

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

Anas Barakat

ANAS.BARAKAT@TELECOM-PARISTECH.FR

Pascal Bianchi

PASCAL.BIANCHI@TELECOM-PARISTECH.FR

*LTCI, Telecom Paris, Institut Polytechnique de Paris
19 Place Marguerite Perey, 91120 Palaiseau, France*

Abstract

Although ADAM is a very popular algorithm for optimizing the weights of neural networks, it has been recently shown that it can diverge even in simple convex optimization examples. Several variants of ADAM have been proposed to circumvent this convergence issue. In this work, we study the ADAM algorithm for smooth nonconvex optimization under a boundedness assumption on the adaptive learning rate. The bound on the adaptive step size depends on the Lipschitz constant of the gradient of the objective function and provides safe theoretical adaptive step sizes. Under this boundedness assumption, we show a novel first order convergence rate result in both deterministic and stochastic contexts. We further establish convergence rates of the function value sequence using the Kurdyka-Łojasiewicz property.

1. Introduction

Consider the unconstrained optimization problem $\min_{x \in \mathbb{R}^d} f(x)$, where $f : \mathbb{R}^d \rightarrow \mathbb{R}$ is a differentiable map and d is an integer. Gradient descent is one of the most classical algorithms to solve this problem. Since the seminal work Robbins and Monro [33], its stochastic counterpart became one of the most popular algorithms to solve machine learning problems (see Bottou et al. [9] for a recent survey). Recently, a class of algorithms called adaptive algorithms which are variants of stochastic gradient descent became very popular in machine learning applications. Using a coordinate-wise step size computed using past gradient information, the step size is adapted to the function to optimize and does not follow a predetermined step size schedule. Among these adaptive algorithms, ADAM [18] is very popular for optimizing the weights of neural networks. However, recently, Reddi et al. [32] exhibited a simple convex stochastic optimization problem over a compact set where ADAM fails to converge because of its short-term gradient memory. Moreover, they proposed an algorithm called AMSGRAD to fix the convergence issue of ADAM. This work opened the way to the emergence of other variants of ADAM to overcome its convergence issues (PADAM[39], YOGI[38], ADABOUND[25], ADASHIFT [40], NADAM [13], QHADAM [26], RADAM [23] etc., see Appendix A for a detailed review). In this work, under a bounded step size assumption, we propose a theoretical analysis of ADAM for nonconvex optimization.

Contributions.

- We establish a convergence rate for ADAM in the deterministic case for nonconvex optimization under a bounded step size. This algorithm can be seen as a deterministic clipped version of ADAM

which guarantees safe theoretical step sizes. More precisely, we show a $O(1/n)$ convergence rate by introducing a suitable Lyapunov function.

- We show a similar convergence result to Zaheer et al. [38, Thm. 1] for nonconvex stochastic optimization up to the limit of the variance of stochastic gradients under an almost surely bounded step size. In comparison to the literature, we relax the hypothesis of the boundedness of the gradients and improve the dependency of the convergence result on the dimension d of the parameters.
- We propose a convergence rate analysis of the objective function of the algorithm using the Kurdyka-Łojasiewicz (KL) property. To the best of our knowledge, this is the first time such a result is established for an adaptive optimization algorithm.

2. A Momentum Algorithm with Adaptive Step Size

Notations. All operations between vectors of \mathbb{R}^d are to read coordinatewise. The vector of ones of \mathbb{R}^d is denoted by $\mathbf{1}$. When a scalar is added to a vector, it is added to each one of its coordinates. Inequalities are also to be read coordinatewise. If $x \in \mathbb{R}^d, x \leq \lambda \in \mathbb{R}$ means that each coordinate of x is smaller than λ .

We investigate the following algorithm defined by two sequences (x_n) and (p_n) in \mathbb{R}^d :

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

where $\nabla f(x)$ is the gradient of f at point x , (a_n) is a sequence of vectors in \mathbb{R}^d with positive coordinates, b is a positive real constant and $x_0, p_0 \in \mathbb{R}^d$.

Algorithm (1) includes the classical Heavy-ball method as a special case, but is much more general. Indeed, we allow the sequence of step sizes (a_n) to be adaptive : $a_n \in \mathbb{R}^d$ may depend on the past gradients $g_k := \nabla f(x_k)$ and the iterates x_k for $k \leq n$. We stress that the step size a_n is a vector of \mathbb{R}^d and that the product $a_{n+1}p_{n+1}$ in (1) is read componentwise (this is equivalent to the formulation with a diagonal matrix preconditioner applied to the gradient [2, 16, 27]). This vector step size formulation includes all adaptive step size algorithms as special cases. In particular, ADAM [18] defined by the iterates :

$$\begin{cases} x_{n+1} = x_n - \frac{a}{\epsilon + \sqrt{v_{n+1}}} p_{n+1} \\ p_{n+1} = p_n + b(\nabla f(x_n) - p_n) \\ v_{n+1} = v_n + c(\nabla f(x_n)^2 - v_n) \end{cases} \quad (2)$$

for constants $a \in \mathbb{R}_+, b, c \in [0, 1]$, can be seen as an instance of this algorithm by setting $a_n = \frac{a}{\epsilon + \sqrt{v_n}}$ where the vector v_n , as defined above, is an exponential moving average of the gradient squared. For simplification, we omit bias correction steps for p_{n+1} and v_{n+1} .

We introduce the main assumption on the objective function which is standard in gradient-based algorithms analysis.

Assumption 2.1 *The mapping $f : \mathbb{R}^d \rightarrow \mathbb{R}$ satisfies the following*

- (i) *f is continuously differentiable and its gradient ∇f is L -Lipschitz continuous,*
- (ii) *f is bounded from below, i.e., $\inf_{x \in \mathbb{R}^d} f(x) > -\infty$.*

3. First Order Convergence Rate

3.1. Deterministic setting

Let $(H_n)_{n \geq 0}$ be a sequence defined for all $n \in \mathbb{N}$ by $H_n := f(x_n) + \frac{1}{2b} \langle a_n, p_n^2 \rangle$. We further assume the following step size growth condition.

Assumption 3.1 *There exists $\alpha > 0$ s.t. $a_{n+1} \leq \frac{a_n}{\alpha}$.*

Note that this assumption is satisfied for ADAM with $\alpha = \sqrt{1-c}$ where c is the parameter in (2). Unlike in AMSGRAD [32], the step size is not necessarily nonincreasing.

We now establish a key lemma which will give us a descent property under mild assumptions. We provide a proof of this result in Appendix B.1.

Lemma 1 *Let Assumptions 2.1 and 3.1 hold true. Then, for all $n \in \mathbb{N}$, for all $u \in \mathbb{R}_+$,*

$$H_{n+1} \leq H_n - \langle a_{n+1} p_{n+1}^2, A_{n+1} \rangle - \frac{b}{2} \langle a_{n+1} (\nabla f(x_n) - p_n)^2, B \mathbf{1} \rangle, \quad (3)$$

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

We can now state a sublinear convergence rate for the minimum of the gradients norms until time n .

Theorem 2 *Let Assumptions 2.1 and 3.1 hold true. Suppose that $1 - \alpha < b \leq 1$. Let $\varepsilon > 0$ s.t. $a_{\text{sup}} := \frac{2}{L} \left(1 - \frac{(b - (1 - \alpha))^2}{2b\alpha} - \frac{1 - \alpha}{2b} - \varepsilon \right)$ is nonnegative. Let $\delta > 0$ s.t. for all $n \in \mathbb{N}$,*

$$\delta \leq a_{n+1} \leq \min \left(a_{\text{sup}}, \frac{a_n}{\alpha} \right). \quad (4)$$

(i) *The sequence (H_n) is nonincreasing and $\sum_n \|p_n\|^2 < \infty$.*

In particular, $\lim x_{n+1} - x_n \rightarrow 0$ and $\lim \nabla f(x_n) \rightarrow 0$ as $n \rightarrow +\infty$.

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

Unlike most of the theoretical results for variants of ADAM [10, 25, 32, 39, 41], we remark that the bound (ii) does not depend on the dimension d of the parameter x_k . Moreover, remark that the step size bound almost matches the classical $2/L$ upperbound on the step size of gradient descent. Perhaps, the closest idea to our algorithm is the recent ADABOUND [25] which considers a dynamic learning rate bound. Our approach is different : we propose a static bound on the adaptive learning rate which depends on the Lipschitz constant of the objective function. In particular, this bound stems naturally from our theoretical derivations and is not to be chosen by the user (see Appendix A.4 for a discussion).

3.2. Stochastic setting

We establish a similar bound in the stochastic setting. Consider the problem of finding a local minimizer of the expectation $F(x) := \mathbb{E}(f(x, \xi))$ w.r.t. $x \in \mathbb{R}^d$, where $f : \mathbb{R}^d \times \Xi \rightarrow \mathbb{R}$ is a measurable map and $f(\cdot, \xi)$ is a possibly nonconvex function depending on some random variable ξ . The distribution of ξ is assumed to be unknown, but revealed online by the observation of iid copies $(\xi_n : n \geq 1)$ of the r.v. ξ . For a fixed value of ξ , the mapping $x \mapsto f(x, \xi)$ is supposed to be differentiable, and its gradient w.r.t. x is denoted by $\nabla f(x, \xi)$. We study a stochastic version of Algorithm (1) by replacing the deterministic gradient $\nabla f(x_n)$ by $\nabla f(x_n, \xi_{n+1})$.

