Convergence of Random Reshuffling Under the Kurdyka-Łojasiewicz Inequality

Xiao Li

with Andre Milzarek and Junwen Qiu

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

https://arxiv.org/pdf/2110.04926

Outline

Overview

Random Reshuffling (RR)

Standard Kurdyka-Łojasiewicz (KL) Analysis Framework

A New Analysis Framework and Main Results

Conclusion

Problem Statement

Optimization Problem

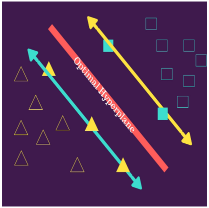
Consider the finite-sum optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{N} \sum_{i=1}^{N} f(x, i),$$

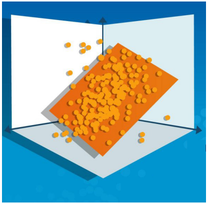
$f(\cdot, i)$ is smooth and possibly nonconvex $\forall i \in [N] := \{1, \cdots, N\}$.

Applications:

Neural Network



Supervised Learning



Matrix Optimization

Random Reshuffling (RR)

RR: Iteration t

1. Set $x_0^t = x^t$ and generate a random permutation σ^t of $[N]$;
2. **For** $i = 1, \cdots, N$ (loop of one epoch)

   $$x_i^t = x_{i-1}^t - \alpha_t \nabla f(x_{i-1}^t, \sigma_i^t);$$

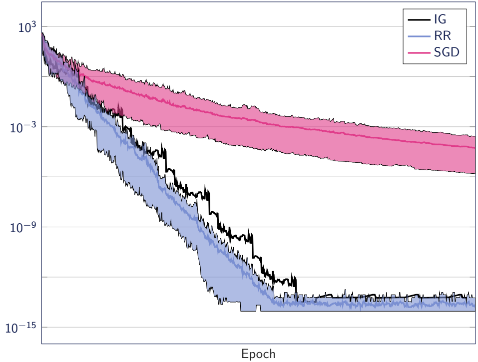
   End
3. Set $x^{t+1} = x_N^t$.

Remark:

- The step size sequence $\{\alpha_t\}$ is diminishing.
- **Incremental gradient (IG) method:** $\sigma^t = [N]$.
- **Stochastic gradient (SGD) method**: sampling with-replacement.

IG	f_1	f_2	f_3	f_4	f_5
RR	f_2	f_5	f_4	f_1	f_3
SGD	f_3	f_2	f_2	f_1	f_3

Why Random Reshuffling (RR)?



- Easy to implement due to simple implementation.
- Better performance than SGD.
- Widely used in large-scale learning software packages.

Different Types of Convergence:

Two main types [Absil et al. (2005)]:

- **Weak convergence:** $\lim_{t\to\infty} \|\nabla f(x^t)\| = 0$;
- **Strong (limit-point) convergence:** $\lim_{t\to\infty} x^t = x^* \in \mathrm{crit}(f)$;

Another usually used notion of "convergence":

- **Iteration complexity:** No convergence guarantees for $\|\nabla f(x^t)\|$,

$$\min_{0\le t\le T} \|\nabla f(x^t)\| \le \frac{1}{\sqrt{T}}, \quad \text{or} \quad \frac{1}{T}\sum_{t=1}^{T} \|\nabla f(x^t)\| \le \frac{1}{\sqrt{T}}.$$

Existing Theoretical Results

IG

- Weak convergence has been shown in [Luo and Tseng, 1994; Tseng, 1998; Solodov and Zavriev, 1998].
- Strong convergence at rate $\mathcal{O}(1/t)$ under strong convexity has been shown in [Gürbüzbalaban et al., 2019].

Existing Theoretical Results

IG

- Weak convergence has been shown in [Luo and Tseng, 1994; Tseng, 1998; Solodov and Zavriev, 1998].
- Strong convergence at rate $\mathcal{O}(1/t)$ under strong convexity has been shown in [Gürbüzbalaban et al., 2019].

RR

- Strong convergence at rate $\mathcal{O}(1/t)$ under strong convexity with Lipschitz continuous $f, \nabla f, \nabla^2 f$ [Gürbüzbalaban et al., 2021].
  - SGD has min-max rate $\mathcal{O}(1/\sqrt{t})$ [Nemirovskij and Yudin, 1983].
- More results about iteration complexity are shown in [HaoChen and Sra, 2019; Nagaraj et al., 2019; Mishchenko et al., 2020, etc.]
- In nonconvex setting, $\liminf_{t\to\infty} \|\nabla f(x^t)\| = 0$ and iteration complexity results [Nguyen et al., 2020].

