Convergence Analysis of a Momentum Algorithm with Adaptive Step Size for Nonconvex Optimization

Anas Barakat and Pascal Bianchi

LTCI, Télécom Paris, Institut Polytechnique de Paris, France
anas.barakat@telecom-paristech.fr

Problem

$$\min_{x\in\mathbb{R}^d} f(x)$$

- f **non-convex** differentiable.
- ∇f is $L-$Lipschitz continuous.
- $\inf_{x\in\mathbb{R}^d} f(x) > -\infty$.

Summary

Main Idea :
clipping the adaptive step size using a bound depending on L(∇F).

Contributions :

- sublinear rates in deterministic and stochastic contexts (no bounded gradients compared to [1], dimension free).
- convergence rates on the function value sequence under Kurdyka-Łojasiewicz (KL) property.

A momentum algorithm

$$\begin{cases} x_{n+1} = x_n - a_{n+1} p_{n+1} \\ p_{n+1} = p_n + b\,(\nabla f(x_n) - p_n) \end{cases}$$

- coordinatewise product.
- $a_n \in \mathbb{R}^d$ may depend on the past gradients $g_k := \nabla f(x_k)$ and the iterates x_k for $k \leq n$.
- includes SGD, Heavy Ball, ADAM and other adaptive algorithms [2].

Mild Assumption (verified for ADAM)

There exists $\alpha > 0$ s.t. $a_{n+1} \leq \frac{a_n}{\alpha}$.

A descent lemma

$$\forall n \in \mathbb{N}, \quad H_n := f(x_n) + \frac{1}{2b}\langle a_n, p_n^2\rangle\,.$$

Lemma. Under previous assumptions, $\forall n \in \mathbb{N}, \forall u \in \mathbb{R}_+$,

$$H_{n+1} \leq H_n - \langle a_{n+1} p_{n+1}^2, A_{n+1}\rangle\,,$$

$$\text{where} \quad A_{n+1} := 1 - \frac{a_{n+1}L}{2} - \frac{|b-(1-\alpha)|}{2u} - \frac{1-\alpha}{2b}$$

Deterministic setting

Theorem. Let previous assumptions hold. Assume $1-\alpha < b \leq 1$ and :

- Let $\varepsilon > 0$ s.t. $a_{\sup} := \frac{2}{L}\left(1 - \frac{(b-(1-\alpha))^2}{2b\alpha} - \frac{1-\alpha}{2b} - \varepsilon\right) \geq 0$.
- Let $\delta > 0$ s.t. $\forall n \in \mathbb{N}$, $\quad \boldsymbol{\delta \leq a_{n+1} \leq \min\left(a_{\sup}, \frac{a_n}{\alpha}\right)}$. (1)

Then (H_n) is nonincreasing, $\lim \nabla f(x_n) \to 0$ as $n \to +\infty$ and

$$\boxed{\forall n \geq 1, \quad \min_{0\leq k\leq n-1} \|\nabla f(x_k)\|^2 \leq \frac{4}{nb^2}\left(\frac{H_0 - \inf f}{\delta\varepsilon} + \|p_0\|^2\right).}$$

Stochastic setting

Theorem. Let previous assumptions hold. Assume $1-\alpha < b \leq 1$ and :

- $\forall x \in \mathbb{R}^d, \quad \mathbb{E}\|\nabla f(x,\xi) - \nabla F(x)\|^2 \leq \sigma^2$.
- Let $\varepsilon > 0$ s.t. $\bar{a}_{\sup} := \frac{1}{L}\left(1 - \frac{(b-(1-\alpha))^2}{b\alpha} - \frac{1-\alpha}{b} - \varepsilon\right) \geq 0$.
- Let $\delta > 0$ s.t. $\forall n \geq 1$, **almost surely,** $\boldsymbol{\delta \leq a_{n+1} \leq \min\left(\bar{a}_{\sup}, \frac{a_n}{\alpha}\right)}$.

$$\boxed{\mathbb{E}[\|\nabla F(x_\tau)\|^2] \leq \frac{4}{n\delta b^2\alpha}\left(\frac{H_0 - \inf f}{\varepsilon} + \|\sqrt{a_0}p_0\|^2 + \frac{n\bar{a}_{\sup}\sigma^2}{2\varepsilon}\right)}$$

where x_τ is an iterate uniformly randomly chosen from $\{x_0, \cdots, x_{n-1}\}$.

