Noname manuscript No.
(will be inserted by the editor)

Global convergence of the Heavy-ball method for convex optimization

Euhanna Ghadimi · Hamid Reza Feyzmahdavian · Mikael Johansson

Received: date / Accepted: date

Abstract This paper establishes global convergence and provides global bounds of the convergence rate of the Heavy-ball method for convex optimization problems. When the objective function has Lipschitz-continuous gradient, we show that the Cesáro average of the iterates converges to the optimum at a rate of $\mathcal{O}(1/k)$ where k is the number of iterations. When the objective function is also strongly convex, we prove that the Heavy-ball iterates converge linearly to the unique optimum.

Keywords Performance of first-order algorithms · Rate of convergence · Complexity · Smooth convex optimization · Heavy-ball method · Gradient method

Mathematics Subject Classification (2000) 90C26 · 90C30 · 49M37

1 Introduction

First-order convex optimization methods have a rich history dating back to 1950's [1–3]. Recently, these methods have attracted significant interest, both in terms of new theory [4–6] and in terms of applications in numerous areas such as signal processing [7], machine learning [8] and control [9]. One reason for this renewed interest is that first-order methods have a small per-iteration cost and are attractive in large-scale and distributed settings. But the development has also been fuelled by the development of accelerated methods with optimal convergence rates [10] and re-discovery of methods that are not only order-optimal, but also have optimal convergence times for smooth convex problems [11]. In spite of all this

This work was sponsored in part by the Swedish Foundation for Strategic Research (SSF) and the Swedish Research Council (VR).

E. Ghadimi · H. R. Feyzmahdavian · M. Johansson
Department of Automatic Control, School of Electrical Engineering and ACCESS Linnaeus Center, Royal Institute of Technology - KTH, Stockholm, Sweden.
E-mail: euhanna@kth.se

H. R. Feyzmahdavian
E-mail: hamidrez@kth.se
M. Johansson
E-mail: mikaelj@kth.se

progress, some very basic questions about the achievable convergence speed of first-order convex optimization methods are still open [6].

The basic first-order method is the gradient descent algorithm. For unconstrained convex optimization problems with objective functions that have Lipschitz-continuous gradient, the method produces iterates that are guaranteed to converge to the optimum at the rate $\mathcal{O}(1/k)$ where k is the number of iterations. When the objective function is also strongly convex, the iterates are guaranteed to converge at a linear rate [12].

In the early 1980's, Nemirovski and Yudin [13] proved that no first-order method can converge at a rate faster than $\mathcal{O}(1/k^2)$ on convex optimization problems with Lipschitz-continuous gradient. This created a gap between the guaranteed convergence rate of the gradient method and what could potentially be achieved. This gap was closed by Nesterov, who presented an accelerated first-order method that converges as $\mathcal{O}(1/k^2)$ [10]. Later, the method was generalized to also attain linear convergence rate for strongly convex objective functions, resulting in the first truly order-optimal first-order method for convex optimization [14]. The accelerated first-order methods combine gradient information at the current and the past iterate, as well as the iterates themselves [14]. For strongly convex problems, Nesterov's method can be tuned to yield a better convergence factor than the gradient iteration, but it is not known how small the convergence factor can be made.

When the objective function is twice continuously differentiable, strongly convex and has Lipschitz continuous gradient, the Heavy-ball method by Polyak [11] has linear convergence rate and better convergence factor than both the gradient and Nesterov's accelerated gradient method. The Heavy-ball method uses previous iterates when computing the next, but in contrast to Nesterov's method it only uses the gradient at the current iterate. Extensions of the Heavy-ball method to constrained and distributed optimization problems have confirmed its performance benefits over the standard gradient-based methods [15–17].

On the other hand, when the objective function is not necessary convex but has Lipschitz continuous gradient, Zavriev et al. [18] provided sufficient conditions for the Heavy-ball trajectories to converge to a stationary point. However, there are virtually no results on the rate of convergence of the Heavy-ball method for convex problems that are not necessarily twice-differentiable. Recently, Lessard et al [19] showed by an example that the Heavy-ball method does not necessarily converge on strongly convex (but not twice differentiable) objective functions even if one chooses step-size parameters according to Polyak's original stability criterion. In general, it is not clear whether the Heavy-ball method performs better than Nesterov's method, or even the basic gradient descent when the objective is not twice continuously differentiable.

The aim of this paper is to contribute to a more complete understanding of first-order methods for convex optimization. We provide a global convergence analysis for the Heavy-ball method on convex optimization problems with Lipschitz-continuous gradient, with and without the additional assumption of strong convexity. We show that if the parameters of the Heavy-ball method are chosen within certain ranges, the running average of the iterates converge to the optimal point at the rate $\mathcal{O}(1/k)$ when the objective function has Lipschitz continuous gradient. Moreover, for the same class of problems, we are able to show that the individual iterates themselves converge at rate $\mathcal{O}(1/k)$ if the Heavy-ball method uses (appropriately chosen) time-varying step-sizes. Finally, if the cost function is also strongly convex, we show that the iterates converge at a linear rate.

The rest of the paper is organized as follows. Section 2 reviews first-order convex optimization algorithms. Global convergence proofs for the Heavy-ball method are presented in Section 3 for objective functions with Lipschitz continuous gradient and in Section 4 for objective functions that are also strongly convex. Concluding remarks are given in Section 5.

1.1 Notation

We let $\mathbb{R}$, $\mathbb{N}$, and $\mathbb{N}_0$ denote the set of real numbers, the set of natural numbers, and the set of natural numbers including zero, respectively. The Euclidean norm is denoted by $\|\cdot\|$.

2 Background

We consider unconstrained convex optimization problems on the form

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x) \tag{1}$$

where $f:\mathbb{R}^n\to\mathbb{R}$ is a continuously differentiable convex function. We will provide convergence bounds for the Heavy-ball method for all functions in the following classes.