Theorem 3 *Let Assumption 2.1 (for F) and Assumption 3.1 hold true. Assume the following bound on the variance in stochastic gradients: $\mathbb{E}\|\nabla f(x, \xi) - \nabla F(x)\|^2 \leq \sigma^2$ for all $x \in \mathbb{R}^d$. Suppose moreover that $1 - \alpha < b \leq 1$. Let $\varepsilon > 0$ s.t. $\bar{a}_{\text{sup}} := \frac{1}{L} \left(1 - \frac{(b-(1-\alpha))^2}{b\alpha} - \frac{1-\alpha}{b} - \varepsilon\right)$ is nonnegative. Let $\delta > 0$ s.t. for all $n \geq 1$, almost surely,*

$$\delta \leq a_{n+1} \leq \min\left(\bar{a}_{\text{sup}}, \frac{a_n}{\alpha}\right). \quad (5)$$

Then,

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

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

The proof is deferred to the appendix. In the special case where there is no momentum in the algorithm (i.e. RMSPROP) and assuming that the gradients are bounded, a similar convergence rate is obtained in Zaheer et al. [38, Thm. 1] (see Appendix A.4).

4. Convergence rate analysis under the KŁ property

Historically introduced by the fundamental works of Łojasiewicz [24] and Kurdyka [19], the KŁ inequality is the key tool of our analysis. We refer to Bolte et al. [6] for an in-depth presentation of this property. The KŁ inequality is satisfied by a broad class of functions including nonsmooth deep neural networks built from activation functions such as ReLU ($\max(0, t)$) and pieces such as quadratics t^2 and log-exp ($\log(1 + e^t)$) (see Bolte et al. [7, Appendix] for more examples of functions). Indeed, this inequality is verified for the large class of functions which are definable in an o-minimal structure [19], a class which is stable under all the typical functional operations in optimization (e.g. sums, compositions, inf-projections).

The KŁ inequality has been used to show the convergence of several first-order optimization methods towards critical points [3–5, 7, 15, 21]. In this section, we use a methodology exposed in Bolte et al. [8, Appendix] to show convergence rates based on the KŁ property. We modify it to encompass momentum methods. Note that although this modification was initiated in Ochs [28], Ochs et al. [29], we use a different separable Lyapunov function. To the best of our knowledge, this is the first time such a KŁ-based analysis is proposed for an adaptive algorithm.

Consider the function $H : \mathbb{R}^d \times \mathbb{R}^d \rightarrow \mathbb{R}$ defined for all $z = (x, y) \in \mathbb{R}^d \times \mathbb{R}^d$ by

$$H(z) = H(x, y) = f(x) + \frac{1}{2b}\|y\|^2. \quad (6)$$

Notice that $H_n = f(x_n) + \frac{1}{2b}\langle a_n, p_n^2 \rangle = H(x_n, y_n)$ where $(y_n)_{n \in \mathbb{N}}$ is a sequence defined for all $n \in \mathbb{N}$ by $y_n = \sqrt{a_n} p_n$.

Assumption 4.1 *f is coercive i.e. $\lim_{\|x\| \rightarrow +\infty} f(x) = +\infty$.*

Define $[\alpha < H < \beta] := \{z \in \mathbb{R}^{2d} : \alpha < H(z) < \beta\}$. Let $\eta > 0$ and define

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

where $C^0[0, \eta]$ is the set of continuous functions on $[0, \eta]$ and $C^1(0, \eta)$ is the set of continuously differentiable functions on $(0, \eta)$.

Definition 4 (*KŁ property, Bolte et al. [8, Appendix]*) A proper and lower semicontinuous (l.s.c) function $H : \mathbb{R}^{2d} \rightarrow (-\infty, +\infty]$ has the KŁ property locally at $\bar{z} \in \text{dom } H$ if there exist $\eta > 0$, $\varphi \in \Phi_\eta$ and a neighborhood $U(\bar{z})$ s.t. for all $z \in U(\bar{z}) \cap [H(\bar{z}) < H < H(\bar{z}) + \eta]$:

$$\varphi'(H(z) - H(\bar{z})) \|\nabla H(z)\| \geq 1. \quad (7)$$

When $H(\bar{z}) = 0$, we can rewrite Equation (7) as : $\|\nabla(\varphi \circ H)(z)\| \geq 1$ for suitable z points. This means that H becomes sharp under a reparameterization of its values through the so-called desingularizing function φ . Indeed, it transforms a region where the gradients are arbitrarily small into a region where gradients are lowerbounded, away from zero. The function H is said to be a KŁ function if it has the KŁ property at each point of the domain of its gradient.

Definition 5 (*KŁ exponent*) If φ can be chosen as $\varphi(s) = \frac{\bar{c}}{\theta} s^\theta$ for some $\bar{c} > 0$ and $\theta \in (0, 1]$ in Theorem 4, then we say that H has the KŁ property at \bar{z} with an exponent of θ ¹. We say that H is a KŁ function with an exponent θ if it has the same exponent θ at any \bar{z} .

Furthermore, if H is a proper closed semialgebraic function, then H is a KŁ function with a suitable exponent $\theta \in (0, 1]$. The slope of φ around the origin informs about the "flatness" of a function around a point. Hence, the KŁ exponent allows to obtain convergence rates. In the light of this remark, we state one of the main results of this work.

Theorem 6 (*Convergence rates*) Let $(z_k)_{k \in \mathbb{N}}$ be the sequence defined for all $k \in \mathbb{N}$ by $z_k = (x_k, y_k)$ where $y_k = \sqrt{a_k} p_k$ and (x_k, p_k) is generated by Algorithm (1) from a starting point z_0 . Let Assumptions 2.1, 3.1 and 4.1 hold true. Assume that Condition (4) holds. Suppose moreover that H is a KŁ function with KŁ exponent θ . Denote by $f(x_*)$ the limit of the sequence $(H(z_k))_{k \in \mathbb{N}}$ where x_* is a critical point of f . Then, the following convergence rates hold:

- (i) If $\theta = 1$, then $f(x_k)$ converges in a finite number of iterations.
- (ii) If $1/2 \leq \theta < 1$, then $f(x_k)$ converges to $f(x_*)$ linearly i.e. there exist $q \in (0, 1)$, $C > 0$ s.t. $f(x_k) - f(x_*) \leq C q^k$.
- (iii) If $0 < \theta < 1/2$, then $f(x_k) - f(x_*) = O(k^{\frac{1}{2\theta-1}})$.

Sketch of the proof. The proof consists of two main steps. The first one is to show that the iterates enter and stay in a region where the KŁ inequality holds. More precisely, this region will be the set of limit points of the sequence (z_k) which we show is included in the set of critical points of the function H . This first step of the proof is achieved using the properties of the limit set (Theorem 11) and the uniformized KŁ property (Theorem 12). Then, the second step is to exploit this inequality in order to derive the sought convergence results. We defer the complete proof to Appendix C.3. It is possible to derive a convergence rate on the objective function values under a KŁ assumption on this same function f instead of an assumption on the Lyapunov function H (see Appendix C.4).

1. $\alpha := 1 - \theta$ is also defined as the KŁ exponent in other papers [20].

References

- [1] P-A. Absil, R. Mahony, and B. Andrews. Convergence of the iterates of descent methods for analytic cost functions. *SIAM Journal on Optimization*, 16(2):531–547, 2005.
- [2] N. Agarwal, B. Bullins, X. Chen, E. Hazan, K. Singh, C. Zhang, and Y. Zhang. Efficient full-matrix adaptive regularization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 102–110, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL <http://proceedings.mlr.press/v97/agarwal19b.html>.
- [3] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116(1-2):5–16, 2009.
- [4] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran. Proximal alternating minimization and projection methods for nonconvex problems: An approach based on the kurdyka-łojasiewicz inequality. *Mathematics of Operations Research*, 35(2):438–457, 2010.
- [5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1-2):91–129, 2013.
- [6] J. Bolte, A. Daniilidis, O. Ley, and L. Mazet. Characterizations of łojasiewicz inequalities: subgradient flows, talweg, convexity. *Transactions of the American Mathematical Society*, 362(6):3319–3363, 2010.
- [7] J. Bolte, S. Sabach, and M. Teboulle. Proximal alternating linearized minimization for non-convex and nonsmooth problems. *Mathematical Programming*, 146(1-2):459–494, 2014.
- [8] 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.
- [9] L. Bottou, F. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.
- [10] X. Chen, S. Liu, R. Sun, and M. Hong. On the convergence of a class of adam-type algorithms for non-convex optimization. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=H1x-x309tm>.
- [11] S. De, A. Mukherjee, and E. Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and their comparison to nesterov acceleration on autoencoders. *arXiv preprint arXiv:1807.06766*, 2018.
- [12] J. Diakonikolas and M. I. Jordan. Generalized momentum-based methods: A hamiltonian perspective. *arXiv preprint arXiv:1906.00436*, 2019.
- [13] T. Dozat. Incorporating nesterov momentum into adam. 2016.

- [14] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(Jul):2121–2159, 2011.
- [15] P. Frankel, G. Garrigos, and J. Peypouquet. Splitting methods with variable metric for kurdyka–łojasiewicz functions and general convergence rates. *Journal of Optimization Theory and Applications*, 165(3):874–900, 2015.
- [16] V. Gupta, T. Koren, and Y. Singer. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv preprint arXiv:1706.06569*, 2017.
- [17] 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. IEEE, 2017.
- [18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.
- [19] K. Kurdyka. On gradients of functions definable in o-minimal structures. In *Annales de l’institut Fourier*, volume 48, pages 769–783, 1998.
- [20] G. Li and T. K. Pong. Calculus of the exponent of kurdyka–łojasiewicz inequality and its applications to linear convergence of first-order methods. *Foundations of computational mathematics*, 18(5):1199–1232, 2018.
- [21] Qunwei Li, Yi Zhou, Yingbin Liang, and Pramod K Varshney. Convergence analysis of proximal gradient with momentum for nonconvex optimization. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2111–2119. JMLR. org, 2017.
- [22] J. Liang, J. Fadili, and G. Peyré. A multi-step inertial forward-backward splitting method for non-convex optimization. In *Advances in Neural Information Processing Systems*, pages 4035–4043, 2016.
- [23] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.
- [24] S. Łojasiewicz. Une propriété topologique des sous-ensembles analytiques réels. *Les équations aux dérivées partielles*, 117:87–89, 1963.
- [25] L. Luo, Y. Xiong, and Y. Liu. Adaptive gradient methods with dynamic bound of learning rate. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=Bkg3g2R9FX>.
- [26] J. Ma and D. Yarats. Quasi-hyperbolic momentum and adam for deep learning. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=S1fUpoR5FQ>.
- [27] H. B. McMahan and M. J. Streeter. Adaptive bound optimization for online convex optimization. In *COLT*, pages 244–256, 2010.
- [28] P. Ochs. Local convergence of the heavy-ball method and ipiano for non-convex optimization. *Journal of Optimization Theory and Applications*, 177(1):153–180, 2018.

