Sharp Analysis of Stochastic Optimization under Global Kurdyka-Łojasiewicz Inequality

Ilyas Fatkhullin [1 *] Jalal Etesami [2 *] Niao He [1] Negar Kiyavash [2]

[1] ETH Zurich [2] EPFL
[*] equal contribution

Problem Setting

Non-convex *stochastic* optimization problem:

$$\min_{x\in\mathbb{R}^d} f(x) := \mathbb{E}_{\xi\sim\mathcal{D}}\left[f_\xi(x)\right], \tag{1}$$

$f(\cdot)$ is smooth, i.e., $\|\nabla f(x) - \nabla f(y)\| \leq L\|x-y\|$ for all $x, y \in \mathbb{R}^d$, $f^* := \inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

Goal: find $\hat{x}$ such that $\mathbb{E}\left[f(\hat{x}) - f^*\right] \leq \varepsilon$.

Assumptions on Landscape

Kurdyka-Łojasiewicz (KŁ) [1]. There exists a continuous function $\phi : \mathbb{R}^+ \to \mathbb{R}^+$ such that $\phi(0) = 0$ and $\phi^2(\cdot)$ is convex, and

$$||\nabla f(x)|| \geq \phi\left(f(x) - f^*\right) \quad \text{for all } x \in \mathbb{R}^d.$$

Polyak-Łojasiewicz (α-PŁ) [2]. There exists $\alpha \in [1,2]$ and $\mu > 0$ such that

$$\|\nabla f(x)\|^\alpha \geq (2\mu)^{\alpha/2}\left(f(x) - f^*\right) \quad \text{for all } x \in \mathbb{R}^d.$$

Simple example: $f(x) = x^{\frac{\alpha}{\alpha-1}}$ is α-PŁ.

Applications: Reinforcement Learning, Optimal Control, GLMs.

Assumptions on Noise

Expected Smoothness of order k (k-ES) [3]. Stochastic gradient estimator $g_k(x,\xi)$ satisfies $\mathbb{E}\left[g_k(x,\xi)\right] = \nabla f(x)$ and

$$\mathbb{E}\left[\|g_k(x,\xi)\|^2\right] \leq 2A\cdot h\big(f(x) - f^*\big) + B\cdot ||\nabla f(x)||^2 + \frac{C}{b_k},$$

for all $x \in \mathbb{R}^d$, where $A, B, C > 0$, $h : \mathbb{R}^+ \to \mathbb{R}^+$ is a concave continuously differentiable with $h'(t) \geq 0$, $h(0) = 0$.

Bounded variance (BV) is a special case of k-ES with $A = 0$, $B = 1$ and $C = \sigma^2$, i.e.,

$$\mathbb{E}\left[\|g_k(x,\xi) - \nabla f(x)\|^2\right] \leq \frac{\sigma^2}{b_k}.$$

Examples of k-ES:

- mini-batch estimator $g_k(x,\xi) = \frac{1}{b_k}\sum_{i=1}^{b_k}\nabla f_{\xi^i}(x)$,
- compression in distributed optimization $g_k(x,\xi) = \mathcal{Q}\left(\nabla f_\xi(x)\right)$.

Main goal

Understand the sample complexity of first-order stochastic optimization methods under global KŁ and PŁ conditions.

Related work: Analysis of SGD: 1. $\alpha = 2$ and k-ES [3], 2. $\alpha \in [1,2]$ and BV [4]. Our analysis recovers the special cases in [3] and [4].

Dynamics of SGD under KŁ

Stochastic Gradient Descent (SGD):

$$x_{t+1} = x_t - \eta_t g_k(x_t, \xi_t).$$

Lemma 1. Let $f(\cdot)$ satisfy KŁ and $g_k(x,\xi)$ satisfy k-ES, then

$$\delta_{t+1} \leq \delta_t + a\eta^2\cdot h\big(\delta_t\big) - \frac{\eta}{2}\phi^2(\delta_t) + \frac{d\eta^2}{b},$$

where $\delta_t := \mathbb{E}\left[f(x_t) - f^\star\right]$, $a := LA$, $d := \frac{LC}{2}$, $\eta := \eta_t$.