Definition 1 We say that $f:\mathbb{R}^n\to\mathbb{R}$ belongs to the class $\mathscr{F}_L^{1,1}$, if it is convex, continuously differentiable, and its gradient is Lipschitz continuous with constant L, *i.e.*,

$$0\leq f(y)-f(x)-\langle\nabla f(x),y-x\rangle\leq\frac{L}{2}\|x-y\|^2,$$

holds for all $x,y\in\mathbb{R}^n$. If f is also strongly convex with modulus $\mu>0$, *i.e.*,

$$\frac{\mu}{2}\|x-y\|^2\leq f(y)-f(x)-\langle\nabla f(x),y-x\rangle,\quad\forall x,y\in\mathbb{R}^n,$$

then, we say that f belongs to $\mathscr{S}_{\mu,L}^{1,1}$.

Our baseline first-order method is gradient descent:

$$x_{k+1}=x_k-\alpha\nabla f(x_k), \tag{2}$$

where α is a positive step-size parameter. Let $x^\star$ be an optimal point of (1) and $f^\star=f(x^\star)$. If $f\in\mathscr{F}_L^{1,1}$, then $f(x_k)-f^\star$ associated with the sequence $\{x_k\}$ in (2) converges at rate $\mathscr{O}(1/k)$. On the other hand, if $f\in\mathscr{S}_{\mu,L}^{1,1}$, then the sequence $\{x_k\}$ generated by the gradient descent method converges linearly, *i.e.*, there exists $q\in[0,1)$ such that

$$\|x_k-x^\star\|\leq q^k\|x_0-x^\star\|,\quad k\in\mathbb{N}_0.$$

The scalar q is called the *convergence factor*. The optimal convergence factor for $f\in\mathscr{S}_{\mu,L}^{1,1}$is $q=(L-\mu)/(L+\mu)$, attained for $\alpha=2/(L+\mu)$ [12].

The convergence of the gradient iterates can be accelerated by accounting for the history of iterates when computing the ones to come. Methods in which the next iterate depends not only on the current iterate but also on the preceding ones are called *multi-step methods*. The simplest multi-step extension of gradient descent is the Heavy-ball method:

$$x_{k+1}=x_k-\alpha\nabla f(x_k)+\beta\,(x_k-x_{k-1})\,, \tag{3}$$

for constant parameters $\alpha>0$ and $\beta>0$ [12]. For the class of twice continuously differentiable strongly convex functions with Lipschitz continuous gradient, Polyak used a local analysis to derive optimal step-size parameters and to show that the optimal convergence factor of the Heavy-ball iterates is $(\sqrt{L}-\sqrt{\mu})/(\sqrt{L}+\sqrt{\mu})$. This convergence factor is always smaller than the one associated with the gradient iterates, and significantly so when

the Hessian of the objective function is poorly conditioned. Note that this local analysis requires twice differentiability of the objective functions, and is, therefore, not valid for all $f \in \mathscr{F}_L^{1,1}$nor for all $f \in \mathscr{S}_{\mu,L}^{1,1}$.

In contrast, Nesterov's fast gradient method [14] is a first-order method with better convergence guarantees than the basic gradient method for objectives in $\mathscr{F}_L^{1,1}$and $\mathscr{S}_{\mu,L}^{1,1}$classes. In it's simplest form, Nesterov's algorithm with constant step-sizes takes the form

$$\begin{aligned} y_{k+1} &= x_k - \frac{1}{L}\nabla f(x_k), \\ x_{k+1} &= y_{k+1} + \beta(y_{k+1} - y_k). \end{aligned} \tag{4}$$

When $f \in \mathscr{S}_{\mu,L}^{1,1}$, the iterates produced by (4) with $\beta = (\sqrt{L}-\sqrt{\mu})/(\sqrt{L}+\sqrt{\mu})$ converge linearly towards the optimal point with a convergence factor $1-\sqrt{\mu/L}$. This factor is smaller than that of the gradient, but larger than that of the Heavy-ball method for twice-differentiable cost functions.

3 Global analysis of Heavy-ball algorithm for the class $\mathscr{F}_L^{1,1}$

In this section, we consider the Heavy-ball iterates (3) for the objective functions $f \in \mathscr{F}_L^{1,1}$. Our first result shows that the method is indeed guaranteed to converge globally and estimates the convergence rate of the Cesáro averages of the iterates.

Theorem 1 *Assume that $f \in \mathscr{F}_L^{1,1}$and that*

$$\beta \in [0,1), \quad \alpha \in \left(0, \frac{2(1-\beta)}{L}\right). \tag{5}$$

Then, the sequence $\{x_k\}$ generated by Heavy-ball iteration (3) *satisfies*

$$f(\bar{x}_T) - f^\star \leq \begin{cases} \frac{\|x_0-x^\star\|^2}{2(T+1)}\left(\frac{L\beta}{1-\beta} + \frac{1-\beta}{\alpha}\right), & \text{if } \alpha \in \big(0, \frac{1-\beta}{L}\big], \\ \frac{\|x_0-x^\star\|^2}{2(T+1)(2(1-\beta)-\alpha L)}\left(L\beta + \frac{(1-\beta)^2}{\alpha}\right), & \text{if } \alpha \in \big[\frac{1-\beta}{L}, \frac{2(1-\beta)}{L}\big), \end{cases} \tag{6}$$

where $\bar{x}_T$ is the Cesáro average of the iterates, i.e.,

$$\bar{x}_T = \frac{1}{T+1}\sum_{k=0}^{T} x_k.$$

Proof. Assume that $\beta \in [0,1)$, and let

$$p_k = \frac{\beta}{1-\beta}(x_k - x_{k-1}), \quad k \in \mathbb{N}_0. \tag{7}$$