Overview

KL Inequality

KL Inequality

$f : \mathbb{R}^n \to \mathbb{R}$ is said to satisfy the KL property at $\bar{x} \in \mathbb{R}^n$ if

- $\exists$ a desingularization function ϱ for all $x \in U$:=neighborhood of $\bar{x}$

$$\varrho'(|f(x) - f(\bar{x})|) \cdot \|\nabla f(x)\| \geq 1. \quad \text{(KL inequality)}$$

Remark:

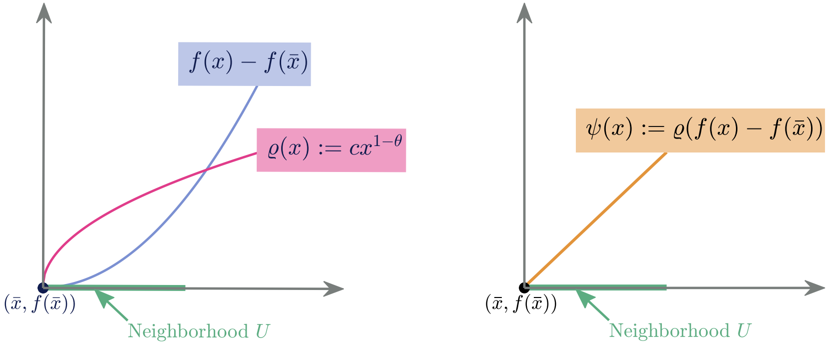
- Popular choice: $\varrho(x) = cx^{1-\theta}$, i.e., the Łojasiewicz inequality

$$|f(x) - f(\bar{x})|^{\theta} \leq c\|\nabla f(x)\|, \quad c > 0.$$

  Here, $\theta \in [0, 1)$ is called the KL exponent.

- Very mild and general [Attouch et al. (2013)].

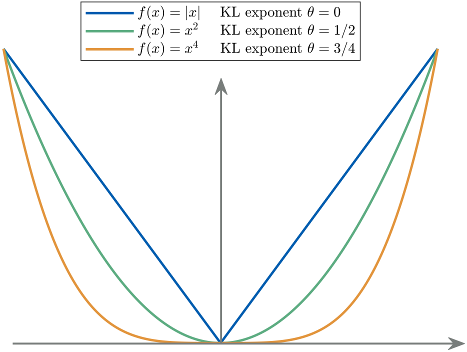
Geometric Interpretation



$$\boxed{\varrho'(f(x) - f(\bar{x})) \cdot \|\nabla f(x)\| \geq 1 \iff \psi'(x) \geq 1.}$$

- ψ is sharp (absolute value like).
- KL inequality characterizes 'curvature' of f around $\bar{x}$.

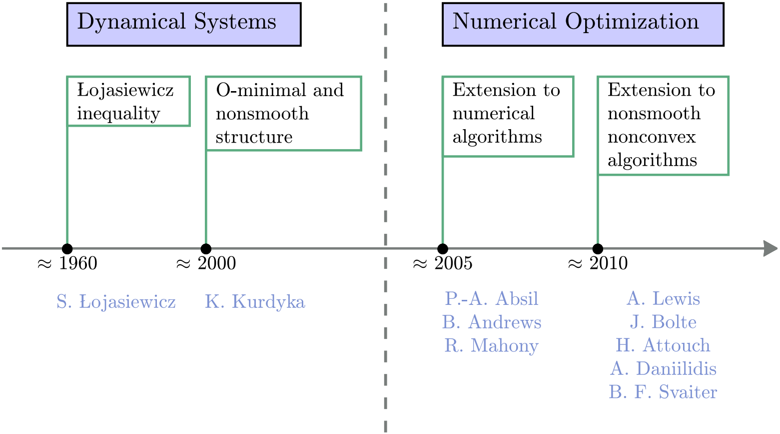
Geometric Interpretation: KL exponent θ



Functions with different KL exponent θ.

- Larger θ implies further from 'sharp' curvature.
- $\theta \in [0, \frac{1}{2}]$: good curvature.