- [29] 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. doi: 10.1137/130942954. URL <https://doi.org/10.1137/130942954>.
- [30] R. Pascanu, T. Mikolov, and Y. Bengio. On the difficulty of training recurrent neural networks. In *International conference on machine learning*, pages 1310–1318, 2013.
- [31] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.
- [32] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018. URL <https://openreview.net/forum?id=ryQu7f-RZ>.
- [33] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.
- [34] H. Robbins and S. Monro. A stochastic approximation method. In *Herbert Robbins Selected Papers*, pages 102–109. Springer, 1985.
- [35] I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- [36] T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *Coursera: Neural networks for machine learning*, 4(2):26–31, 2012.
- [37] Z. Wu and M. Li. General inertial proximal gradient method for a class of nonconvex nonsmooth optimization problems. *Computational Optimization and Applications*, 73(1):129–158, 2019.
- [38] 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.
- [39] D. Zhou, Y. Tang, Z. Yang, Y. Cao, and Q. Gu. On the convergence of adaptive gradient methods for nonconvex optimization. *arXiv preprint arXiv:1808.05671*, 2018.
- [40] Z. Zhou, Q. Zhang, G. Lu, H. Wang, W. Zhang, and Y. Yu. Adashift: Decorrelation and convergence of adaptive learning rate methods. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=HkgTkhRcKQ>.
- [41] F. Zou, L. Shen, Z. Jie, W. Zhang, and W. Liu. A sufficient condition for convergences of adam and rmsprop. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 11127–11135, 2019.

Appendix A. Appendix A : Related works

A.1. The Heavy-Ball Algorithm.

Adaptive algorithms as Heavy Ball. Thanks to its small per-iteration cost and its acceleration properties (at least in the strongly convex case), the Heavy-ball method, also called gradient descent with momentum, recently regained popularity in large-scale optimization [35]. This speeding up idea dates back to the sixties with the seminal work of Polyak [31]. In order to tackle nonconvex optimization problems, Ochs et al. [29] proposed iPiano, a generalization of the well known heavy-ball in the form of a forward-backward splitting algorithm with an inertial force for the sum of a smooth possibly nonconvex and a convex function. In the particular case of the Heavy-ball method, this algorithm writes for two sequences of reals (α_n) and (β_n) :

$$x_{n+1} = x_n - \alpha_n \nabla f(x_n) + \beta_n (x_n - x_{n-1}). \quad (8)$$

We remark that Algorithm (1) can be written in a similar fashion by choosing step sizes $\alpha_n = ba_{n+1}$ and inertial parameters $\beta_n = (1 - b)a_{n+1}/a_n$. Ochs et al. [29] only consider the case where α_n and β_n are real-valued. Moreover, the latter does not consider adaptive step sizes, i.e step sizes depending on past gradient information. We can show some improvement with respect to Ochs et al. [29] with weaker convergence conditions in terms of the step size of the algorithm (see Appendix A.2) while allowing adaptive vector-valued step sizes a_n (see Proposition 7).

It is shown in Ochs et al. [29] that the sequence of function values converges and that every limit point is a critical point of the objective function. Moreover, supposing that the Lyapunov function has the KŁ property at a cluster point, they show the finite length of the sequence of iterates and its global convergence to a critical point of the objective function. Similar results are shown in Wu and Li [37] for a more general version than iPiano [29] computing gradients at an extrapolated iterate like in Nesterov’s acceleration.

Convergence rate. Ochs et al. [29] determines a $O(1/n)$ convergence rate (where n is the number of iterations of the algorithm) with respect to the proximal residual which boils down to the gradient for noncomposite optimization. Furthermore, a recent work introduces a generalization of the Heavy-ball method (and Nesterov’s acceleration) to constrained convex optimization in Banach spaces and provides a non-asymptotic hamiltonian based analysis with $O(1/n)$ convergence rate [12]. In the same vein, in Section 3, we establish a similar convergence result for an adaptive step size instead of a fixed predetermined step size schedule like in the Heavy-ball algorithm (see Theorem 2).

Convergence rates under the Kurdyka-Łojasiewicz property. The KŁ property is a powerful tool to analyze gradient-like methods. We elaborate on this property in Section 4. It is for example possible to derive convergence rates assuming that the objective function satisfies this geometric property. Indeed, some recent progress has been made to study convergence rates of the Heavy-ball algorithm in the nonconvex setting. Ochs [28] establishes local convergence rates for the iterates and the function values sequences under the KŁ property. The convergence proof follows a general method that is often used in non-convex optimization convergence theory. This framework was used for gradient descent [1], for proximal gradient descent (see Attouch and Bolte [3] for an analysis with the Łojasiewicz inequality) and further generalized to a class of descent methods called *gradient-like descent* algorithms [5](see also for ex. Bolte et al. [8, Appendix]). KŁ-based

asymptotic convergence rates were established for constant Heavy-ball parameters [28]. Asymptotic convergence rates based on the KŁ property were also shown [17] for a general algorithm solving nonconvex nonsmooth optimization problems called Multi-step Inertial Forward-Backward splitting [22] which has iPiano and Heavy-ball methods as special cases. In this work, step sizes and momentum parameter vary along the algorithm run and are not supposed constant. However, specific values are chosen and consequently, their analysis does not encompass adaptive step sizes i.e. stepsizes that can possibly depend on past gradient information. In the present work, we establish similar convergence rates for methods such as ADAM under a bounded step size assumption (see Theorem 6). We also mention [21] which analyzes the accelerated proximal gradient method for nonconvex programming (APGnc) and establishes convergence rates of the function value sequence by exploiting the KŁ property. This algorithm is a descent method i.e. the function value sequence is shown to decrease over time. In the present work, we analyze adaptive algorithms which are not descent methods. Note that even Heavy-ball is not a descent method. Hence, our analysis requires additional treatments to exploit the KŁ property : we introduce a suitable Lyapunov function which is not the objective function.

A.2. Comparison to Ochs et al. [29]

We recall the conditions satisfied by α_n and β_n in Ochs et al. [29] in order to traduce them in terms of the algorithm (1) at stake. Define :

$$\delta_n := \frac{1}{\alpha_n} - \frac{L}{2} - \frac{\beta_n}{2\alpha_n} \quad \gamma_n := \delta_n - \frac{\beta_n}{2\alpha_n}.$$

Conditions of Ochs et al. [29] write: $\alpha_n \geq c_1$ $\beta_n \geq 0$ $\delta_n \geq \gamma_n \geq c_2$ where c_1, c_2 are positive constants and (δ_n) is monotonically decreasing.

One can remark that algorithm (1) can be written as (8) with step sizes $\alpha_n = ba_{n+1}$ and inertial parameters $\beta_n = (1 - b)\frac{a_{n+1}}{a_n}$. Conditions on these parameters can be expressed in terms of a_n . Supposing $c_2 = 0$, the condition $\gamma_n \geq c_2$ is equivalent to

$$\frac{a_{n+1}}{a_n} \leq \frac{2}{2 - b(2 - a_n L)}. \quad (9)$$

Note that the classical condition $a_n \leq 2/L$ shows up consequently. Moreover, the condition on (δ_n) is equivalent to

$$\frac{1}{a_{n+1}} \leq \frac{3 - b}{2} \frac{1}{a_n} - \frac{1 - b}{2a_{n-1}} \quad \text{for } n \geq 1. \quad (10)$$

Note that we get rid of condition (10) while allowing adaptive step sizes a_n (see Proposition 7).

A.3. Vector step size formulation

We present in the following table how to recover some of the famous algorithms with a vector step size formulation:

Algorithm	Effective step size a_{n+1}	Momentum
SGD [34]	$a_{n+1} \equiv a$	$b = 1$ (no momentum)
ADAGRAD [14]	$a_{n+1} = a \left(\sum_{i=0}^n g_i^2 \right)^{-1/2}$	$b = 1$
RMSPROP [36]	$a_{n+1} = a \left[\epsilon + \left(c \sum_{i=0}^n (1-c)^{n-i} g_i^2 \right)^{1/2} \right]^{-1}$	$b = 1$
ADAM [18]	$a_{n+1} = a \left[\epsilon + \left(c \sum_{i=0}^n (1-c)^{n-i} g_i^2 \right)^{1/2} \right]^{-1}$	$0 \leq b \leq 1$ (close to 0)

A.4. Variants of ADAM

Theoretical guarantees of variants of ADAM. We list most of the existing variants of the ADAM algorithms together with their theoretical convergence guarantees in the following table. The gradient is supposed L -lipschitz continuous in all the convergence results. $g_{1:T,i} = [g_{1,i}, g_{2,i}, \dots, g_{T,i}]^T$.