KL inequality

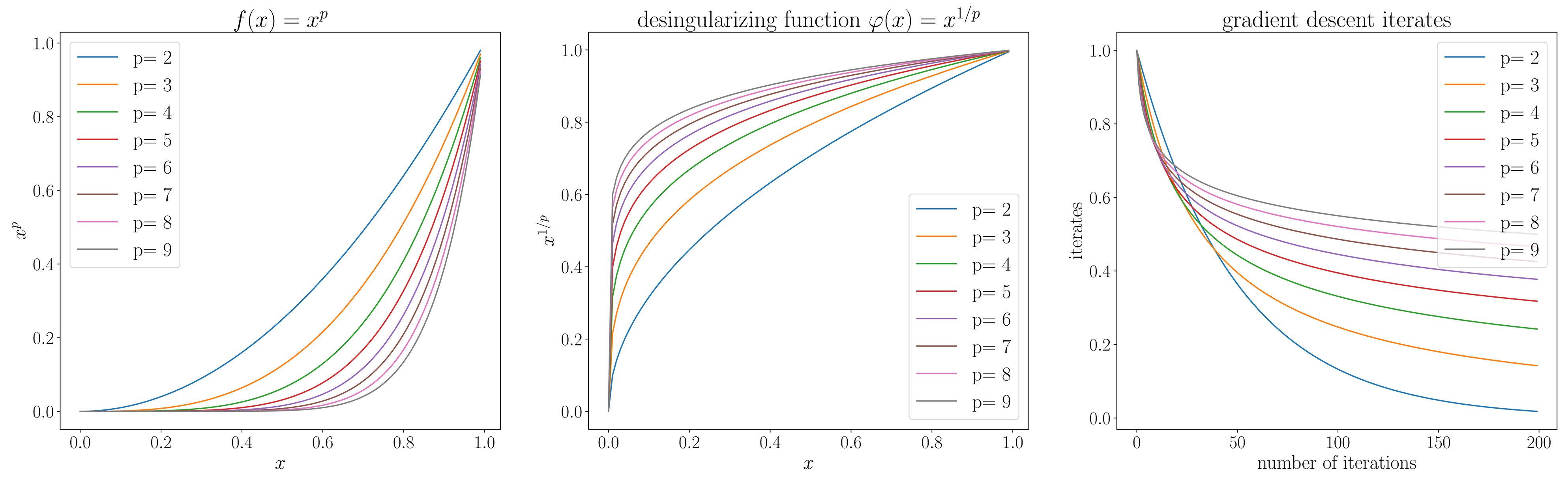
- satisfied by nonsmooth deep neural networks built from activations ReLU ($\max(0,t)$), $t \mapsto t^2$ and log-exp ($\log(1+e^t)$).

$$\Phi_\eta := \{\varphi \in C^0[0,\eta) \cap C^1(0,\eta) : \varphi(0) = 0\,,\, \varphi \,\text{concave and}\, \varphi' > 0\}\,.$$

Definition. (KŁ property, [3, Appendix]) A proper l.s.c function $H : \mathbb{R}^{2d} \to (-\infty, +\infty]$ has the KŁ property locally at $\bar{z} \in \operatorname{dom} H$ if there exist $\eta > 0$, $\varphi \in \Phi_\eta$ and a neighborhood $U(\bar{z})$ s.t. for all $z \in U(\bar{z}) \cap [H(\bar{z}) < H < H(\bar{z}) + \eta]$:

$$\|\nabla(\varphi \circ (H(\cdot) - H(\bar{z}))\,(z)\| \geq 1\,.$$

- H becomes sharp under a reparameterization of its values through the so-called desingularizing function φ.



KL rates (similar techniques to [3, 4])

$$\forall z \in \mathbb{R}^d \times \mathbb{R}^d\,,\; H(z) = H(x,y) = f(x) + \frac{1}{2b}\|y\|^2\,.$$

Theorem. Let $z_k = (x_k, y_k)$ where $y_k = \sqrt{a_k}p_k$, $f(x_*) = \lim H(z_k)$ where $\nabla f(x_*) = 0$. Suppose that f is coercive, condition (1) holds and

- H **is a KŁ function with KŁ exponent** θ i.e. $\varphi(s) = \frac{\bar{c}}{\theta}s^\theta$, $\theta \in (0,1]$.

(i) If $\theta = 1$, then $\boldsymbol{f(x_k)}$ **converges in a finite number of iterations.**

(ii) If $1/2 \leq \theta < 1$, then $\boldsymbol{\exists q \in (0,1), C > 0}$ s.t. $\boldsymbol{f(x_k) - f(x_*) \leq C\,q^k}$.

(iii) If $0 < \theta < 1/2$, then $\boldsymbol{f(x_k) - f(x_*) = O(k^{\frac{1}{2\theta-1}})}$.

References

[1] M. Zaheer, S. Reddi, D. Sachan, S. Kale, and S. Kumar. Adaptive methods for nonconvex optimization. In *Advances in Neural Information Processing Systems*, pages 9793–9803, 2018.

[2] P. Ochs, Y. Chen, T. Brox, and T. Pock. ipiano: Inertial proximal algorithm for nonconvex optimization. *SIAM Journal on Imaging Sciences*, 7(2):1388–1419, 2014.

[3] J. Bolte, S. Sabach, M. Teboulle, and Y. Vaisbourd. First order methods beyond convexity and lipschitz gradient continuity with applications to quadratic inverse problems. *SIAM Journal on Optimization*, 28(3):2131–2151, 2018.

[4] P. R. Johnstone and P. Moulin. Convergence rates of inertial splitting schemes for nonconvex composite optimization. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4716–4720.