Appendix A. First Order Convergence Rate

A.1. Variants of Adam

We list most of the existing variants of the Adam algorithm together with their theoretical convergence guarantees in Table 2.

Remark 13 *The average regret bound result in the last line of Table 2 figures in Luo et al. (2019). Actually, according to Savarese (2019), slightly different assumptions on the bound functions should be considered to guarantee this regret rate.*

A.2. Proof of Lemma 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:

$$f(x_{n+1}) \leq f(x_n) - \langle \nabla f(x_n), a_{n+1}p_{n+1} \rangle + \frac{L}{2}\|a_{n+1}p_{n+1}\|^2 \tag{10}$$

Moreover,

$$\frac{1}{2b}\langle a_{n+1}, p_{n+1}^2 \rangle - \frac{1}{2b}\langle a_n, p_n^2 \rangle = \frac{1}{2b}\langle a_{n+1}, p_{n+1}^2 - p_n^2 \rangle + \frac{1}{2b}\langle a_{n+1} - a_n, p_n^2 \rangle. \tag{11}$$

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 :

$$H_{n+1} \leq H_n + \frac{L}{2}\|a_{n+1}p_{n+1}\|^2 - \frac{b}{2}\langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle - \langle a_{n+1}p_{n+1}, p_n \rangle + \frac{1}{2b}\langle a_{n+1} - a_n, p_n^2 \rangle. \tag{12}$$

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

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

Under Assumption 2, 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:

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

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

$$(b-(1-\alpha))a_{n+1}p_{n+1}(\nabla f(x_n) - p_n) \leq \frac{|b-(1-\alpha)|}{2u}\langle a_{n+1}, p_{n+1}^2 \rangle + \frac{|b-(1-\alpha)|u}{2}\langle a_{n+1}, (\nabla f(x_n) - p_n)^2 \rangle. \tag{13}$$

Table 2: **Theoretical guarantees of variants of Adam.** The gradient is supposed L-lipschitz continuous in all the convergence results. $g_{1:T,i} = [g_{1,i}, g_{2,i}, \cdots, g_{T,i}]^T$.