Corollary 1. Let $f(\cdot)$ satisfy α-PŁ, $b_k = \Theta\left(k^\tau\right)$. Then the sample complexity of SGD is

$$T\cdot\sum_{k=0}^{K-1} b_k = \begin{cases}\mathcal{O}\left(\epsilon_f^{-\frac{4-\alpha}{\alpha}}\right) & \text{for } 0 \leq \tau \leq \frac{\gamma}{4-\alpha-\gamma},\\ \mathcal{O}\left(\epsilon_f^{-\frac{(4-\alpha-\gamma)(\tau+1)}{\alpha}}\right) & \text{for } \tau > \frac{\gamma}{4-\alpha-\gamma}.\end{cases}$$

Table 1: Sample complexity to achieve $\mathbb{E}\left[f(x_k) - f^\star\right] \leq \epsilon_f$, $\kappa = \mathcal{L}/\mu$.

Method	Online (1)	Finite sum (2)
GD	–	$n\kappa\left(\frac{1}{\epsilon_f}\right)^{\frac{2-\alpha}{\alpha}}$
SGD	$\frac{\kappa\sigma^2}{\mu}\left(\frac{1}{\epsilon_f}\right)^{\frac{4-\alpha}{\alpha}}$	$\frac{\kappa\sigma^2}{\mu}\left(\frac{1}{\epsilon_f}\right)^{\frac{4-\alpha}{\alpha}}$
PAGER	$\left(\frac{\sigma^2}{\mu} + \kappa^2\right)\left(\frac{1}{\epsilon_f}\right)^{\frac{2}{\alpha}}$	$n + \sqrt{n}\kappa\left(\frac{1}{\epsilon_f}\right)^{\frac{2-\alpha}{\alpha}}$

Table 2: Sample complexity to achieve $\mathbb{E}\left[dist\left(x, X^\star\right)\right] \leq \epsilon_x$, where $X^\star \neq \emptyset$ is the set of optimal points of $f(\cdot)$.

Method	Online (1)	Finite sum (2)
GD	–	$n\kappa\left(\frac{1}{\epsilon_x}\right)^{\frac{2-\alpha}{\alpha-1}}$
SGD	$\kappa\sigma^2\left(\frac{1}{\epsilon_x}\right)^{\frac{4-\alpha}{\alpha-1}}$	$\kappa\sigma^2\left(\frac{1}{\epsilon_x}\right)^{\frac{4-\alpha}{\alpha-1}}$
PAGER	$\left(\frac{\sigma^2}{\mu} + \kappa^2\right)\left(\frac{1}{\epsilon_x}\right)^{\frac{2}{\alpha-1}}$	$n + \sqrt{n}\kappa\left(\frac{1}{\epsilon_x}\right)^{\frac{2-\alpha}{\alpha-1}}$

Variance Reduction for α-PŁ

Additional assumption: ***k*-Average $\mathcal{L}$-smoothness (*k*-AS)**.

Let $g'_k(x,\xi) := \frac{1}{b'_k}\sum_{i=1}^{b'_k}\nabla f_{\xi^i}(x)$ and $g'_k(y,\xi) := \frac{1}{b'_k}\sum_{i=1}^{b'_k}\nabla f_{\xi^i}(y)$ be unbiased estimators of $\nabla f(\cdot)$ at points x and y, $\xi = (\xi^1, \ldots, \xi^{b'_k})$, let $\widetilde{\Delta}(x,y) := g'_k(x,\xi) - g'_k(y,\xi)$. There exists $\mathcal{L} \geq 0$ such that

$$\mathbb{E}\left[\left\|\widetilde{\Delta}(x,y) - \Delta(x,y)\right\|^2\right] \leq \frac{\mathcal{L}^2}{b'_k}\|x-y\|^2$$

for all $x, y \in \mathbb{R}^d$, where $\Delta(x,y) := \nabla f(x) - \nabla f(y)$.

- if $b'_k = n$ in finite sum case, then $\mathcal{L} = 0$,
- if each $\nabla f_{\xi^i}(x)$ is $\bar{L}$ Lipschitz, then k-AS holds with $\mathcal{L} \leq \bar{L}$.