Then

$$x_{k+1} + p_{k+1} = \frac{1}{1-\beta}x_{k+1} - \frac{\beta}{1-\beta}x_k \overset{(3)}{=} x_k + p_k - \frac{\alpha}{1-\beta}\nabla f(x_k),$$

which implies that

$$\begin{aligned}\|x_{k+1}+p_{k+1}-x^\star\|^2 =&\|x_k+p_k-x^\star\|^2-\frac{2\alpha}{1-\beta}\langle x_k+p_k-x^\star,\nabla f(x_k)\rangle+\left(\frac{\alpha}{1-\beta}\right)^2\|\nabla f(x_k)\|^2\\ \overset{(7)}{=}&\|x_k+p_k-x^\star\|^2-\frac{2\alpha}{1-\beta}\langle x_k-x^\star,\nabla f(x_k)\rangle\\ &-\frac{2\alpha\beta}{(1-\beta)^2}\langle x_k-x_{k-1},\nabla f(x_k)\rangle+\left(\frac{\alpha}{1-\beta}\right)^2\|\nabla f(x_k)\|^2. \end{aligned}\tag{8}$$

Since $f\in\mathscr{F}_L^{1,1}$, it follows from [14, Theorem 2.1.5] that

$$\begin{aligned}\frac{1}{L}\|\nabla f(x_k)\|^2&\le\langle x_k-x^\star,\nabla f(x_k)\rangle,\\ f(x_k)-f^\star+\frac{1}{2L}\|\nabla f(x_k)\|^2&\le\langle x_k-x^\star,\nabla f(x_k)\rangle,\\ f(x_k)-f(x_{k-1})&\le\langle x_k-x_{k-1},\nabla f(x_k))\rangle.\end{aligned}\tag{9}$$

Substituting the above inequalities into (8) yields

$$\begin{aligned}\|x_{k+1}+p_{k+1}-x^\star\|^2\le&\|x_k+p_k-x^\star\|^2-\frac{2\alpha(1-\lambda)}{L(1-\beta)}\|\nabla f(x_k)\|^2-\frac{2\alpha\lambda}{1-\beta}\big(f(x_k)-f^\star\big)\\ &-\frac{\alpha\lambda}{L(1-\beta)}\|\nabla f(x_k)\|^2-\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_k)-f(x_{k-1})\big)\\ &+\left(\frac{\alpha}{1-\beta}\right)^2\|\nabla f(x_k)\|^2,\end{aligned}$$

where $\lambda\in(0,1]$ is a parameter which we will use to balance the weights between the first two inequities in (9). Collecting the terms in the preceding inequality, we obtain

$$\begin{aligned}\frac{2\alpha}{(1-\beta)}\big(\lambda+\frac{\beta}{1-\beta}\big)\big(f(x_k)-f^\star\big)+\|x_{k+1}+p_{k+1}-x^\star\|^2&\\ \le\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_{k-1})-f^\star\big)+\|x_k+p_k-x^\star\|^2&\\ +\left(\frac{\alpha}{1-\beta}\right)\left(\frac{\alpha}{1-\beta}-\frac{2-\lambda}{L}\right)\|\nabla f(x_k)\|^2.&\end{aligned}\tag{10}$$

Note that when $\alpha\in[0,(2-\lambda)(1-\beta)/L]$, the last term of (10) becomes non-positive and, therefore, can be eliminated from the right-hand-side. Summing (10) over $k=0,\dots,T$ gives

$$\begin{aligned}\frac{2\alpha\lambda}{(1-\beta)}\sum_{k=0}^{T}\big(f(x_k)-f^\star\big)+\sum_{k=0}^{T}\left(\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_k)-f^\star\big)+\|x_{k+1}+p_{k+1}-x^\star\|^2\right)&\\ \le\sum_{k=0}^{T}\left(\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_{k-1})-f^\star\big)+\|x_k+p_k-x^\star\|^2\right),&\end{aligned}$$

which implies that

$$\frac{2\alpha\lambda}{(1-\beta)}\sum_{k=0}^{T}\big(f(x_k)-f^\star\big)\le\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_0)-f^\star\big)+\|x_0-x^\star\|^2.$$

Note that as f is convex, we have

$$(T+1)f(\bar{x}_T) \leq \sum_{k=0}^{T} f(x_k). \tag{11}$$

It now follows that

$$f(\bar{x}_T) - f^\star \leq \frac{1}{T+1}\left(\frac{\beta}{\lambda(1-\beta)}(f(x_0)-f^\star) + \frac{1-\beta}{2\alpha\lambda}\|x_0-x^\star\|^2\right). \tag{12}$$

Additionally, according to [14, Lemma 1.2.3], $f(x_0)-f^\star \leq (L/2)\|x_0-x^\star\|^2$. The proof is completed by replacing this upper bound in (12) and setting $\lambda = 1$ for $\alpha \in \big(0,(1-\beta)/L\big]$ and $\lambda = 2-(\alpha L)/(1-\beta)$ for $\alpha \in \big[(1-\beta)/L, 2(1-\beta)/L\big)$. □

A few remarks regarding the results of Theorem 1 are in order: first, a similar convergence rate can be proved for the minimum function values within T number of Heavy-ball iterates. More precisely, the sequence $\{x_k\}$ generated by (3) satisfies

$$\min_{0\leq k\leq T} f(x_k) - f^\star \leq \mathscr{O}\left(\frac{\|x_0-x^\star\|^2}{T}\right),$$

for all $T \in \mathbb{N}_0$. Second, for any fixed $\bar{\alpha} \in (0, 1/L]$, one can verify that the $\beta \in [0,1)$ which minimizes the convergence factor (6) is $\beta^\star = 1-\sqrt{\bar{\alpha}L}$ which yields the convergence factor

$$\min_{0\leq k\leq T} f(x_k) - f^\star \leq \frac{1}{2(T+1)}\left(\frac{2\sqrt{\bar{\alpha}L}-\bar{\alpha}L}{\bar{\alpha}}\right)\|x_0-x^\star\|^2.$$

Note that this convergence factor is always smaller than the one for the gradient descent method obtained by setting $\beta = 0$ in (6), i.e.,

$$f(x_T) - f^\star \leq \frac{1}{2\bar{\alpha}(T+1)}\|x_0-x^\star\|^2.$$

Finally, setting $\bar{\alpha} = 1/L$ in the preceding upper bounds, we see that the factors coincide and equal the best convergence factor of the gradient descent method reported in [7].