Algorithm	Effective step size a_{n+1}	b_n	c_n	Assumptions	Convergence Result
AMSGRAD ⁽¹⁾ , ADAMNC ⁽²⁾ [32]	$\frac{a_0}{\sqrt{n}} \frac{1}{\sqrt{\hat{v}_n}}$ (1) $\hat{v}_{n+1} = \max(\hat{v}_n, (1-c_n)v_n + c_n g_n^2)$ (2) $\hat{v}_{n+1} = (1-c_n)\hat{v}_n + c_n g_n^2$	$1 - b_1 \lambda^{n-1}$ or $1 - \frac{b_1}{n}$	$c_n \equiv c_1$ $\frac{c_1}{n}$ (for ADAMNC)	<ul style="list-style-type: none"> convex functions bounded gradients bounded feasible set $\sum_{i=1}^d \hat{v}_{T,i}^{1/2} \leq d$ (AMSGRAD) $\sum_{i=1}^d \ g_{1:T,i}\ _2 \leq \sqrt{dT}$ $b_1 < \sqrt{c_1}$ (ADAMNC) 	$R_T/T = O(\sqrt{\log T/T})$ $\frac{R_T}{T} = O(1/\sqrt{T})$ (ADAMNC)
ADAM [11]	$\frac{4\ g_n\ ^2 \eta}{3L(1-(1-b)^n)(\eta+2\sigma)^2} \frac{1}{\epsilon + \sqrt{v_n}}$ $v_{n+1} = (1-c_1)v_n + c_1 g_n^2$	$b_n \equiv b_1$ $= 1 - \frac{\eta}{\eta+2\sigma}$	$c_n \equiv c_1$	<ul style="list-style-type: none"> σ-bounded gradients $\epsilon = 2\sigma$ 	$\forall \eta > 0 \exists n \leq \frac{9L\sigma^2(f(x_2) - f(x_*))}{\eta^6}$ s.t. $\ g_n\ \leq \eta$
PADAM, AMSGRAD [39]	$\frac{1}{\sqrt{N}} \frac{\hat{v}_n^p}{\hat{v}_n^{\frac{p}{2}}}$ (AMSGRAD) $\frac{1}{\sqrt{dN}} \frac{1}{\hat{v}_n^{\frac{p}{2}}}$ $\hat{v}_n = \max(\hat{v}_{n-1}, (1-c)v_{n-1} + cg_n^2)$	$b_n \equiv b$	$c_n \equiv c$	<ul style="list-style-type: none"> bounded gradients For PADAM: $\bullet p \in [0, \frac{1}{4}]$ $1-b < (1-c)^{2p}$ $\sum_{i=1}^d \ g_{1:N,i}\ _2 \leq \sqrt{dN}$ 	$\mathbb{E}[\ g_T\ ^2] = O(\frac{1+\sqrt{d}}{\sqrt{N}} + \frac{d}{N})$ $= O(\sqrt{\frac{d}{N}} + \frac{d}{N})$ (AMSGRAD) τ uniform r.v in $\{1, \dots, N\}$
RMSProp ⁽¹⁾ , Yogi ⁽²⁾ [38]	$\frac{a_1}{\epsilon + \sqrt{v_n}}$ (1) $v_{n+1} = (1-c)v_n + cg_n^2$ (2) $v_n = v_{n-1} - \text{csign}(v_{n-1} - g_n^2)$	$b_n \equiv b$	$c_n \equiv c$	<ul style="list-style-type: none"> G-bounded gradients $a_1 \leq \frac{\epsilon\sqrt{1-c}}{2L}$ (Yogi) $a_1 \leq \frac{\epsilon}{2L}$ $\bullet c \leq \frac{\epsilon^2}{16G^2}$ σ^2-bounded variance 	$\mathbb{E}[\ g_T\ ^2] = O(\frac{1}{N} + \sigma^2)$ τ uniform r.v in $\{1, \dots, N\}$ $O(\frac{1}{N})$ if minibatch $\Theta(N)$
AMSGRAD ⁽¹⁾ , ADAFOU ⁽²⁾ [10]	$\frac{1}{\sqrt{n}} \frac{1}{\sqrt{\hat{v}_n}}$ (1) $\hat{v}_{n+1} = \max(\hat{v}_n, (1-c_n)v_n + c_n g_n^2)$ (2) $\hat{v}_{n+1} = (1 - \frac{1}{n})\hat{v}_n + \frac{1}{n} g_n^2$	non-increasing	$c_n \equiv c_1$	<ul style="list-style-type: none"> bounded gradients $\exists c > 0$ s.t. $g_{1,i} \geq c$ 	$\min_{n \in [0, N]} \mathbb{E}[\ g_n\ ^2] = O(\frac{\log N + d^2}{\sqrt{N}})$
GENERIC ADAM [41]	$\frac{\alpha_n}{\sqrt{v_n}}$ $v_{n+1} = (1-c_n)v_n + c_n g_n^2$ $\alpha_n = \hat{\alpha} \frac{\sqrt{1-(1-c)^n}}{1-(1-b)^n}$	$b_n \geq b > 0$	$0 < c_n < 1$ non-increasing $\lim c_n = c > b^2$	<ul style="list-style-type: none"> bounded gradients in expectation $d_n \leq \frac{c_n}{\sqrt{c_n}} \leq c_0 d_n$ d_n non-increasing 	$\mathbb{E}[\ g_T\ ^{\frac{4}{3}}]^{\frac{3}{2}} \leq \frac{C+C'\sum_{n=1}^N \alpha_n \sqrt{c_n}}{N\alpha_N}$ τ uniform r.v in $\{1, \dots, N\}$
ADABOUND ⁽¹⁾ , AMSBOUND ⁽²⁾ [25]	$\frac{1}{\sqrt{n}} \text{clip}(\frac{\alpha}{\sqrt{v_n}}, \eta_l(n), \eta_u(n))$ $\eta_l(n)$ non-decreasing to α_* $\eta_u(n)$ non-increasing to α_* (1) $v_{n+1} = (1-c)v_n + cg_n^2$ (2) $v_{n+1} = \max(v_n, (1-c)v_n + cg_n^2)$	$1 - (1-b)\lambda^{n-1}$ or $1 - \frac{1-b}{n}$ $b_n \geq b$	$c_n \equiv c$	<ul style="list-style-type: none"> bounded gradients closed convex bounded feasible set $1-b < \sqrt{1-c}$ 	$R_T/T = O(1/\sqrt{T})$

Discussion of theoretical results. The first type of convergence results uses the online optimization framework which controls the convergence rate of the average regret. This framework was adopted for AMSGRAD, ADAMNC [32], ADABOUND and AMSBOUND [25]. In this setting, it is assumed that the feasible set containing the iterates is bounded by adding a projection step to the algorithm if needed. We do not make such an assumption in our analysis. [32] establishes a regret bound in the convex setting. The second type of theoretical results is based on the control of the norm of the (stochastic) gradients. We remark that some of these results depend on the dimension of the parameters. Zhou et al. [39] improves this dependency in comparison to Chen et al. [10]. The convergence result in De et al. [11] is established under quite specific values of a_{n+1} , b_n and ϵ . Zaheer et al. [38] show a $O(1/n)$ convergence rate for an increasing mini-batch size. However, the proof is provided for RMSPROP and seems difficult to adapt to ADAM which involves a momentum term. Indeed, unlike RMSPROP, ADAM does not admit the objective function as a Lyapunov function. Although we assume boundedness of the step size by Condition (5), we do not suppose that $a_1 \leq \frac{\epsilon}{2L}$ (see table in Appendix A.4) which can impose a very small step size and result in a slow convergence. The step size assumption $a_1 \leq \frac{\epsilon}{2L}$ imposes a very small step size which may result in a slow convergence. We also remark that all the available theoretical results assume boundedness of the (stochastic) gradients. We do not make such an assumption. Furthermore, we do not add any decreasing $1/\sqrt{n}$ factor in front of the adaptive step size as it is considered in Luo et al. [25], Reddi et al. [32] and Chen et al. [10]. Although constant hyperparameters b and c are used in practice, theoretical results are often established for non constant b_n and c_n [25, 32]. We also mention that most of the theoretical bounds depend on the dimension of the parameter [10, 25, 32, 39, 41].

Other variants of ADAM. Recently, several other algorithms were proposed in the literature to enhance ADAM. Although these algorithms lack theoretical guarantees, they present interesting ideas and show good practical performance. For instance, ADASHIFT [40] argues that the convergence issue of ADAM is due to its unbalanced step sizes. To solve this issue, they propose to use temporally shifted gradients to compute the second moment estimate in order to decorrelate it from the first moment estimate. NADAM [13] incorporates Nesterov’s acceleration into ADAM in order to improve its speed of convergence. Moreover, originally motivated by variance reduction, QHADAM [26] replaces both ADAM’s moment estimates by quasi-hyperbolic terms and recovers ADAM, RMSPROP and NADAM as particular cases (modulo the bias correction). Guided by the same variance reduction principle, RADAM [23] estimates the variance of the effective step size of the algorithm and proposes a multiplicative variance correction to the update rule.

Step size bound. Perhaps, the closest idea to our algorithm is the recent ADABOUND [25] which considers a dynamic learning rate bound. Luo et al. [25] show that extremely small and large learning rates can cause convergence issues to ADAM and exhibit empirical situations where such an issue shows up. Inspired by the gradient clipping strategy proposed in Pascanu et al. [30] to tackle the problem of vanishing and exploding gradients in training recurrent neural networks, Luo et al. [25] apply clipping to the effective step size of the algorithm in order to circumvent step size instability. More precisely, authors propose dynamic bounds on the learning rate of adaptive methods such as ADAM or AMSGRAD to solve the problem of extreme learning rates which can lead to poor performance. Initialized respectively at 0 and ∞ , lower and upper bounds both converge smoothly to a constant final step size following a predetermined formula defined by the user. Consequently, the algorithm resembles an adaptive algorithm in the first iterations and becomes progressively similar

to a standard SGD algorithm. Our approach is different : we propose a static bound on the adaptive learning rate which depends on the Lipschitz constant of the objective function. This bound stems naturally from our theoretical derivations.