Algorithm	**Effective step size a_{n+1}**	b_n	c_n	Assumptions	Convergence Result
AMSGRAD$^{(1)}$, ADAMNC$^{(2)}$ (Reddi et al., 2018)	$\frac{a_0}{\sqrt{n}}\frac{1}{\sqrt{\hat{v}_n}}$ $^{(1)}\hat{v}_{n+1} = \max(\hat{v}_n, (1-c_n)v_n + c_n g_n^2)$ $^{(2)}\hat{v}_{n+1} = (1-c_n)\hat{v}_n + c_n g_n^2$	$1 - b_1\lambda^{n-1}$ or $1 - \frac{b_1}{n}$	$c_n \equiv c_1$ $\frac{c_1}{n}$ (for ADAMNC)	• convex functions • bounded gradients • bounded feasible set • $\sum_{i=1}^d \hat{v}_{T,i}^{1/2} \leq d$ (AMSGRAD) • $\sum_{i=1}^d \lVert g_{1:T,i}\rVert_2 \leq \sqrt{dT}$ • $b_1 < \sqrt{c_1}$(ADAMNC)	$R_T/T = O(\sqrt{logT/T})$ $\frac{R_T}{T} = O(1/\sqrt{T})$ (ADAMNC)
ADAM (De et al., 2018)	$\frac{4\lVert g_n\rVert^2\eta}{3L(1-(1-b)^n)^2(\eta+2\sigma)^2}\frac{1}{\epsilon+\sqrt{v_n}}$ $v_{n+1} = (1-c_1)v_n + c_1 g_n^2$	$b_n \equiv b_1$ $= 1 - \frac{\eta}{\eta+2\sigma}$	$c_n \equiv c_1$	• 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. $\lVert g_n\rVert \leq \eta$
PADAM,AMSGRAD (Zhou et al., 2018)	$\frac{1}{\sqrt{N}}\frac{1}{\hat{v}_n^p}$ $\frac{1}{\sqrt{dN}}\frac{1}{\hat{v}_n^{\frac{1}{2}}}$ (AMSGRAD) $\hat{v}_n = \max(\hat{v}_{n-1}, (1-c)v_{n-1} + cg_n^2)$	$b_n \equiv b$	$c_n \equiv c$	• bounded gradients For PADAM: • $p \in [0, \frac{1}{4}]$ • $1 - b < (1-c)^{2p}$ • $\sum_{i=1}^d \lVert g_{1:N,i}\rVert_2 \leq \sqrt{dN}$ AMSGRAD: $p = \frac{1}{2}$ and $1 - b < 1 - c$	$\mathbb{E}[\lVert g_\tau\rVert^2] = O(\frac{1+\sqrt{d}}{\sqrt{N}} + \frac{d}{N})$ $= O(\sqrt{\frac{d}{N}} + \frac{d}{N})$ (AMSGRAD) τ uniform r.v in $\{1, \cdots, N\}$
RMSPROP$^{(1)}$, YOGI$^{(2)}$ (Zaheer et al., 2018)	$\frac{a_1}{\epsilon+\sqrt{v_n}}$ $^{(1)}v_{n+1} = (1-c)v_n + cg_n^2$ $^{(2)}v_n = v_{n-1} - c\text{sign}(v_{n-1} - g_n^2)$	$b_n \equiv b$	$c_n \equiv c$	• G-bounded gradients • $a_1 \leq \frac{\epsilon\sqrt{1-c}}{2L}$ (YOGI) • $a_1 \leq \frac{\epsilon}{2L}$ • $c \leq \frac{\epsilon^2}{16G^2}$ • σ^2-bounded variance	$\mathbb{E}[\lVert g_\tau\rVert^2] = O(\frac{1}{N} + \sigma^2)$ τ uniform r.v in $\{1, \cdots, N\}$ $O(\frac{1}{N})$ if minibatch $\Theta(N)$
AMSGRAD$^{(1)}$, ADAFOM$^{(2)}$ (Chen et al., 2019)	$\frac{1}{\sqrt{n}}\frac{1}{\sqrt{\hat{v}_n}}$ $^{(1)}\ \hat{v}_{n+1} = \max(\hat{v}_n, (1-c_n)v_n + c_n g_n^2)$ $^{(2)}\ \hat{v}_{n+1} = (1-\frac{1}{n})\hat{v}_n + \frac{1}{n}g_n^2)$	non-increasing	$c_n \equiv c_1$	• bounded gradients • $\exists c > 0$ s.t. $\lvert g_{1,i}\rvert \geq c$	$\min_{n\in[0,N]} \mathbb{E}[\lVert g_n\rVert^2] = O(\frac{logN+d^2}{\sqrt{N}})$
GENERIC ADAM (Zou et al., 2019)	$\frac{\alpha_n}{\sqrt{v_n}}$ $v_{n+1} = (1-c_n)v_n + c_n g_n^2$ $\alpha_n = \hat{\alpha}\frac{\sqrt{1-(1-c)^n}}{1-(1-b)^n}$	$b_n \geq b > 0$	$0 < c_n < 1$ non-increasing $\lim c_n = c > b^2$	• bounded gradients in expectation • $d_n \leq \frac{\alpha_n}{\sqrt{c_n}} \leq c_0 d_n$ d_n non-increasing	$\mathbb{E}[\lVert g_\tau\rVert^{\frac{4}{3}}]^{\frac{3}{2}} \leq \frac{C+C'\sum_{n=1}^N \alpha_n\sqrt{c_n}}{N\alpha_N}$ τ uniform r.v in $\{1, \cdots, N\}$
ADABOUND$^{(1)}$, AMSBOUND$^{(2)}$ (Luo et al., 2019)	$\frac{1}{\sqrt{n}}\text{clip}(\frac{\alpha}{\sqrt{v_n}}, \eta_l(n), \eta_u(n))$ $\eta_l(n)$ 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 $1 - \frac{1-b}{n}$ $b_n \geq b$	$c_n \equiv c$	• bounded gradients • closed convex bounded feasible set • $1 - b < \sqrt{1-c}$	$R_T/T = O(1/\sqrt{T})$

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

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

This concludes the proof.

A.3. A first result under an upperbound of the step size

Proposition 14 *Let Assumption 1 hold true. Suppose moreover that $1-\alpha < b \le 1$. Let $\varepsilon > 0$ s.t. $a_{\sup} := \frac{2}{L}\left(1 - \frac{(b-(1-\alpha))^2}{2b\alpha} - \frac{1-\alpha}{2b} - \varepsilon\right)$ is nonnegative. Assume for all $n \in \mathbb{N}$,*

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

Then, for all $n \ge 1$,

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

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

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

We get the assumption made in the proposition by injecting the second condition into the first one and adding the assumption $\frac{a_{n+1}}{a_n} \le \frac{1}{\alpha}$ made in the lemma. Under this assumption, we sum over $0 \le k \le n-1$ Equation (4), rearrange it and use $A_{n+1} \ge \varepsilon$, $B \ge 0$ to obtain :