Next, we show that our analysis can be strengthened when we use (appropriately chosen) time-varying step-sizes in the Heavy-ball method. In this case, the individual iterates x_k (and not just their running average) converge with rate $\mathscr{O}(1/k)$.

Theorem 2 *Assume that $f \in \mathscr{F}_L^{1,1}$ and that*

$$\beta_k = \frac{k}{k+2}, \quad \alpha_k = \frac{\alpha_0}{k+2}, \quad k \in \mathbb{N}, \tag{13}$$

where $\alpha_0 \in (0, 1/L]$. Then, the sequence $\{x_k\}$ generated by Heavy-ball iteration (3) *satisfies*

$$f(x_T) - f^\star \leq \frac{\|x_0-x^\star\|^2}{2\alpha_0(T+1)}, \quad T \in \mathbb{N}. \tag{14}$$

Proof. The proof is similar to that of Theorem 1, so we will be somewhat terse. For $k \in \mathbb{N}_0$, let $p_k = k(x_k - x_{k-1})$. It is easy to verify that

$$x_{k+1} + p_{k+1} = x_k + p_k - \alpha_0 \nabla f(x_k),$$

which together with the inequalities in (8) implies that

$$2\alpha_0(k+1)\big(f(x_k) - f^\star\big) + \|x_{k+1} + p_{k+1} - x^\star\|^2 \leq 2\alpha_0 k\big(f(x_{k-1}) - f^\star\big) + \|x_k + p_k - x^\star\|^2.$$

Summing this inequality over $k = 0, \ldots, T$ gives

$$2\alpha_0(T+1)\big(f(x_T) - f^\star\big) + \|x_{T+1} + p_{T+1} - x^\star\|^2 \leq \|x_0 - x^\star\|^2.$$

The proof is complete. □

To illustrate our results, we evaluate the gradient method and the two variations of the Heavy-ball method on a numerical example. In this example, the objective function is the Moreau proximal envelope of the function $f(x) = (1/c)\|x\|$:

$$f(x) = \begin{cases} \dfrac{1}{c}\|x\| - \dfrac{1}{2c^2} & \|x\| \geq \dfrac{1}{c}, \\ \dfrac{1}{2}\|x\|^2 & \|x\| \leq \dfrac{1}{c}, \end{cases} \tag{15}$$

with $c = 5$ and $x \in \mathbb{R}^{50}$. One can verify that $f(x) \in \mathscr{F}_L^{1,1}$, i.e., it is convex and continuously differentiable with Lipschitz constant $L = 1$ [20]. First-order methods designed to find the minimum of this cost function are expected to pertain very poor convergence behavior [6]. For the Heavy-ball algorithm with constant step-sizes (3) we chose $\beta = 0.5$ and $\alpha = 1/L$, for the variant with time varying step-sizes (13) we used $\alpha_0 = 1/L$ whereas the gradient algorithm was implemented with the step-size $\alpha = 1/L$. Fig. 1 shows the progress of the objective values towards the optimal solution. The plot suggests that $\mathscr{O}(1/k)$ is a quite accurate convergence rate estimate for the Heavy-ball and the gradient method.

3.1 Convergence analysis of Nesterov's method with constant step-sizes

For objective functions $f \in \mathscr{S}_{\mu,L}^{1,1}$, it is possible to use constant step-sizes in Nesterov's method and still guarantee a linear rate of convergence [14]. For the objective functions on the class $\mathscr{F}_L^{1,1}$, however, to the best of our knowledge no convergence result exists for Nesterov's method with fixed step-sizes. Using a similar analysis as in the previous section, we can derive the following convergence rate bound.

Theorem 3 *Assume that $f \in \mathscr{F}_L^{1,1}$ and that $\beta \in [0,1)$. Then the sequence $\{x_k\}$ generated by Nesterov's iteration* (4) *satisfies*

$$f(\bar{x}_T) - f^\star \leq \frac{1}{T+1}\left(\frac{\beta}{1-\beta}(f(x_0) - f^\star) + \frac{L(1-\beta)}{2}\|x_0 - x^\star\|^2\right). \tag{16}$$

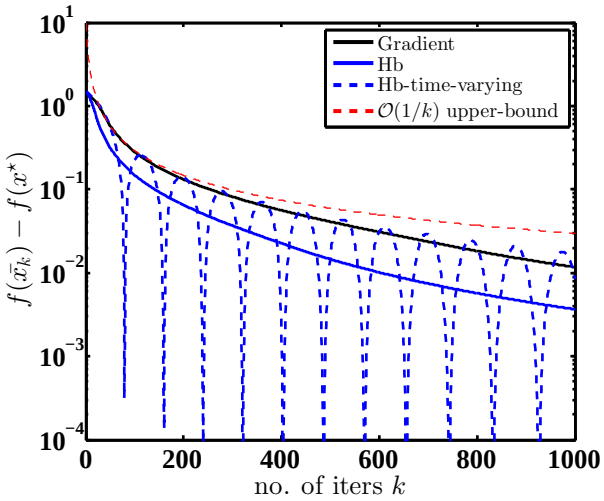

Fig. 1 Comparison of the progress of the objective values evaluated at the Cesáro average of the iterates of the gradient descent and Heavy-ball methods, and of the primal variable itself for Heavy-ball iterates with time-varying step-sizes. Included for reference is also an $\mathcal{O}(1/k)$ upper bound.