Appendix B. Appendix B: proofs of Section 3

B.1. Proof of Theorem 1

Supposing that ∇f is L -Lipschitz, using Taylor's expansion and the expression of p_n in the algorithm, we obtain the following inequality:

$$f(x_{n+1}) \leq f(x_n) - \langle \nabla f(x_n), a_{n+1}p_{n+1} \rangle + \frac{L}{2} \|a_{n+1}p_{n+1}\|^2 \quad (11)$$

Moreover,

$$\frac{1}{2b} \langle a_{n+1}, p_{n+1}^2 \rangle - \frac{1}{2b} \langle a_n, p_n^2 \rangle = \frac{1}{2b} \langle a_{n+1}, p_{n+1}^2 - p_n^2 \rangle + \frac{1}{2b} \langle a_{n+1} - a_n, p_n^2 \rangle. \quad (12)$$

Observing that $p_{n+1}^2 - p_n^2 = -b^2(\nabla f(x_n) - p_n)^2 + 2bp_{n+1}(\nabla f(x_n) - p_n)$, we obtain after simplification :

$$H_{n+1} \leq H_n + \frac{L}{2} \|a_{n+1}p_{n+1}\|^2 - \frac{b}{2} \langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle - \langle a_{n+1}p_{n+1}, p_n \rangle + \frac{1}{2b} \langle a_{n+1} - a_n, p_n^2 \rangle. \quad (13)$$

Using again $p_n = p_{n+1} - b(\nabla f(x_n) - p_n)$, we replace p_n :

$$\begin{aligned} H_{n+1} &\leq H_n + \frac{L}{2} \|a_{n+1}p_{n+1}\|^2 - \frac{b}{2} \langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle \\ &\quad - \langle a_{n+1}, p_{n+1}^2 \rangle + b \langle a_{n+1}p_{n+1}, \nabla f(x_n) - p_n \rangle + \frac{1}{2b} \langle a_{n+1} - a_n, p_n^2 \rangle. \end{aligned}$$

Under Assumption 3.1, we write: $\langle a_{n+1} - a_n, p_n^2 \rangle \leq (1 - \alpha) \langle a_{n+1}, p_n^2 \rangle$ and using $p_n^2 = p_{n+1}^2 + b^2(\nabla f(x_n) - p_n)^2 - 2bp_{n+1}(\nabla f(x_n) - p_n)$, it holds that:

$$\begin{aligned} H_{n+1} &\leq H_n - \langle a_{n+1}, p_{n+1}^2 \rangle - \frac{b}{2} \langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle \\ &\quad + \frac{L}{2} \|a_{n+1}p_{n+1}\|^2 + (b - (1 - \alpha)) \langle a_{n+1}p_{n+1}, \nabla f(x_n) - p_n \rangle \\ &\quad + \frac{1 - \alpha}{2b} \langle a_{n+1}, p_{n+1}^2 \rangle + \frac{b(1 - \alpha)}{2} \langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle. \end{aligned}$$

Using the classical inequality $xy \leq \frac{x^2}{2u} + \frac{uy^2}{2}$, we have :

$$(b - (1 - \alpha)) \langle a_{n+1}p_{n+1}, \nabla f(x_n) - p_n \rangle \leq \frac{|b - (1 - \alpha)|}{2u} \langle a_{n+1}, p_{n+1}^2 \rangle + \frac{|b - (1 - \alpha)|u}{2} \langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle. \quad (14)$$

Hence, after using this inequality and rearranging the terms, we derive the following inequality:

$$H_{n+1} \leq H_n - \langle a_{n+1} p_{n+1}^2, 1 - \frac{a_{n+1} L}{2} - \frac{|b - (1 - \alpha)|}{2u} - \frac{1 - \alpha}{2b} \rangle - \frac{b}{2} \langle a_{n+1} (\nabla f(x_n) - p_n)^2, \left(1 - \frac{|b - (1 - \alpha)| u}{b} - (1 - \alpha) \right) \mathbf{1} \rangle.$$

This concludes the proof.

B.2. An additional proposition

Proposition 7 *Let Assumption 2.1 hold true. Suppose moreover that $1 - \alpha < b \leq 1$. Let $\varepsilon > 0$ s.t. $a_{\text{sup}} := \frac{2}{L} \left(1 - \frac{(b - (1 - \alpha))^2}{2b\alpha} - \frac{1 - \alpha}{2b} - \varepsilon \right)$ is nonnegative. Let $\delta > 0$ s.t. for all $n \in \mathbb{N}$,*

$$a_{n+1} \leq \min \left(a_{\text{sup}}, \frac{a_n}{\alpha} \right).$$

Then, for all $n \geq 1$,

$$\sum_{k=0}^{n-1} \langle a_{k+1}, \nabla f(x_k)^2 \rangle \leq \frac{4}{b^2 \alpha} \left(\frac{H_0 - \inf f}{\varepsilon} + \langle a_0, p_0^2 \rangle \right)$$

Proof This is a consequence of Theorem 1. Conditions $A_{n+1} \geq \varepsilon$ and $B \geq 0$ write as follow :

$$a_{n+1} \leq \frac{2}{L} \left(1 - \frac{b - (1 - \alpha)}{2u} - \frac{1 - \alpha}{2b} - \varepsilon \right) \quad \text{and} \quad u \leq \frac{\alpha b}{b - (1 - \alpha)}.$$

We get the assumption made in the proposition by injecting the second condition into the first one and adding the assumption $\frac{a_{n+1}}{a_n} \leq \alpha$ made in the lemma. Under this assumption, we sum over $0 \leq k \leq n - 1$ Equation (3), rearrange it and use $A_{n+1} \geq \varepsilon$, $B \geq 0$ to obtain :

$$\sum_{k=0}^{n-1} \varepsilon \langle a_{k+1}, p_{k+1}^2 \rangle \leq H_0 - H_n,$$

Then, observe that $H_n \geq f(x_n) \geq \inf f$. Therefore, we derive :

$$\sum_{k=0}^{n-1} \langle a_{k+1}, p_{k+1}^2 \rangle \leq \frac{H_0 - \inf f}{\varepsilon}. \quad (15)$$

Moreover, from the Algorithm 1 second update rule, we get $\nabla f(x_k) = \frac{1}{b} p_{k+1} - \frac{1-b}{b} p_k$. Hence, we have for all $k \geq 0$:

$$\nabla f(x_k)^2 \leq 2 \left(\frac{1}{b^2} p_{k+1}^2 + \frac{(1-b)^2}{b^2} p_k^2 \right) \leq \frac{2}{b^2} (p_{k+1}^2 + p_k^2).$$

We deduce that :

$$\begin{aligned}
\sum_{k=0}^{n-1} \langle a_{k+1}, \nabla f(x_k)^2 \rangle &\leq \frac{2}{b^2} \sum_{k=0}^{n-1} \langle a_{k+1}, p_{k+1}^2 + p_k^2 \rangle \\
&= \frac{2}{b^2} \sum_{k=0}^{n-1} \langle a_{k+1}, p_{k+1}^2 \rangle + \frac{2}{b^2} \sum_{k=0}^{n-1} \langle a_{k+1}, p_k^2 \rangle \\
&\leq \frac{2}{b^2} \sum_{k=0}^{n-1} \langle a_{k+1}, p_{k+1}^2 \rangle + \frac{2}{b^2 \alpha} \sum_{k=0}^{n-1} \langle a_k, p_k^2 \rangle \\
&\leq \frac{2}{b^2 \alpha} \left(2 \sum_{k=1}^{n-1} \langle a_k, p_k^2 \rangle + \langle a_n, p_n^2 \rangle + \langle a_0, p_0^2 \rangle \right) \\
&\leq \frac{4}{b^2 \alpha} \sum_{k=0}^n \langle a_k, p_k^2 \rangle \\
&\leq \frac{4}{b^2 \alpha} \left(\frac{H_0 - \inf f}{\varepsilon} + \langle a_0, p_0^2 \rangle \right).
\end{aligned}$$

■

B.3. Proof of Theorem 2

This is a consequence of Theorem 1. Conditions $A_{n+1} \geq \varepsilon$ and $B \geq 0$ write as follow :

$$a_{n+1} \leq \frac{2}{L} \left(1 - \frac{b - (1 - \alpha)}{2u} - \frac{1 - \alpha}{2b} - \varepsilon \right) \quad \text{and} \quad u \leq \frac{\alpha b}{b - (1 - \alpha)}.$$

We get the assumption made in the proposition by injecting the second condition into the first one and adding the assumption $\frac{a_{n+1}}{a_n} \leq \alpha$ made in the lemma. Under this assumption, we sum over $0 \leq k \leq n - 1$ Equation (3), rearrange it and use $A_{n+1} \geq \varepsilon$, $B \geq 0$ and $a_{k+1} \geq \delta$ to obtain :

$$\sum_{k=0}^{n-1} \delta \varepsilon \|p_{k+1}\|^2 \leq H_0 - H_n,$$

Then, observe that $H_n \geq f(x_n) \geq \inf f$. Therefore, we derive :

$$\sum_{k=0}^{n-1} \|p_{k+1}\|^2 \leq \frac{H_0 - \inf f}{\delta \varepsilon}. \quad (16)$$

Moreover, from the algorithm 1 second update rule, we get $\nabla f(x_k) = \frac{1}{b} p_{k+1} - \frac{1-b}{b} p_k$. Hence, we have for all $k \geq 0$:

$$\|\nabla f(x_k)\|^2 \leq 2 \left(\frac{1}{b^2} \|p_{k+1}\|^2 + \frac{(1-b)^2}{b^2} \|p_k\|^2 \right) \leq \frac{2}{b^2} (\|p_{k+1}\|^2 + \|p_k\|^2).$$

We deduce that :

$$\sum_{k=0}^{n-1} \|\nabla f(x_k)\|^2 \leq \frac{2}{b^2} \sum_{k=0}^{n-1} (\|p_{k+1}\|^2 + \|p_k\|^2) = \frac{2}{b^2} \left(2 \sum_{k=1}^{n-1} \|p_k\|^2 + \|p_n\|^2 + \|p_0\|^2 \right) \leq \frac{4}{b^2} \sum_{k=0}^n \|p_k\|^2. \quad (17)$$

Finally, using Equations (16) and (17), we have :

$$\min_{0 \leq k \leq n-1} \|\nabla f(x_k)\|^2 \leq \frac{1}{n} \sum_{k=0}^{n-1} \|\nabla f(x_k)\|^2 \leq \frac{4}{nb^2} \left(\frac{H_0 - \inf f}{\delta \varepsilon} + \|p_0\|^2 \right)$$

B.4. A convergence result for gradient descent in the nonconvex setting.

