduction, see, e.g. (Ghadimi et al. '16, Reddi et al. '16).
- All existing asymptotic results ("accumulation points of $\{\boldsymbol{x}^k\}_k$ are stationary points of ψ") rely on differential inclusion approaches (Majewski et al. '18, Davis et al. '20, Geiersbach and Scarinci '21).

$\rightsquigarrow$ This requires stringent (a.s.) boundedness assumptions on $\{\boldsymbol{x}^k\}_k$, Sard-type conditions, ...

- The convex case is much easier and has been covered, e.g., in (Ghadimi et al. '16, Atchadé et al. '17, Rosasco et al. '20).

Related Works

- Results for general nonconvex f and under bounded variance conditions are surprisingly rare.
- (Davis and Drusvyatskiy '19) established one of the first complexity results for prox-SGD using the Moreau envelope.
- However, this complexity bound again cannot be easily extended to asymptotic convergence results.
- Earlier studies rely on variance reduction, see, e.g. (Ghadimi et al. '16, Reddi et al. '16).
- All existing asymptotic results ("accumulation points of $\{\boldsymbol{x}^k\}_k$ are stationary points of ψ") rely on differential inclusion approaches (Majewski et al. '18, Davis et al. '20, Geiersbach and Scarinci '21).

$\rightsquigarrow$ This requires stringent (a.s.) boundedness assumptions on $\{\boldsymbol{x}^k\}_k$, Sard-type conditions, ...

- The convex case is much easier and has been covered, e.g., in (Ghadimi et al. '16, Atchadé et al. '17, Rosasco et al. '20).

Standard Assumptions for prox-SGD

We consider the following assumptions for prox-SGD:

Standard Assumptions for prox-SGD

(B.1) There is $\bar{f}$ such that $f(x) \geq \bar{f}$ for all $x \in \mathbb{R}^n$ and $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz continuous on $\mathbb{R}^n$ with modulus $\mathsf{L} > 0$.

(B.2) The function φ is τ-weakly convex, proper, lsc., and bounded from below on $\operatorname{dom} \varphi$, i.e., we have $\varphi(x) \geq \bar{\varphi}$ for all $x \in \operatorname{dom} \varphi$.

(B.3) The mapping φ is L_φ-Lipschitz on $\operatorname{dom} \boldsymbol{p}$.

(B.4) Each $\boldsymbol{g}^k$ defines an unbiased estimator of $\nabla f(\boldsymbol{x}^k)$, i.e., we have $\mathbb{E}[\boldsymbol{g}^k \mid \mathcal{F}_k] = \nabla f(\boldsymbol{x}^k)$ a.s., and there are $\mathsf{C}, \mathsf{D} \geq 0$ with:

$$\mathbb{E}[\|\boldsymbol{g}^k - \nabla f(\boldsymbol{x}^k)\|^2 \mid \mathcal{F}_k] \leq \mathsf{C}[f(\boldsymbol{x}^k) - \bar{f}] + \mathsf{D} \quad \text{a.s.} \quad \forall\, k.$$

(B.5) The step sizes $\{\alpha_k\}_k$ satisfy $\sum_{k=0}^{\infty} \alpha_k = \infty$ and $\sum_{k=0}^{\infty} \alpha_k^2 < \infty$.

- Analysis under the ABC-type condition (**B.4**) seems to be new.

Convergence of prox-SGD

Based on our convergence framework, we obtain:

Corollary: Convergence of prox-SGD

Let $\{\boldsymbol{x}^k\}_k$ be generated by prox-SGD and let the conditions (**B.1**)–(**B.5**) be satisfied. Then, it holds that $\lim_{k\to\infty} \mathbb{E}[\|F_{\text{nat}}(\boldsymbol{x}^k)\|] = 0$ and $\lim_{k\to\infty} \|F_{\text{nat}}(\boldsymbol{x}^k)\| = 0$ almost surely.

- This result is new to prox-SGD.
- Full asymptotic results for prox-SGD without boundedness conditions.
- Similar strong convergence results can be derived for stochastic model-based methods (Davis and Drusvyatskiy '19).

Conclusion

- Novel unified convergence theorem to derive expected and almost convergence results.
- Applicable to a vast class of stochastic methodologies under state-of-the-art assumptions.
- New insights and (simpler) convergence results for prox-SGD and SMM.

$\rightsquigarrow$ Our framework can serve as a plugin-type tool to facilitate the convergence analysis of stochastic methods.

Thank You

Thank you very much for your attention$\sim$