KL History



Dynamical Systems
Numerical Optimization
Łojasiewicz inequality
O-minimal and nonsmooth structure
Extension to numerical algorithms
Extension to nonsmooth nonconvex algorithms
≈ 1960
≈ 2000
≈ 2005
≈ 2010
S. Łojasiewicz
K. Kurdyka
P.-A. Absil
B. Andrews
R. Mahony
A. Lewis
J. Bolte
H. Attouch
A. Daniilidis
B. F. Svaiter

Standard KL Framework [Attouch et al. (2013); Absil et al. (2005)]

Conditions

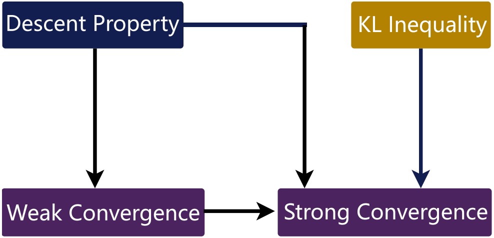
- **Sufficient decrease property:** (algorithmic)

$$f(x^{t+1}) \leq f(x^t) - c\|\nabla f(x^t)\|^2.$$

  - This leads to weak convergence.

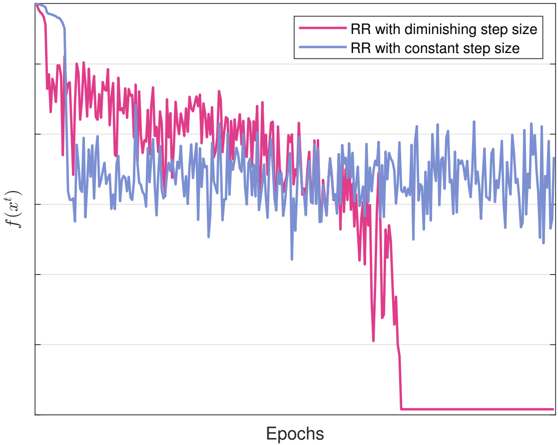
- f satisfies the **KL inequality** (problem intrinsic).

Proof Flow



The Fundamental Difference

- **Non-descent nature** and needs **diminishing step size**.



Overview

An Analysis Framework I

New Conditions (Algorithmic)

- **Approximate descent property:** $\exists\ b_1 > 0, b_2 > 0, \mu \geq 2$ s.t.

$$f(x^{t+1}) \leq f(x^t) - b_1\alpha_t\|\nabla f(x^t)\|^2 + b_2\alpha_t^{\mu}.$$

- **Relative error:** $\exists\ c_1 > 0, c_2 > 0, \nu \geq 2$ s.t.

$$\|x^{t+1} - x^t\| \leq c_1\alpha_t\|\nabla f(x^t)\| + c_2\alpha_t^{\nu}.$$

Observations:

- '$b_2\alpha_t^{\mu}$' is the non-descent term.
- RR only converges to a neighborhood if α_t is constant step size.

An Analysis Framework II

Diminishing step sizes (Algorithmic)

$$\alpha_t > 0, \quad \sum \alpha_t = \infty, \quad \text{and} \quad \sum \alpha_t^{\min\{\mu,\nu\}} < \infty.$$

- $\alpha_t = \mathcal{O}(\frac{1}{t^\gamma})$ with $\gamma = 1$ and $\frac{1}{2}$ are popular choices.

Consequence:

$$\boxed{\text{Weak Convergence: } \lim_{t\to\infty} \|\nabla f(x^t)\| = 0.}$$

Highlights:

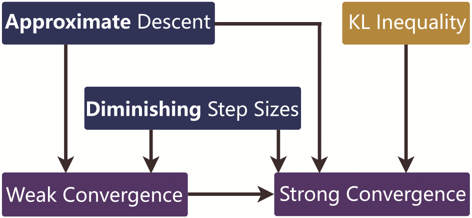
- Approximate descent property + Diminishing step sizes $\Longrightarrow$ special 'descent property'.
- Needs an elementary analysis (ε-δ arguments).

New Analysis Framework III

Additional Condition (Problem Intrinsic)

- **KL inequality** of f.

New Proof Flow



Highlights:

- A novel descent-type condition for the iterations.
- A variant of KL inequality (taking absolute value of "$f(x) - f(\bar{x})$").
- Combining KL framework with the dynamics of diminishing step sizes, etc.

Main Results

Assumptions

A.1 Each component function $f(\cdot, i)$ has Lipschitz continuous gradient.

A.2 Objective function f satisfies the KL inequality.