Consider the gradient descent algorithm defined by : $x_{k+1} = x_k - \gamma \nabla f(x_k)$. Assume that $\gamma > 0$ and $1 - \frac{\gamma L}{2} > 0$.

Supposing that ∇f is L -Lipschitz, using Taylor's expansion and regrouping the terms, we obtain the following inequality:

$$f(x_{k+1}) \leq f(x_k) - \gamma \left(1 - \frac{\gamma L}{2} \right) \|\nabla f(x_k)\|_2^2.$$

Then, we sum the inequalities for $0 \leq k \leq n-1$, lower bound the gradients norms in the sum by their minimum and we obtain :

$$\min_{0 \leq k \leq n-1} \|\nabla f(x_k)\|_2^2 \leq \frac{f(x_0) - \inf f}{n\gamma(1 - \frac{\gamma L}{2})}.$$

Comparison of Theorem 2 to gradient descent. A similar result holds for deterministic gradient descent. If γ is a fix step size for gradient descent and there exist $\delta > 0, \varepsilon > 0$ s.t. $\gamma > \delta$ and $1 - \frac{\gamma L}{2} > \varepsilon$, then (see Appendix B.4) for all $n \geq 1$:

$$\min_{0 \leq k \leq n-1} \|\nabla f(x_k)\|^2 \leq \frac{f(x_0) - \inf f}{n\gamma(1 - \frac{\gamma L}{2})} \leq \frac{f(x_0) - \inf f}{n\delta\varepsilon}.$$

When $p_0 = 0$ (this is the case for ADAM), the bound in Theorem 2 coincides with the gradient descent bound, up to the constant $4/b^2$. We mention however that ε for Algorithm (1) is defined by a slightly more restrictive condition than for gradient descent : when $b = 1$, there is no momentum and $a_{\text{sup}} = \frac{1}{L}(1 - 2\varepsilon) < 2/L$. Hence, under the boundedness of the effective step size, the algorithm has a similar convergence guarantee to gradient descent. Remark that the step size bound almost matches the classical $2/L$ upperbound on the step size of gradient descent. As it is already known for gradient descent, a large step size, even if it is adaptive, can harm the convergence of the algorithm.

B.5. Proof of Theorem 3

The proof of this proposition mainly follows the same path as its deterministic counterpart. However, due to stochasticity, a residual term (the last term in Equation (18)) quantifying the difference between the stochastic gradient estimate and the true gradient of the objective function (compare

Equation (18) to Theorem 1). Following the exact same steps of Appendix B.1, we obtain by replacing the deterministic gradient $\nabla f(x_n)$ by its stochastic estimate $\nabla f(x_n, \xi_{n+1})$:

$$\begin{aligned} H_{n+1} &\leq H_n - \langle a_{n+1}p_{n+1}^2, 1 - \frac{a_{n+1}L}{2} - \frac{|b - (1 - \alpha)|}{2u} - \frac{1 - \alpha}{2b} \rangle \\ &\quad - \frac{b}{2} \langle a_{n+1}(\nabla f(x_n, \xi_{n+1}) - p_n)^2, \left(1 - \frac{|b - (1 - \alpha)|u}{b} - (1 - \alpha)\right) \mathbf{1} \rangle \\ &\quad + \langle \nabla f(x_n, \xi_{n+1}) - \nabla F(x_n), a_{n+1}p_{n+1} \rangle. \end{aligned} \quad (18)$$

Using the classical inequality $xy \leq \frac{x^2}{2} + \frac{y^2}{2}$ and the almost sure boundedness of the step size a_{n+1} , we get :

$$\begin{aligned} \langle \nabla f(x_n, \xi_{n+1}) - \nabla F(x_n), a_{n+1}p_{n+1} \rangle &\leq \langle \frac{1}{2}(\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n))^2 + \frac{1}{2}p_{n+1}^2, a_{n+1} \rangle \\ &\leq \frac{a_{\text{sup}}}{2} \|\nabla f(x_n, \xi_{n+1}) - \nabla F(x_n)\|^2 + \frac{1}{2} \langle a_{n+1}, p_{n+1}^2 \rangle. \end{aligned}$$

Therefore, taking the expectation and using the boundedness of the variance, we obtain from Equation (18) :

$$\mathbb{E}[H_{n+1}] - \mathbb{E}[H_n] \leq -\mathbb{E} \left[\langle a_{n+1}p_{n+1}^2, 1 - \frac{a_{n+1}L}{2} - \frac{|b - (1 - \alpha)|}{2u} - \frac{1 - \alpha}{2b} \rangle \right] + \frac{a_{\text{sup}}\sigma^2}{2}.$$

Then, the proof follows the lines of Appendix B.2. Hence, we have

$$\mathbb{E}[H_{n+1}] - \mathbb{E}[H_n] \leq -\mathbb{E} [\langle a_{n+1}p_{n+1}^2, \varepsilon \mathbf{1} \rangle] + \frac{a_{\text{sup}}\sigma^2}{2}.$$

We sum these inequalities for $k = 0, \dots, n - 1$ and rearrange the terms to obtain

$$\mathbb{E} \left[\sum_{k=0}^{n-1} \langle a_{k+1}, p_{k+1}^2 \rangle \right] \leq \frac{H_0 - \inf f}{\varepsilon} + \frac{na_{\text{sup}}\sigma^2}{2\varepsilon}.$$

Then following the derivations in Appendix B.2 using $\nabla f(x_k, \xi_{k+1}) = \frac{1}{b}p_{k+1} - \frac{1-b}{b}p_k$, we establish the following inequality

$$\mathbb{E} \left[\sum_{k=0}^{n-1} \langle a_{k+1}, \nabla f(x_k, \xi_{k+1})^2 \rangle \right] \leq \frac{4}{b^2\alpha} \left(\frac{H_0 - \inf f}{\varepsilon} + \langle a_0, p_0^2 \rangle + \frac{na_{\text{sup}}\sigma^2}{2\varepsilon} \right).$$

Finally, we apply Jensen's inequality to $\|\cdot\|^2$, we inject the assumption $a_{n+1} \geq \delta$ and we divide the previous inequality by n to obtain the sought result

$$\frac{1}{n} \sum_{k=0}^{n-1} \mathbb{E} [\|\nabla F(x_k)\|^2] \leq \frac{4}{n\delta b^2\alpha} \left(\frac{H_0 - \inf f}{\varepsilon} + \langle a_0, p_0^2 \rangle + \frac{na_{\text{sup}}\sigma^2}{2\varepsilon} \right).$$

Appendix C. Appendix C: proofs of Section 4

C.1. Conditions similar to "gradient-like descent sequences" conditions [8, Appendix][29]

We introduce a lemma which will be useful in our analysis later.

Lemma 8 *Let $(z_k)_{k \in \mathbb{N}}$ be the sequence defined for all $k \in \mathbb{N}$ by $z_k = (x_k, y_k)$ where $y_k = \sqrt{a_k} p_k$ and (x_k, p_k) is generated by Algorithm (1) from a starting point z_0 . Let Assumptions 2.1 and 3.1 hold true. Assume moreover that condition (4) holds. Then,*

(i) (sufficient decrease property) *There exists a positive scalar ρ_1 s.t. :*

$$H(z_{k+1}) - H(z_k) \leq -\rho_1 \|x_{k+1} - x_k\|^2 \quad \forall k \in \mathbb{N}.$$

(ii) *There exists a positive scalar ρ_2 s.t. :*

$$\|\nabla H(z_{k+1})\| \leq \rho_2 (\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|) \quad \forall k \geq 1.$$

(iii) (continuity condition) *If \bar{z} is a limit point of a subsequence $(z_{k_j})_{j \in \mathbb{N}}$, then $\lim_{j \rightarrow +\infty} H(z_{k_j}) = H(\bar{z})$.*

Remark 9 *Conditions in Bolte et al. [8, Appendix] corresponding to Theorem 8 are more general. Authors consider a nonsmooth objective function and introduce the Fréchet subdifferential instead of the gradient. For our purposes, this is sufficient.*

Proof

(i) From Theorems 1 and 2, we get for all $k \in \mathbb{N}$:

$$H(z_{k+1}) - H(z_k) \leq -\varepsilon \langle a_{k+1}, p_{k+1}^2 \rangle \leq -\varepsilon \langle a_{k+1}, \left(\frac{x_{k+1} - x_k}{-a_{k+1}} \right)^2 \rangle \leq -\frac{\varepsilon}{a_{\text{sup}}} \|x_{k+1} - x_k\|^2.$$

We set $\rho_1 := \frac{\varepsilon}{a_{\text{sup}}}$.

(ii) First, observe that for all $k \in \mathbb{N}$

$$\|\nabla H(z_{k+1})\| \leq \|\nabla f(x_{k+1})\| + \frac{1}{b} \|y_{k+1}\|. \quad (19)$$

Now, let us upperbound each one of these two terms. Recall that we can rewrite our algorithm under a "Heavy-ball"-like form as follows:

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) + \beta_k (x_k - x_{k-1}) \quad \forall k \geq 1.$$

where $\alpha_k := b a_{k+1}$ and $\beta_k = (1 - b) \frac{a_{k+1}}{a_k}$ are vectors.