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

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

$$\sum_{k=0}^{n-1}\langle a_{k+1}, p_{k+1}^2\rangle \le \frac{H_0 - \inf f}{\varepsilon}\,. \tag{14}$$

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 $k \ge 0$:

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

We deduce that :

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

■

A.4. Proof of Theorem 2

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

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

We get the assumption made in the proposition by injecting the second condition into the first one and adding the assumption $\frac{a_{n+1}}{a_n} \le \alpha$ made in the lemma. Under this assumption, we sum over $0 \le k \le n-1$ Equation (4), rearrange it and use $A_{n+1} \ge \varepsilon$, $B \ge 0$ and $a_{k+1} \ge \delta$ to obtain :

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

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

$$
\sum_{k=0}^{n-1} \|p_{k+1}\|^2 \le \frac{H_0 - \inf f}{\delta\varepsilon} . \tag{15}
$$

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 $k \ge 0$:

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

We deduce that :

$$
\sum_{k=0}^{n-1}\|\nabla f(x_k)\|^2 \le \frac{2}{b^2}\sum_{k=0}^{n-1}(\|p_{k+1}\|^2 + \|p_k\|^2) = \frac{2}{b^2}\left(2\sum_{k=1}^{n-1}\|p_k\|^2 + \|p_n\|^2 + \|p_0\|^2\right) \le \frac{4}{b^2}\sum_{k=0}^{n}\|p_k\|^2 . \tag{16}
$$

Finally, using Equations (15) and (16), we have :

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

A.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 (17)) quantifying the difference between the stochastic gradient estimate and the true gradient of the objective function (compare Equation (17) to Lemma 1) remains. Following the exact same steps of Appendix A.2, we obtain by replacing the deterministic gradient $\nabla f(x_n)$ by its stochastic estimate $\nabla f(x_n,\xi_{n+1})$:

$$\begin{aligned} H_{n+1} \leq H_n &- \langle a_{n+1}p_{n+1}^2, 1-\frac{a_{n+1}L}{2}-\frac{|b-(1-\alpha)|}{2u}-\frac{1-\alpha}{2b}\rangle \\ &- \frac{b}{2}\langle a_{n+1}(\nabla f(x_n,\xi_{n+1})-p_n)^2, \left(1-\frac{|b-(1-\alpha)|u}{b}-(1-\alpha)\right)\mathbf{1}\rangle \\ &+ \langle \nabla f(x_n,\xi_{n+1})-\nabla F(x_n), a_{n+1}p_{n+1}\rangle . \end{aligned} \tag{17}$$

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

$$\begin{aligned} \langle \nabla f(x_n,\xi_{n+1})-\nabla F(x_n), a_{n+1}p_{n+1}\rangle &\leq \langle (\nabla f(x_n,\xi_{n+1})-\nabla F(x_n))^2+\frac{1}{4}p_{n+1}^2, a_{n+1}\rangle \\ &\leq \bar{a}_{\sup}\|\nabla f(x_n,\xi_{n+1})-\nabla F(x_n)\|^2+\frac{1}{4}\langle a_{n+1}, p_{n+1}^2\rangle . \end{aligned}$$

Therefore, taking the expectation and using the boundedness of the variance, we obtain from Equation (17) :

$$\mathbb{E}[H_{n+1}]-\mathbb{E}[H_n] \leq -\mathbb{E}\left[\langle a_{n+1}p_{n+1}^2, \frac{3}{4}-\frac{a_{n+1}L}{2}-\frac{|b-(1-\alpha)|}{2u}-\frac{1-\alpha}{2b}\rangle\right]+\bar{a}_{\sup}\sigma^2 .$$

Then, the proof follows the lines of Appendix A.3. Hence, we have

$$\mathbb{E}[H_{n+1}]-\mathbb{E}[H_n] \leq -\mathbb{E}\left[\langle a_{n+1}p_{n+1}^2, \varepsilon\mathbf{1}\rangle\right]+\bar{a}_{\sup}\sigma^2 .$$

We sum these inequalities for $k = 0,\cdots,n-1$, inject the assumption $a_{n+1} \geq \delta$ and rearrange the terms to obtain

$$\delta\,\mathbb{E}\left[\sum_{k=0}^{n-1}\|p_{k+1}\|^2\right] \leq \mathbb{E}\left[\sum_{k=0}^{n-1}\langle a_{k+1}, p_{k+1}^2\rangle\right] \leq \frac{H_0-\inf f}{\varepsilon}+\frac{n\bar{a}_{\sup}\sigma^2}{\varepsilon} . \tag{18}$$