Proof. Assume that $\beta \in [0,1)$, and let

$$p_k = \frac{\beta}{1-\beta}\left(x_k - x_{k-1} + \frac{1}{L}\nabla f(x_{k-1})\right), \quad k \in \mathbb{N}_0. \tag{17}$$

Considering (4) and substituting the y-th iterates in the x-th iterates yields

$$x_{k+1} + p_{k+1} = \frac{1}{1-\beta}x_{k+1} + \frac{\beta}{1-\beta}(\frac{1}{L}\nabla f(x_k) - x_k) \overset{(4)}{=} x_k + p_k - \frac{1}{L(1-\beta)}\nabla f(x_k),$$

which implies that

$$\begin{aligned}
\|x_{k+1} + p_{k+1} - x^\star\|^2 =& \|x_k + p_k - x^\star\|^2 - \frac{2}{L(1-\beta)}\langle x_k + p_k - x^\star, \nabla f(x_k)\rangle + \frac{1}{L^2(1-\beta)^2}\|\nabla f(x_k)\|^2 \\
\overset{(17)}{=}& \|x_k + p_k - x^\star\|^2 - \frac{2}{L(1-\beta)}\langle x_k - x^\star, \nabla f(x_k)\rangle - \frac{2\beta}{L(1-\beta)^2}\langle x_k - x_{k-1}, \nabla f(x_k)\rangle \\
& - \frac{2\beta}{L^2(1-\beta)^2}\langle \nabla f(x_{k-1}), \nabla f(x_k)\rangle + \frac{1}{L^2(1-\beta)^2}\|\nabla f(x_k)\|^2 \\
\overset{(9)}{\leq}& \|x_k + p_k - x^\star\|^2 - \frac{2}{L(1-\beta)}(f(x_k) - f^\star) - \frac{1}{L^2(1-\beta)}\|\nabla f(x_k)\|^2 \\
& - \frac{2\beta}{L(1-\beta)^2}(f(x_k) - f(x_{k-1})) - \frac{\beta}{L^2(1-\beta)^2}\|\nabla f(x_k) - \nabla f(x_{k-1})\|^2 \\
& - \frac{2\beta}{L^2(1-\beta)^2}\langle \nabla f(x_{k-1}), \nabla f(x_k)\rangle + \frac{1}{L^2(1-\beta)^2}\|\nabla f(x_k)\|^2.
\end{aligned}$$

After rearrangement of terms, we thus have

$$\frac{2}{L(1-\beta)^2}(f(x_k)-f^\star)+\|x_{k+1}+p_{k+1}-x^\star\|^2 \le \frac{2\beta}{L(1-\beta)^2}(f(x_{k-1})-f^\star)+\|x_k+p_k-x^\star\|^2 - \frac{\beta}{L^2(1-\beta)^2}\|\nabla f(x_{k-1})\|^2 \tag{18}$$

Multiplying the sides of (18) in $L/2$ and summing over $k=0,\dots,T$ gives

$$\frac{1}{1-\beta}\sum_{k=0}^{T}\big(f(x_k)-f^\star\big)+\sum_{k=0}^{T}\left(\frac{\beta}{(1-\beta)^2}\big(f(x_k)-f^\star\big)+\frac{L}{2}\|x_{k+1}+p_{k+1}-x^\star\|^2\right) \le \sum_{k=0}^{T}\left(\frac{\beta}{(1-\beta)^2}\big(f(x_{k-1})-f^\star\big)+\frac{L}{2}\|x_k+p_k-x^\star\|^2\right),$$

which implies that

$$\frac{1}{1-\beta}\sum_{k=0}^{T}\big(f(x_k)-f^\star\big) \le \frac{\beta}{(1-\beta)^2}\big(f(x_0)-f^\star\big)+\frac{L}{2}\|x_0-x^\star\|^2.$$

Using the convexity inequality (11) concludes the proof. □

Recently, Allen-Zou and Orrechia [21] demonstrated that another fast gradient method due to Nesterov [22] converges with constant step-sizes for all $f\in\mathscr{F}_L^{1,1}$. That method generates iterates in the following manner

$$\begin{aligned} y_{k+1} &= x_k-\frac{1}{L}\nabla f(x_k),\\ z_{k+1} &= \arg\min_{z\in\mathbb{R}^n}\{V_x(z)+\alpha\langle\nabla f(x_k),z-z_k\rangle\},\\ x_{k+1} &= \tau z_{k+1}+(1-\tau)y_{k+1}, \end{aligned} \tag{19}$$

where $\tau\in[0,1]$, and $V_x(\cdot)$ is the Bergman divergence function [21].Similar to Theorem 3, it has been shown in [21] that the Cesáro average of the iterates generated by (19) converges to the optimum at a rate of $\mathscr{O}(1/k)$. Note that while both iterations (4) and (19) enjoy the same global rate of convergence, the two schemes are remarkably different computationally. In particular, (19) requires two gradient computations per iteration, as opposed to one gradient computation needed in (4).

4 Global analysis of Heavy-ball algorithm for the class $\mathscr{S}_{\mu,L}^{1,1}$

In this section, we focus on objective functions in the class $\mathscr{S}_{\mu,L}^{1,1}$and derive a global linear rate of convergence for the Heavy-ball algorithm. In our convergence analysis, we will use the following simple lemma on convergence of sequences.