Thm: Convergence Results

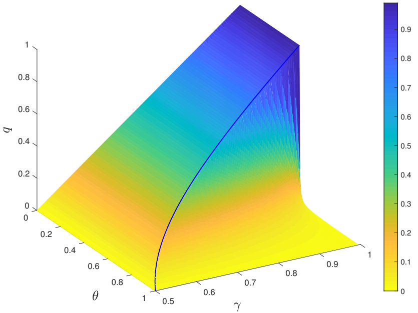
Assume **A.1**-**A.2**. Using diminishing step sizes $\alpha_t = \mathcal{O}(1/t^\gamma)$, we obtain

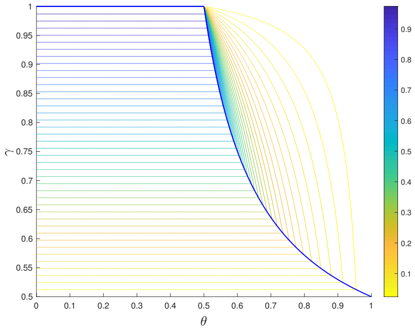
- strong (limit-point) convergence $\lim_{t\to\infty} x^t = x^* \in \mathrm{crit}(f)$.
- We have

$$\|x^t - x^*\| = \begin{cases} \mathcal{O}\left(1/t\right), & \text{if } \theta \in [0, 1/2], \\ \mathcal{O}\left(1/t^p\right), p \in (0,1), & \text{if } \theta \in (1/2, 1). \end{cases}$$

  Here, $\theta \in [0, 1)$ is the KL exponent.

Convergence Rate





Surface and Contour plot of the rates $\mathcal{O}(t^{-q})$ as a multifunction of the step size parameter $\alpha_t = \mathcal{O}(1/t^{\gamma})$ with $\gamma \in (\frac{1}{2}, 1]$ and the KL exponent $\theta \in [0, 1)$.

- $\mathcal{O}(t^{-1})$ rate matches that of the the strongly convex setting.

Standing Out

	Nonconvex	Convex	Strongly Convex
Strong Convergence	This paper		IG: Gürbüzbalaban et al. 19' RR: Gürbüzbalaban et al. 21'
Weak Convergence	IG:[bounded gradient] Luo & Tseng 94' Solodov & Zavriev 98' Tseng 98' RR: This paper		
Iteration Complexity	Nagaraj et al. 19' Mishchenko et al. 20' Nguyen et al. 21'	Shamir 16' Mishchenko et al. 20'	Nagaraj et al. 19' HaoChen & Sra 19' Mishchenko et al. 20' etc..

Overview

Conclusion

- Random Reshuffling has been shown to has strong limit-point convergence in nonconvex setting under KL inequality.

- When KL exponent $\theta = 1/2$, we obtain rate $\|x^t - x^*\| = \mathcal{O}(1/t)$, which coincides with the rate under strong convexity.

- Significantly, our techniques/framework can potentially be utilized for a large class of non-descent algorithms with diminishing step sizes.

Reference I

Absil, P.-A., R. Mahony, and B. Andrews
2005. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547.

Attouch, H., J. Bolte, and B. F. Svaiter
2013. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized Gauss-Seidel methods. *Math. Program.*, 137(1-2, Ser. A):91–129.

Gürbüzbalaban, M., A. Ozdaglar, and P. Parrilo
2021. Why random reshuffling beats stochastic gradient descent. *Mathematical Programming*, 186(1-2):49–84.

Gürbüzbalaban, M., A. Ozdaglar, and P. A. Parrilo
2019. Convergence rate of incremental gradient and incremental Newton methods. *SIAM Journal on Optimization*, 29(4):2542–2565.

HaoChen, J. Z. and S. Sra
2019. Random shuffling beats SGD after finite epochs. In *International Conference on Machine Learning*, Pp. 2624–2633.

Reference II

Luo, Z.-Q. and P. Tseng
1994. Analysis of an approximate gradient projection method with applications to the backpropagation algorithm. *Optimization Methods and Software*, 4(2):85–101.

Mishchenko, K., A. Khaled Ragab Bayoumi, and P. Richtárik
2020. Random reshuffling: Simple analysis with vast improvements. *Advances in Neural Information Processing Systems*, 33.

Nagaraj, D., P. Jain, and P. Netrapalli
2019. Sgd without replacement: Sharper rates for general smooth convex functions. In *International Conference on Machine Learning*, Pp. 4703–4711. PMLR.

Nemirovskij, A. S. and D. B. Yudin
1983. *Problem complexity and method efficiency in optimization*, Wiley-Interscience Series in Discrete Mathematics. Wiley, Chichester.

Nguyen, L. M., Q. Tran-Dinh, D. T. Phan, P. H. Nguyen, and M. van Dijk
2020. A unified convergence analysis for shuffling-type gradient methods. *arXiv preprint arXiv:2002.08246*.

Reference III

Solodov, M. V. and S. Zavriev
1998. Error stability properties of generalized gradient-type algorithms. *Journal of Optimization Theory and Applications*, 98(3):663–680.

Tseng, P.
1998. An incremental gradient(-projection) method with momentum term and adaptive stepsize rule. *SIAM Journal on Optimization*, 8(2):506–531.