Then, using $\nabla f(x_k,\xi_{k+1}) = \frac{1}{b}p_{k+1}-\frac{1-b}{b}p_k$ and a similar upperbound to Equation (16) we show that

$$\sum_{k=0}^{n-1} \|\nabla f(x_k, \xi_{k+1})\|^2 \le \frac{4}{b^2} \sum_{k=0}^{n} \|p_k\|^2 \,. \tag{19}$$

Therefore, combining Equations (18) and (19), we establish the following inequality

$$\mathbb{E}\left[\sum_{k=0}^{n-1} \|\nabla f(x_k, \xi_{k+1})\|^2\right] \le \frac{4}{b^2}\left(\frac{H_0 - \inf f}{\delta\varepsilon} + \|p_0\|^2\right) + \frac{4\bar{a}_{\sup} n}{\delta\varepsilon b^2}\sigma^2 \,.$$

Finally, we apply Jensen's inequality to $\|\cdot\|^2$ and divide the previous inequality by n to obtain the sought result

$$\frac{1}{n}\sum_{k=0}^{n-1} \mathbb{E}\left[\|\nabla F(x_k)\|^2\right] \le \frac{4}{n\delta b^2}\left(\frac{H_0 - \inf f}{\delta\varepsilon} + \|p_0\|^2\right) + \frac{4\bar{a}_{\sup}}{\delta\varepsilon b^2}\sigma^2 \,.$$

Remark 15 *Following the derivations in Appendix A.3, note that we also obtain the following result*

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

A.6. Comparison to Ochs et al. (2014)

We recall the conditions satisfied by α_n and β_n in Ochs et al. (2014) in order to traduce them in terms of the algorithm (1) at stake. Define :

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

Conditions of Ochs et al. (2014) write: $\alpha_n \ge c_1$ $\beta_n \ge 0$ $\delta_n \ge \gamma_n \ge c_2$ where c_1, c_2 are positive constants and (δ_n) is monotonically decreasing.

One can remark that algorithm (1) can be written as (3) with step sizes $\alpha_n = b a_{n+1}$ 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 $\gamma_n \ge c_2$ is equivalent to

$$\frac{a_{n+1}}{a_n} \le \frac{2}{2 - b(2 - a_n L)}. \tag{20}$$

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

$$\frac{1}{a_{n+1}} \le \frac{3-b}{2}\frac{1}{a_n} - \frac{1-b}{2a_{n-1}} \qquad \text{for} \qquad n \ge 1. \tag{21}$$

Note that we get rid of condition (21) while allowing adaptive step sizes a_n (see Proposition 14).

A.7. Performance of gradient descent in the nonconvex setting.

In the nonconvex setting, for a smooth function f, we cannot say anything about the convergence rate of the sequences $(f(x_k))$ and (x_k). Nevertheless, as exposed in (Nesterov, 2004, p.28), we can control the minimum of the gradients norms. We prove this result in the following for completeness.

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 for $n \geq 1$:

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

Appendix B. KŁ Convergence Analysis

B.1. Three abstract conditions

Inspired from the abstract convergence mechanism of Bolte et al. (2018, Appendix), we show that similar conditions hold in our case. We highlight that these conditions are slightly different here, since we do not deal with *gradient-like descent sequences* (for which the objective function is nonincreasing over the iterations). Conditions below are closer to those of Ochs et al. (2014) which studies a non-descent algorithm. Note however that the Lyapunov function H and the sequence (z_k) we consider are different.

Lemma 16 *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 1 and 2 hold true. Assume moreover that condition (5) 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. :

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

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

Remark 17 *Note that the conditions in Lemma 16 can be generalized to a nonsmooth objective function. Indeed, in Bolte et al. (2018, Appendix), the Fréchet subdifferential replaces the gradient.*

Proof

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

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

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

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

$$\|\nabla H(z_{k+1})\| \leq \|\nabla f(x_{k+1})\| + \frac{1}{b}\left\|y_{k+1}\right\|. \tag{22}$$

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, using the L-Lipschitz continuity of the gradient, we obtain

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

Moreover,

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

Hence,

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

Therefore, the following inequality holds :

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

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 (22), we obtain

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

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.