Lemma 1. *Let $\{A_k\}_{k\ge0}$ and $\{B_k\}_{k\ge0}$ be nonnegative sequences of real numbers satisfying*

$$A_{k+1}+bB_{k+1}\le a_1A_k+a_2A_{k-1}+cB_k,\quad k\in\mathbb{N}_0 \tag{20}$$

with constants $a_1,a_2,b\in\mathbb{R}_+$ and $c\in\mathbb{R}$. Moreover, assume that

$$A_{-1}=A_0,\quad a_1+a_2<1,\quad c<b.$$

Then, the sequence $\{A_k\}_{k\geq 0}$ *generated by* (20) *satisfies*

$$A_k \leq q^k((q-a_1+1)A_0+cB_0), \tag{21}$$

where $q \in [0,1)$ *is given by*

$$q = \max\left\{\frac{c}{b}, \frac{a_1+\sqrt{a_1^2+4a_2}}{2}\right\}.$$

Proof. It is easy to check that (21) holds for $k=0$. Let $\gamma \geq 0$. From (20), we have

$$\begin{aligned} A_{t+1}+\gamma A_t + bB_{t+1} &\leq (a_1+\gamma)A_t + a_2A_{t-1} + cB_t \\ &= (a_1+\gamma)(A_t + \frac{a_2}{a_1+\gamma}A_{t-1} + \frac{c}{a_1+\gamma}B_t) \\ &\leq (a_1+\gamma)(A_t+\gamma A_{t-1}+bB_t). \end{aligned} \tag{22}$$

Note that the last inequality holds if

$$\frac{a_2}{a_1+\gamma} \leq \gamma, \quad \frac{c}{a_1+\gamma} \leq b. \tag{23}$$

The first term in (23) along with $\gamma \geq 0$ is equivalent to have $(-a_1+\sqrt{a_1^2+4a_2})/2 \leq \gamma$. Moreover, the second condition in (23) can be rewritten as $c/b - a_1 \leq \gamma$. Thus, if

$$\gamma = \max\left\{\frac{-a_1+\sqrt{a_1^2+4a_2}}{2}, \frac{c}{b}-a_1, 0\right\}, \tag{24}$$

then (23) holds. Denoting $q \triangleq a_1+\gamma < 1$, it follows from (22) that

$$A_{t+1}+\gamma A_t + bB_{t+1} \leq q(A_t+\gamma A_{t-1}+cB_t) \leq \cdots \leq q^{t+1}((1+\gamma)A_0+cB_0).$$

Since A_t and B_{t+1} are nonnegative, (21) holds. The proof is complete.

□

We are now ready for the main result in this section.

Theorem 4 *Assume that* $f \in \mathscr{S}^{1,1}_{\mu,L}$ *and that*

$$\alpha \in (0, \frac{2}{L}), \quad 0 \leq \beta < \frac{1}{2}\left(\frac{\mu\alpha}{2} + \sqrt{\frac{\mu^2\alpha^2}{4} + 4(1-\frac{\alpha L}{2})}\right). \tag{25}$$

Then, the Heavy-ball method (3) *converges linearly to a unique optimizer* $x^\star$. *In particular,*

$$f(x_k) - f^\star \leq q^k(f(x_0)-f^\star), \tag{26}$$

where $q \in [0,1)$.

Proof. For the heavy-ball iterates (3), we have

$$\|x_{k+1}-x_k\|^2=\alpha^2\|\nabla f(x_k)\|^2+\beta^2\|x_k-x_{k-1}\|^2-2\alpha\beta\langle\nabla f(x_k),x_k-x_{k-1}\rangle. \tag{27}$$

Moreover, since f belongs to $\mathscr{F}_L^{1,1}$, it follows from [14, Theorem 2.1.5] and (3) that

$$\begin{aligned} f(x_{k+1})-f^\star \leq f(x_k)-f^\star-\alpha(1-\frac{\alpha L}{2})\|\nabla f(x_k)\|^2+\frac{L\beta^2}{2}\|x_k-x_{k-1}\|^2 \\ +\beta(1-\alpha L)\langle\nabla f(x_k),x_k-x_{k-1}\rangle. \end{aligned} \tag{28}$$

Let $\theta\in(0,1)$, multiply both sides of (27) by $L\theta/(2-2\theta)$, and add the resulting identity to (28) to obtain

$$\begin{aligned} f(x_{k+1})-f^\star+\frac{L\theta}{2(1-\theta)}\|x_{k+1}-x_k\|^2 \leq f(x_k)-f^\star \\ +\alpha\left(\frac{L}{2(1-\theta)}\alpha-1\right)\|\nabla f(x_k)\|^2+\frac{L\beta^2}{2(1-\theta)}\|x_k-x_{k-1}\|^2 \\ +\beta(1-\frac{\alpha L}{1-\theta})\langle\nabla f(x_k),x_k-x_{k-1}\rangle. \end{aligned} \tag{29}$$

Assume that $(1-\theta)/L\leq\alpha<2(1-\theta)/L$. Then, since $f\in\mathscr{S}_{\mu,L}^{1,1}$, it follows from [14, Theorem 2.1.10] that

$$\begin{aligned} f(x_{k+1})-f^\star+\frac{L\theta}{2(1-\theta)}\|x_{k+1}-x_k\|^2 \leq f(x_k)-f^\star \\ +2\alpha\mu\left(\frac{L}{2(1-\theta)}\alpha-1\right)(f(x_k)-f^\star)+\frac{L\beta^2}{2(1-\theta)}\|x_k-x_{k-1}\|^2 \\ +\beta(1-\frac{\alpha L}{1-\theta})(f(x_k)-f(x_{k-1}))+\frac{\beta\mu}{2}(1-\frac{\alpha L}{1-\theta})\|x_k-x_{k-1}\|^2. \end{aligned}$$