Algorithm 1: PAGER (PAGE [5] with restarts)

for $k = 0, \ldots, K-1$ **do**
- $(x_0, g_0) \leftarrow (\bar{x}_k, \bar{g}_k)$
- $(\eta, p, b, b') \leftarrow (\eta_k, p_k, b_k, b'_k)$
- **for** $t = 0, \ldots, T_k - 1$ **do**
  - $x^{t+1} = x^t - \eta g_t$
  - Sample $\chi \sim$ Bernoulli(p)
  - $$g_{t+1} = \begin{cases}\frac{1}{b}\sum_{i=1}^{b}\nabla f_{\xi^i_{t+1}}(x_{t+1}) & \text{if } \chi = 1\\ g_t + \frac{1}{b'}\sum_{i=1}^{b'}\nabla f_{\xi^i_{t+1}}(x_{t+1}) - \frac{1}{b'}\sum_{i=1}^{b'}\nabla f_{\xi^i_{t+1}}(x_t) & \text{if } \chi = 0\end{cases}$$
- $(\bar{x}_{k+1}, \bar{g}_{k+1}) \leftarrow (x_{t+1}, g_{t+1})$

Return: $\bar{x}_K$

Online case

Theorem 1 Let α-PŁ, BV, and k-AS hold. Set the sequences in Algorithm 1 as $b'_k = \Theta\left(2^{\frac{(2-\alpha)k}{\alpha}}\right)$, $p_k = \Theta\left(2^{\frac{-(2-\alpha)k}{\alpha}}\right)$, $b_k = \Theta\left(2^{\frac{2k}{\alpha}}\right)$, $T_k = \Theta\left(2^{\frac{(2-\alpha)k}{\alpha}}\right)$, $\eta_k = \Theta(1)$. Then the sample complexity of PAGER is $\mathcal{O}\left(\left(\frac{\sigma^2}{\mu} + \kappa^2\right)\epsilon_f^{-\frac{2}{\alpha}}\right)$.

Finite sum case

$$\min_{x\in\mathbb{R}^d}\left[f(x) := \frac{1}{n}\sum_{i=1}^{n} f_i(x)\right]. \tag{2}$$

Theorem 2 Let $f(\cdot)$ have the form (2) and α-PŁ, BV, and k-AS hold. Set the sequences in Algorithm 1 as $p_k = \frac{1}{n+1}$, $b'_k = 1$, $b_k = n$, $T_k = \Theta\left(2^{\frac{(2-\alpha)k}{\alpha}}\right)$, $\eta_k = \Theta(1)$. Then the sample complexity of PAGER is $\widetilde{\mathcal{O}}\left(n + \sqrt{n}\,\kappa\,\epsilon_f^{-\frac{2-\alpha}{\alpha}}\right)$.

Implications and Discussion

- PAGER improves SGD for all $\alpha \in [1,2)$.
- For $\alpha = 1$, PAGER achieves $\mathcal{O}\left(\epsilon_f^{-2}\right)$ compared to $\mathcal{O}\left(\epsilon_f^{-3}\right)$ for SGD.
- PAGER is optimal for 1-PŁ.
- 1-PŁ functions appear in applications such as reinforcement learning.

References

[1] J. Bolte, A. Daniilidis, A. Lewis. The Łojasiewicz inequality for non-smooth subanalytic functions with applications to subgradient dynamical systems. SIAM Journal on Optimization, 2007.

[2] B. Polyak. Gradient methods for the minimization of functionals. USSR Computational Maththematics and Mathematical Physics, 1963.

[3] A. Khaled, P. Richtárik. Better theory for SGD in the non-convex world. arXiv:2002.03329, 2020.

[4] X. Fontaine, V. Bortoli, A. Durmus. Convergence rates and approximation results for SGD and its continuous-time counterpart. COLT 2021.

[5] Z. Li, H. Bao, X. Zhang, P. Richtárik. PAGE: A simple and optimal probabilistic gradient estimator for nonconvex optimization. ICML 2021.