On the one hand,

$$\begin{aligned} \|\nabla f(x_{k+1})\|^2 &\leq 2 (\|\nabla f(x_{k+1}) - \nabla f(x_k)\|^2 + \|\nabla f(x_k)\|^2) \\ &\leq 2 (L^2 \|x_{k+1} - x_k\|^2 + \|\nabla f(x_k)\|^2) \end{aligned} \quad (\text{L-Lipschitz continuity of the gradient})$$

Moreover,

$$\begin{aligned}
\|\nabla f(x_k)\|^2 &= \left\| \frac{x_k - x_{k+1}}{\alpha_k} + \frac{\beta_k}{\alpha_k} (x_k - x_{k-1}) \right\|^2 \\
&\leq 2 \left\| \frac{x_k - x_{k+1}}{b a_{k+1}} \right\|^2 + 2 \left\| \frac{1-b}{b} \frac{1}{a_k} (x_k - x_{k-1}) \right\|^2 \\
&\leq \frac{2}{b^2 \delta^2} \|x_{k+1} - x_k\|^2 + \frac{2(1-b)^2}{b^2 \delta^2} \|x_k - x_{k-1}\|^2 \\
&\leq \frac{2}{b^2 \delta^2} (\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2).
\end{aligned}$$

Hence,

$$\begin{aligned}
\|\nabla f(x_{k+1})\|^2 &\leq 2 (L^2 \|x_{k+1} - x_k\|^2 + \|\nabla f(x_k)\|^2) \\
&\leq 2 \left(L^2 + \frac{2}{b^2 \delta^2} \right) \|x_{k+1} - x_k\|^2 + \frac{4}{b^2 \delta^2} \|x_k - x_{k-1}\|^2 \\
&\leq 2 \left(L^2 + \frac{2}{b^2 \delta^2} \right) (\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2).
\end{aligned}$$

Therefore, the following inequality holds :

$$\|\nabla f(x_{k+1})\| \leq \sqrt{2 \left(L^2 + \frac{2}{b^2 \delta^2} \right)} (\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|).$$

On the otherhand,

$$\|y_{k+1}\| = \|\sqrt{a_{k+1}} p_{k+1}\| = \left\| \frac{x_{k+1} - x_k}{\sqrt{a_{k+1}}} \right\| \leq \frac{1}{\sqrt{\delta}} \|x_{k+1} - x_k\|.$$

Finally, combining the inequalities for both terms in Equation (19), we obtain

$$\|\nabla H(z_{k+1})\| \leq \rho_2 (\|x_{k+1} - x_k\| + \|x_k - x_{k-1}\|) \quad \forall k \geq 1.$$

$$\text{with } \rho_2 := \left(\sqrt{2 \left(L^2 + \frac{2}{b^2 \delta^2} \right)} + \frac{1}{b\sqrt{\delta}} \right).$$

(iii) This is a consequence of the continuity of H . ■

C.2. Properties of the limit point set

Notations and definitions. If (E, d) is a metric space, $z \in E$ and A is a non-empty subset of E , we use the notation $d(z, A) := \inf\{d(z, z') : z' \in A\}$. The set of critical points of the function H is defined by $\text{crit } H := \{z \in \mathbb{R}^{2d} \text{ s.t. } \nabla H(z) = 0\}$.

Definition 10 (*Set of limit points*) The set of all limit points of $(z_k)_{k \in \mathbb{N}}$ initialized at z_0 is defined by $\omega(z_0) := \{\bar{z} \in \mathbb{R}^{2d} : \exists \text{ an increasing sequence of integers } (k_j)_{j \in \mathbb{N}} \text{ s.t. } z_{k_j} \rightarrow \bar{z} \text{ as } j \rightarrow \infty\}$.

Lemma 11 (*Properties of the limit point set*) Let $(z_k)_{k \in \mathbb{N}}$ be the sequence defined for all $k \in \mathbb{N}$ by $z_k = (x_k, y_k)$ where $y_k = \sqrt{\alpha_k} p_k$ and (x_k, p_k) is generated by Algorithm (1) from a starting point z_0 . Let Assumptions 2.1, 3.1 and 4.1 hold true. Assume that Condition (4) holds. Then,

- (i) $\omega(z_0)$ is a nonempty compact set.
- (ii) $\omega(z_0) \subset \text{crit}H = \text{crit}f \times \{0\}$.
- (iii) $\lim_{k \rightarrow +\infty} d(z_k, \omega(z_0)) = 0$.
- (iv) H is finite and constant on $\omega(z_0)$.

Proof

- (i) By Theorem 2, the sequence $(H(z_n))_{n \in \mathbb{N}}$ is nonincreasing. Therefore, for all $n \in \mathbb{N}$, $H(z_n) \leq H(z_0)$ and hence $z_n \in \{z : H(z) \leq H(z_0)\}$. Since f is coercive, H is also coercive and its level sets are bounded. As a consequence, $(z_n)_{n \in \mathbb{N}}$ is bounded and there exist $z_* \in \mathbb{R}^d$ and a subsequence $(z_{k_j})_{j \in \mathbb{N}}$ s.t. $z_{k_j} \rightarrow z_*$ as $j \rightarrow \infty$. Hence, $\omega(z_0) \neq \emptyset$. Furthermore, $\omega(z_0) = \bigcap_{q \in \mathbb{N}} \overline{\bigcup_{k \geq q} \{z_k\}}$ is compact as an intersection of compact sets.
- (ii) First, $\text{crit}H = \text{crit}f \times \{0\}$ because $\nabla H(z) = (\nabla f(x), y/b)^T$. Let $z_* \in \omega(z_0)$. Recall that $x_{k+1} - x_k \rightarrow 0$ as $k \rightarrow \infty$ by Theorem 2. We deduce from the second assertion of Theorem 8 that $\nabla H(z_k) \rightarrow 0$ as $k \rightarrow \infty$. As $z_* \in \omega(z_0)$, there exists a subsequence $(z_{k_j})_{j \in \mathbb{N}}$ converging to z_* . Then, by Lipschitz continuity of ∇H , we get that $\nabla H(z_{k_j}) \rightarrow \nabla H(z_*)$ as $j \rightarrow \infty$. Finally, $\nabla H(z_*) = 0$ since $\nabla H(z_k) \rightarrow 0$ and $(\nabla H(z_{k_j}))_{j \in \mathbb{N}}$ is a subsequence of $(\nabla H(z_n))_{n \in \mathbb{N}}$.
- (iii) This point stems from the definition of limit points. Every subsequence of the sequence $(d(z_k, \omega(z_0)))_{k \in \mathbb{N}}$ converges to zero as a consequence of the definition of $\omega(z_0)$.
- (iv) The sequence $(H(z_n))_{n \in \mathbb{N}}$ is nonincreasing by Theorem 2. It is also bounded from below because $H(z_k) \geq f(x_k) \geq \inf f$ for all $k \in \mathbb{N}$. Hence we can denote by l its limit. Let $\bar{z} \in \omega(z_0)$. There there exists a subsequence $(z_{k_j})_{j \in \mathbb{N}}$ converging to \bar{z} as $j \rightarrow \infty$. By the third assertion of Theorem 8, $\lim_{j \rightarrow +\infty} H(z_{k_j}) = H(\bar{z})$. Hence this limit equals l since $(H(z_n))_{n \in \mathbb{N}}$ converges towards l . Therefore, the restriction of H to $\omega(z_0)$ equals l . ■

C.3. Proof of Theorem 6

We introduce now a uniformized version of the KŁ property which will be useful for our analysis.

Lemma 12 (*Uniformized KŁ property, Bolte et al. [7, Lemma 6, p 478]*) Let Ω be a compact set and let $H : \mathbb{R}^{2d} \rightarrow (-\infty, +\infty]$ be a proper l.s.c function. Assume that H is constant on Ω and satisfies the KŁ property at each point of Ω . Then, there exist $\varepsilon > 0, \eta > 0$ and $\varphi \in \Phi_\eta$ such that for all $\bar{z} \in \Omega$, for all $z \in \{z \in \mathbb{R}^d : d(z, \Omega) < \varepsilon\} \cap [H(\bar{z}) < H < H(\bar{z}) + \eta]$, one has

$$\varphi'(H(z) - H(\bar{z})) \|\nabla H(z)\| \geq 1 \quad (20)$$

The first step of this proof follows the same path as Bolte et al. [8, Proof of Theorem 6.2, Appendix]. Since f is coercive, H is also coercive. The sequence $(H(z_k))_{k \in \mathbb{N}}$ is nonincreasing. Hence, (z_k) is bounded and there exists a subsequence $(z_{k_q})_{q \in \mathbb{N}}$ and $\bar{z} \in \mathbb{R}^{2d}$ s.t. $z_{k_q} \rightarrow \bar{z}$ as $q \rightarrow \infty$. Then, since $(H(z_k))_{k \in \mathbb{N}}$ is nonincreasing and lowerbounded by $\inf f$, it is convergent and we obtain by continuity of H ,

$$\lim_{k \rightarrow +\infty} H(z_k) = H(\bar{z}). \quad (21)$$

If there exists $\bar{k} \in \mathbb{N}$ s.t. $H(z_{\bar{k}}) = H(\bar{z})$, then $H(z_{\bar{k}+1}) = H(\bar{z})$ and by the first point of Theorem 8, $x_{\bar{k}+1} = x_{\bar{k}}$ and then $(x_k)_{k \in \mathbb{N}}$ is stationary and for all $k \geq \bar{k}$, $H(z_k) = H(\bar{z})$ and the results of the theorem hold in this case (note that $\bar{z} \in \text{crit}H$ by Theorem 11). Therefore, we can assume now that $H(\bar{z}) < H(z_k) \forall k > 0$ since $(H(z_k))_{k \in \mathbb{N}}$ is nonincreasing and Equation (21) holds. One more time, from Equation (21), we have that for all $\eta > 0$, there exists $k_0 \in \mathbb{N}$ s.t. $H(z_k) < H(\bar{z}) + \eta$ for all $k > k_0$. From Theorem 11, we get $d(z_k, \omega(z_0)) \rightarrow 0$ as $k \rightarrow +\infty$. Hence, for all $\varepsilon > 0$, there exists $k_1 \in \mathbb{N}$ s.t. $d(z_k, \omega(z_0)) < \varepsilon$ for all $k > k_1$. Moreover, $\omega(z_0)$ is a nonempty compact set and H is finite and constant on it. Therefore, we can apply the uniformization Theorem 12 with $\Omega = \omega(z_0)$. Hence, for any $k > l := \max(k_0, k_1)$, we get

$$\varphi'(H(z_k) - H(\bar{z}))^2 \|\nabla H(z_k)\|^2 \geq 1. \quad (22)$$

This completes the first step of the proof. In the second step, we follow the proof of Johnstone and Moulin [17, Theorem 2]. Using Theorem 8 .(i)-(ii), we can write for all $k \geq 1$,

$$\|\nabla H(z_{k+1})\|^2 \leq 2\rho_2^2 (\|x_{k+1} - x_k\|^2 + \|x_k - x_{k-1}\|^2) \leq \frac{2\rho_2^2}{\rho_1} (H(z_{k-1}) - H(z_{k+1})).$$