Collecting terms yields

$$\begin{aligned} f(x_{k+1})-f^\star+\underbrace{\frac{L\theta}{2(1-\theta)}}_{b}\|x_{k+1}-x_k\|^2 \leq \underbrace{\left(1-2\alpha\mu(1-\frac{\alpha L}{2(1-\theta)})-\beta(\frac{\alpha L}{1-\theta}-1)\right)}_{a_1}(f(x_k)-f^\star) \\ +\underbrace{\beta(\frac{\alpha L}{1-\theta}-1)}_{a_2}(f(x_{k-1})-f^\star)+\underbrace{\frac{\beta}{2}\left(\mu(1-\frac{\alpha L}{1-\theta})+\frac{L\beta}{1-\theta}\right)}_{c}\|x_k-x_{k-1}\|^2, \end{aligned} \tag{30}$$

which is on the form of Lemma 1 if we identify A_k with $f(x_k)-f^\star$ and B_k with $\|x_k-x^\star\|^2$. It is easy to verify that for $\theta\in(0,1)$ and $(1-\theta)/L\leq\alpha<2(1-\theta)/L$, one has

$$b>0,\quad a_1+a_2<1.$$

Moreover, provided that

$$0\leq\beta<\frac{1}{2}\left(\frac{\mu}{L}(\alpha L+\theta-1)+\sqrt{\frac{\mu^2}{L^2}(\alpha L+\theta-1)^2+4\theta}\right),$$

it holds that $c < b$ and consequently one can apply Lemma 1 with constants a_1, a_2, b, and c to conclude the linear convergence (26). Defining $\lambda \triangleq 1-\theta$ the stability criteria reads

$$\lambda \in (0,1), \quad \frac{\lambda}{L} \leq \alpha < \frac{2\lambda}{L}, \quad 0 \leq \beta < \frac{1}{2}\left(\frac{\mu}{L}(\alpha L-\lambda)+\sqrt{\frac{\mu^2}{L^2}(\alpha L-\lambda)^2+4(1-\lambda)}\right).$$

The first two conditions can be rewritten as

$$\alpha \in (0, \frac{2}{L}), \quad \lambda \in \left(\frac{\alpha L}{2}, \min(\alpha L, 1)\right).$$

Substituting $\lambda = \alpha L/2$ into the upper stability bound on β completes the proof. □

This result extends earlier theoretical results for $\mathscr{S}^{2,1}_{L,\mu}$ to $\mathscr{S}^{1,1}_{\mu,L}$and demonstrates that the Heavy-ball method has the same rate of convergence as the gradient method and Nesterov's fast gradient method for this class of objective functions. A few comments regarding our stability criteria (25) are in order.

First, we observe that (25) guarantees stability for a wider range of parameters than the stability criteria (5) for $f \in \mathscr{F}^{1,1}_L$, and wider ranges of parameters than the stability analysis of the Heavy-ball method for non-convex cost functions presented in [18]. In particular, when α tends to $2/L$, our stability criterion allows β to be as large as μ/L, whereas the stability condition (5) requires that β tends to zero when α reaches $2/L$; see Fig. 2.

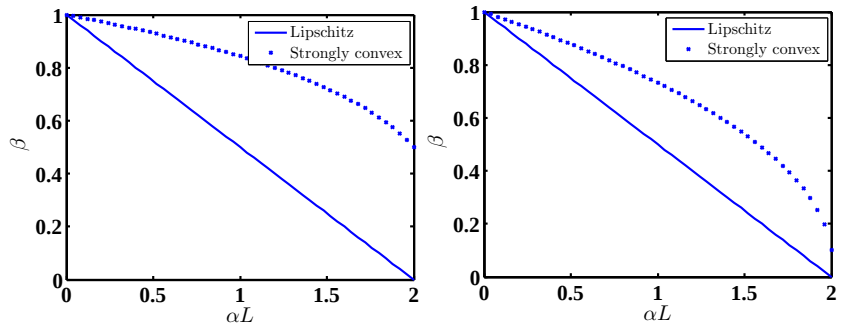

Fig. 2 The set of parameters (α, β) which guarantee convergence of the Heavy-ball algorithm (3) for objective functions $f \in \mathscr{F}^{1,1}_L$(Theorem. 1) and $f \in \mathscr{S}^{1,1}_{\mu,L}$(Theorem. 4). The left figure uses $L = 2$, $\mu = 1$ and in the right figure $L = 10$, $\mu = 1$.

Second, by comparing (25) with α and β that guarantee stability for twice differentiable strongly convex functions [12]:

$$\beta \in [0,1), \quad \alpha \in \left(0, \frac{2(1+\beta)}{L}\right), \tag{31}$$

our stability criteria may appear restrictive at first. However, motivated by [19], we consider a counter example where the original stability criteria (31) for twice differentiable strongly convex functions do not hold for the class $\mathscr{S}^{1,1}_{\mu,L}$. In particular, let us consider

$$\nabla f(x) = \begin{cases} 50x+45 & x < -1, \\ 5x & -1 \leq x < 0, \\ 50x & 0 \leq x. \end{cases} \tag{32}$$

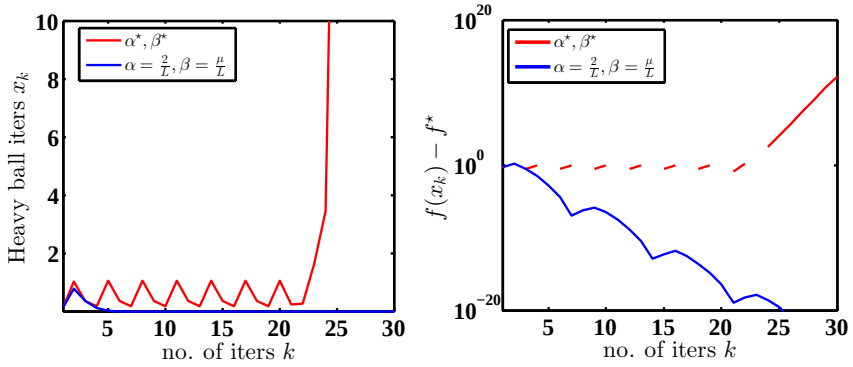

Fig. 3 Heavy-ball iterates with optimal step-sizes for $f \in \mathscr{S}^{2,1}_{\mu,L}$ do not converge for the example in (32). However, parameters that satisfy our new global stability criteria ensure convergence of the iterates.