■

B.2. Proof of Lemma 6

(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. $z_{k_j} \to z_*$ as $j \to \infty$. 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, $\mathrm{crit} H = \mathrm{crit} f \times \{0\}$ because $\nabla H(z) = (\nabla f(x), y/b)^T$. Let $z_* \in \omega(z_0)$. Recall that $x_{k+1} - x_k \to 0$ as $k \to \infty$ by Theorem 2. We deduce from the second assertion of Lemma 16 that $\nabla H(z_k) \to 0$ as $k \to \infty$. 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 $\nabla H(z_{k_j}) \to \nabla H(z_*)$ as $j \to \infty$. Finally, $\nabla H(z_*) = 0$ since $\nabla H(z_k) \to 0$ 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 $(\mathsf{d}(z_k, \omega(z_0)))_{k\in\mathbb{N}}$ 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 $j \to \infty$. By the third assertion of Lemma 16, $\lim_{j\to+\infty} H(z_{k_j}) = H(\bar{z})$. 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.

B.3. Proof of Theorem 10

The first step of this proof follows the same path as Bolte et al. (2018, 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. $z_{k_q} \to \bar{z}$ as

$q \to \infty$. 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,

$$\lim_{k\to+\infty} H(z_k) = H(\bar{z})\,. \tag{23}$$

Using Theorem 2, observe that the sequence (y_k) converges to zero since (a_k) is bounded and $p_k \to 0$. 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 Lemma 16, $x_{\bar{k}+1} = x_{\bar{k}}$ 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 $\bar{z} \in \operatorname{crit} H$ by Lemma 6). 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 (23) holds. One more time, from Equation (23), 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 Lemma 6, we get $\mathsf{d}(z_k, \omega(z_0)) \to 0$ as $k \to +\infty$. Hence, for all $\varepsilon > 0$, there exists $k_1 \in \mathbb{N}$ s.t. $\mathsf{d}(z_k, \omega(z_0)) < \varepsilon$ 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 Lemma 8 with $\Omega = \omega(z_0)$. Hence, for any $k > l := \max(k_0, k_1)$, we get

$$\varphi'(H(z_k) - H(\bar{z}))^2 \, \|\nabla H(z_k)\|^2 \geq 1\,. \tag{24}$$

This completes the first step of the proof. In the second step, we follow the proof of Johnstone and Moulin (2017, Theorem 2). Using Lemma 16 .(i)-(ii), we can write for all $k \geq 1$,

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

Injecting the last inequality in Equation (24), 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 $k > k_2$:

$$[H(z_{k-2}) - H(\bar{z})] - [H(z_k) - H(\bar{z})] \geq \frac{\rho_1}{2\rho_2^2\, \bar{c}^2}[H(z_k) - H(\bar{z})]^{2(1-\theta)}\,. \tag{25}$$

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 (25) as

$$r_{k-2} - r_k \geq C_1 r_k^{2(1-\theta)} \quad \forall k > k_2\,. \tag{26}$$

We distinguish three different cases to obtain the sought results.

(i) $\underline{\theta = 1}$:
Suppose $r_k > 0$ for all $k > k_2$. Then, since we know that $r_k \to 0$ by Equation (23), 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) $\underline{\theta \geq \frac{1}{2}}$:
As $r_k \to 0$, 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

$$r_k \leq (1+C_1)^{-1} r_{k-2} \leq (1+C_1)^{-p_1} r_{k_4}\,. \tag{27}$$

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 (27) holds for all $k > k_2$.

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

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

We disentangle now two cases :

(a) Suppose $2r_{k-2}^{2\theta-2} \geq r_k^{2\theta-2}$. Then, by Equation (26), we get

$$h(r_k) - h(r_{k-2}) = D\,(r_{k-2} - r_k)\, r_{k-2}^{2\theta-2} \geq \frac{C_1\, D}{2}\,. \tag{28}$$

(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

$$h(r_k) - h(r_{k-2}) = \frac{D}{1-2\theta}(r_k^{2\theta-1} - r_{k-2}^{2\theta-1}) > \frac{D}{1-2\theta}\,(q-1) r_{k-2}^{2\theta-1} > \frac{D}{1-2\theta}\,(q-1) r_{k_2}^{2\theta-1} := C_2\,. \tag{29}$$

Combining Equation (28) and Equation (29) yields $h(r_k) \geq h(r_{k-2}) + C_3$ where $C_3 := \min(C_2, \frac{C_1\, D}{2})$. Consequently, $h(r_k) \geq h(r_{k-2\,p_2}) + p_2\, C_3$ where $p_2 := \lfloor \frac{k-k_2}{2} \rfloor$. We deduce from this inequality that

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

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 $\bar{z} \in \operatorname{crit} H$.

B.4. Proof of Lemma 11

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

$$\|\nabla f(x)\|^{\frac{1}{1-\theta}} \geq c(f(x) - f(\bar{x})) \tag{30}$$

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 (30) 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 (30) holds for these x.

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

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

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

Hence, using Equation (30), we get

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

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

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

This completes the proof.