Injecting the last inequality in Equation (22), we obtain for all $k > k_2 := \max(l, 2)$,

$$\frac{2\rho_2^2}{\rho_1} \varphi'(H(z_k) - H(\bar{z}))^2 (H(z_{k-2}) - H(z_k)) \geq 1.$$

Now, use $\varphi'(s) = \bar{c}s^{\theta-1}$ to derive the following for all $k > k_2$:

$$[H(z_{k-2}) - H(\bar{z})] - [H(z_k) - H(\bar{z})] \geq \frac{\rho_1}{2\rho_2^2 \bar{c}^2} [H(z_k) - H(\bar{z})]^{2(1-\theta)}. \quad (23)$$

Let $r_k := H(z_k) - H(\bar{z})$ and $C_1 = \frac{\rho_1}{2\rho_2^2 \bar{c}^2}$. Then, we can rewrite Equation (23) as

$$r_{k-2} - r_k \geq C_1 r_k^{2(1-\theta)} \quad \forall k > k_2. \quad (24)$$

We distinguish three different cases to obtain the sought results.

(i) $\theta = 1$:

Suppose $r_k > 0$ for all $k > k_2$. Then, since we know that $r_k \rightarrow 0$ by Equation (21), C_1 must be equal to 0. This is a contradiction. Therefore, there exist $k_3 \in \mathbb{N}$ s.t. $r_k = 0$ for all $k > k_3$ (recall that $(r_k)_{k \in \mathbb{N}}$ is nonincreasing).

(ii) $\theta \geq \frac{1}{2}$:

As $r_k \rightarrow 0$, there exists $k_4 \in \mathbb{N}$ s.t. for all $k \geq k_4$, $r_k \leq 1$. Observe that $2(1 - \theta) \leq 1$ and hence $r_{k-2} - r_k \geq C_1 r_k$ for all $k > k_2$ and then

$$r_k \leq (1 + C_1)^{-1} r_{k-2} \leq (1 + C_1)^{-p_1} r_{k_4}. \quad (25)$$

where $p_1 := \lfloor \frac{k-k_4}{2} \rfloor$. Notice that $p_1 > \frac{k-k_4-2}{2}$. Thus, the linear convergence result follows. Note also that if $\theta = 1/2$, $2(1 - \theta) = 1$ and Equation (25) holds for all $k > k_2$.

(iii) $\theta < \frac{1}{2}$:

Define the function h by $h(t) = \frac{D}{1-2\theta} t^{2\theta-1}$ where $D > 0$ is a constant. Then,

$$h(r_k) - h(r_{k-2}) = \int_{r_{k-2}}^{r_k} h'(t) dt = D \int_{r_{k-2}}^{r_k} t^{2\theta-2} dt \geq D (r_{k-2} - r_k) r_{k-2}^{2\theta-2}.$$

We disentangle now two cases :

(a) Suppose $2r_{k-2}^{2\theta-2} \geq r_k^{2\theta-2}$. Then, by Equation (24), we get

$$h(r_k) - h(r_{k-2}) = D (r_{k-2} - r_k) r_{k-2}^{2\theta-2} \geq \frac{C_1 D}{2}. \quad (26)$$

(b) Suppose now the opposite inequation $2r_{k-2}^{2\theta-2} < r_k^{2\theta-2}$. We can suppose without loss of generality that r_k are all positive. Otherwise, if there exists p such that $r_p = 0$, the sequence $(r_k)_{k \in \mathbb{N}}$ will be stationary at 0 for all $k \geq p$. Observe that $2\theta - 2 < 2\theta - 1 < 0$, thus $\frac{2\theta-1}{2\theta-2} > 0$. As a consequence, we can write in this case $r_k^{2\theta-1} > q r_{k-2}^{2\theta-1}$ where $q := 2^{\frac{2\theta-1}{2\theta-2}} > 1$. Therefore, using moreover that the sequence $(r_k)_{k \in \mathbb{N}}$ is nonincreasing and $2\theta - 1 < 0$, we derive the following

$$h(r_k) - h(r_{k-2}) = \frac{D}{1-2\theta} (r_k^{2\theta-1} - r_{k-2}^{2\theta-1}) > \frac{D}{1-2\theta} (q-1) r_{k-2}^{2\theta-1} > \frac{D}{1-2\theta} (q-1) r_{k_2}^{2\theta-1} := C_2. \quad (27)$$

Combining Equation (26) and Equation (27) yields $h(r_k) \geq h(r_{k-2}) + C_3$ where $C_3 := \min(C_2, \frac{C_1 D}{2})$. Consequently, $h(r_k) \geq h(r_{k-2p_2}) + p_2 C_3$ where $p_2 := \lfloor \frac{k-k_2}{2} \rfloor$. We deduce from this inequality that

$$h(r_k) \geq h(r_k) - h(r_{k-2p_2}) \geq p_2 C_3.$$

Therefore, rearranging this inequality using the definition of h , we obtain $r_k^{1-2\theta} \leq \frac{D}{1-2\theta} (C_3 p_2)^{-1}$.

Then, since $p_2 > \frac{k-k_2-2}{2}$,

$$r_k \leq C_4 p_2^{\frac{1}{2\theta-1}} \leq C_4 \left(\frac{k - k_2 - 2}{2} \right)^{\frac{1}{2\theta-1}}.$$

where $C_4 := \left(\frac{C_3 (1-2\theta)}{D} \right)^{\frac{1}{2\theta-1}}$.

We conclude the proof by observing that $f(x_k) \leq H(z_k)$ and recalling that $\bar{z} \in \text{crit}H$.

C.4. Corollary to Theorem 6

We introduce a lemma in order to make the KL assumption on the objective function f instead of the auxiliary function H .

Lemma 13 *Let f be a continuously differentiable function satisfying the KL property at \bar{x} with an exponent of $\theta \in (0, 1/2]$. Then the function H defined in Equation (6) has also the KL property at $(\bar{x}, 0)$ with an exponent of θ .*

Proof

Since f has the KL property at \bar{x} with an exponent $\theta \in (0, 1/2]$, there exist c, ε and $\nu > 0$ s.t.

$$\|\nabla f(x)\|^{1-\theta} \geq c(f(x) - f(\bar{x})) \quad (28)$$

for all $x \in \mathbb{R}^d$ s.t. $\|x - \bar{x}\| \leq \varepsilon$ and $f(x) < f(\bar{x}) + \nu$ where condition $f(\bar{x}) - f(x)$ is dropped because Equation (28) holds trivially otherwise. Let $z = (x, y) \in \mathbb{R}^{2d}$ be s.t. $\|x - \bar{x}\| \leq \varepsilon$, $\|y\| \leq \varepsilon$ and $H(\bar{x}, 0) < H(x, y) < H(\bar{x}, 0) + \nu$. We assume that $\varepsilon < b$ (ε can be shrunk if needed). We have $f(x) \leq H(x, y) < H(\bar{x}, 0) + \nu = f(\bar{x}) + \nu$. Hence Equation (28) holds for these x .

By concavity of $u \mapsto u^{\frac{1}{2(1-\theta)}}$, we obtain

$$\|\nabla H(x, y)\|^{1-\theta} \geq C_0 \left(\|\nabla f(x)\|^{1-\theta} + \left\| \frac{y}{b} \right\|^{1-\theta} \right)$$

where $C_0 := 2^{\frac{1}{2(1-\theta)}-1}$.

Hence, using Equation (28), we get

$$\|\nabla H(x, y)\|^{1-\theta} \geq C_0 \left(c(f(x) - f(\bar{x})) + \left\| \frac{y}{b} \right\|^{1-\theta} \right).$$

Observe now that $\frac{1}{1-\theta} \geq 2$ and $\left\| \frac{y}{b} \right\| \leq \frac{\varepsilon}{b} \leq 1$. Therefore, $\left\| \frac{y}{b} \right\|^{1-\theta} \geq \|y/b\|^2$. Finally,

$$\begin{aligned} \|\nabla H(x, y)\|^{1-\theta} &\geq C_0 \left(c(f(x) - f(\bar{x})) + \frac{2}{b} \frac{1}{2b} \|y\|^2 \right) \\ &\geq C_0 \min \left(c, \frac{2}{b} \right) \left(f(x) - f(\bar{x}) + \frac{1}{2b} \|y\|^2 \right) \\ &= C_0 \min \left(c, \frac{2}{b} \right) (H(x, y) - H(\bar{x}, 0)). \end{aligned}$$

This completes the proof. ■

The following result derives a convergence rate on the objective function values under a KL assumption on this same function instead of an assumption on the Lyapunov function H . The result is an immediate consequence of Theorem 13 and Theorem 6.

Corollary 14 *Let $(z_k)_{k \in \mathbb{N}}$ be the sequence defined for all $k \in \mathbb{N}$ by $z_k = (x_k, y_k)$ where $y_k = \sqrt{a_k} p_k$ and (x_k, p_k) is generated by Algorithm (1) from a starting point z_0 . Let Assumptions 2.1, 3.1 and 4.1 hold true. Assume that Condition (4) holds. Suppose moreover that f is a KL function with KL exponent $\theta \in (0, 1/2)$. Denote by $f(x_*)$ the limit of the sequence $(H(z_k))_{k \in \mathbb{N}}$ where x_* is a critical point of f . Then $f(x_k) - f(x_*) = O(k^{\frac{1}{2\theta-1}})$.*