It is easy to check that ∇f is continuous and $f \in \mathscr{S}^{1,1}_{\mu,L}$ with $\mu = 5$ and $L = 50$. According to our numerical tests, for initial conditions in the interval $x_0 < -0.8$ or $x_0 > 0.15$, the Heavy-ball method with parameters $\alpha^\star = 4/(\sqrt{L}+\sqrt{\mu})^2$ and $\beta^\star = (\sqrt{L}-\sqrt{\mu})^2/(\sqrt{L}+\sqrt{\mu})^2$ (the optimal step-sizes for the class $\mathscr{S}^{2,1}_{L,\mu}$ in [12]) produces non-converging sequences. However, Fig. 3 shows that using the maximum value of α permitted by our global analysis results in iterates that converge to the optimum.

Finally, note that Lemma 1 also provides an estimate of the convergence factor of the iterates. In particular, after a few simplifications one can find that when

$$\alpha \in (0, \frac{1}{L}], \quad \beta = \sqrt{(1-\alpha\mu)(1-\alpha L)},$$

and $\theta = 1 - \alpha L$ in (30), the convergence factor of the Heavy-ball method (3) is given by $q = 1 - \alpha\mu$. Note that this factor coincides with the best known convergence factor for the gradient method on $\mathscr{S}^{1,1}_{\mu,L}$ [12, Theorem 2, Chapter 1]. However, supported by the numerical simulations we envisage that the convergence factor could be strengthened even further. This is indeed left as a future work.

5 CONCLUSIONS

Global stability of the Heavy-ball method has been established for two important classes of convex optimization problems. Specifically, we have shown that when the objective function is convex and has a Lipschitz-continuous gradient, then the Cesáro-averages of the iterates converge to the optimum at a rate no slower than $\mathscr{O}(1/k)$, where k is the number of iterations. When the objective function is also strongly convex, we established that the Heavy-ball iterates converge linearly to the unique optimum.

In our future work, we hope to extend the present results to the constrained optimization problems and derive sharper bounds on the guaranteed convergence factor when $f \in \mathscr{S}^{1,1}_{\mu,L}$.

References

1. M. R. Hestenes and E. Stiefel, *Methods of conjugate gradients for solving linear systems*. National Bureau of Standards Washington, DC, 1952.

2. M. Frank and P. Wolfe, "An algorithm for quadratic programming," *Naval research logistics quarterly*, vol. 3, pp. 95–110, 1956.
3. K. J. Arrow, *Studies in Linear and Non-linear Programming*. Stanford mathematical studies in the social sciences, 1958.
4. O. Devolder, F. Glineur, and Y. Nesterov, "First-order methods of smooth convex optimization with inexact oracle," *Mathematical Programming*, pp. 1–39, 2013.
5. C. Guzman and A. Nemirovski, "On lower complexity bounds for large-scale smooth convex optimization," *submitted to Journal of Complexity*, 2014.
6. Y. Drori and M. Teboulle, "Performance of first-order methods for smooth convex minimization: a novel approach," *Mathematical Programming, Series A*, vol. 145, pp. 451–482, 2014.
7. A. Beck and M. Teboulle, "A fast iterative shrinkage-thresholding algorithm for linear inverse problems," *SIAM Journal on Imaging Sciences*, vol. 2, no. 1, pp. 183–202, 2009.
8. Q. Lin, Z. Lu, and L. Xiao, "An Accelerated Proximal Coordinate Gradient Method and its Application to Regularized Empirical Risk Minimization," *ArXiv e-prints*, Jul. 2014.
9. I. Necoara and V. Nedelcu, "Rate analysis of inexact dual first-order methods application to dual decomposition," *Automatic Control, IEEE Transactions on*, vol. 59, no. 5, pp. 1232–1243, May 2014.
10. Y. Nesterov, "A method of solving a convex programming problem with convergence rate $O(1/k^2)$," *Soviet Mathematics Doklady*, vol. 27, pp. 372–376, 1983.
11. B. T. Polyak, "Some methods of speeding up the convergence of iteration methods," *USSR Computational Mathematics and Mathematical Physics*, vol. 4, no. 5, pp. 1–17, 1964.
12. ——, *Introduction to Optimization*. Optimization Software, 1987.
13. A. Nemirovsky and D. Yudin, *Problem complexity and method efficiency in optimization Problem*, ser. Interscience Series in Discrete Mathematics. John Wiley, 1983.
14. Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer, 2004.
15. E. Ghadimi, I. Shames, and M. Johansson, "Multi-step gradient methods for networked optimization," *IEEE Transactions on Signal Processing*, vol. 61, no. 21, pp. 5417–5429, Nov 2013.
16. P. Ochs, T. Brox, and T. Pock, "iPiasco: Inertial proximal algorithm for strongly convex optimization," *Technical Report*, 2014. [Online]. Available: http://lmb.informatik.uni-freiburg.de//Publications/2014/OB14a
17. H. Wang and P. Miller, "Scaled Heavy-Ball acceleration of the Richardson-Lucy algorithm for 3D microscopy image restoration," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 848–854, Feb 2014.
18. S. Zavriev and F. Kostyuk, "Heavy-ball method in nonconvex optimization problems," *Computational Mathematics and Modeling*, vol. 4, no. 4, pp. 336–341, 1993.
19. L. Lessard, B. Recht, and A. Packard, "Analysis and design of optimization algorithms via integral quadratic constraints," *eprint arXiv:1408.3595*, 2014.
20. R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer, 1998, vol. 317.
21. Z. Allen-Zhu and L. Orecchia, "Linear Coupling of Gradient and Mirror Descent: A Novel, Simple Interpretation of Nesterov's Accelerated Method," *ArXiv e-prints*, Jul. 2014.
22. Y. Nesterov, "Gradient methods for minimizing composite functions," *Mathematical Programming*, vol. 140, no. 1, pp. 125–161, 2013.