2015 European Control Conference (ECC)
July 15-17, 2015. Linz, Austria

Global convergence of the Heavy-ball method for convex optimization

Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson

Abstract— **This paper establishes global convergence and provides global bounds of the rate of convergence for the Heavy-ball method for convex optimization. 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. Numerical examples validate our theoretical findings.**

I. Introduction

First-order convex optimization methods have a rich history dating back to 1950's [1]–[3]. Their mild computational burden makes them suitable for large scale applications. Recently, these techniques have attracted significant interest, both in terms of new theory [4]–[6] and in terms of applications in areas such as signal processing [7], machine learning [8] and control [9]. Arguably, this development has 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 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 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 $\mathcal{O}(1/k^2)$ for convex optimization problems with Lipschitz-continuous gradient. This created a gap between the convergence rate of the gradient method and what could potentially be achieved. This gap was closed by Nesterov, who presented a 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 if this is the best that can be achieved.

When, the objective function is twice continuously differentiable, strongly convex and has Lipschitz continuous gradient, the Heavy-ball method by Polyak [11] has local 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 global rate of convergence of the Heavy-ball method for convex problems. 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.

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. 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 II reviews first-order convex optimization algorithms. Global convergence proofs for the Heavy-ball method are presented in Section III for objective functions with Lipschitz continuous gradient and in Section IV for objective functions that are also strongly convex. Numerical studies are performed in Section V and concluding remarks are given in Section VI.

A. Notation

We let $\mathbb{R}$, $\mathbb{R}_+$, $\mathbb{N}$, and $\mathbb{N}_0$ denote the set of real numbers, real positive numbers, the set of natural numbers, and the

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

Authors are with the Department of Automatic Control, School of Electrical Engineering and ACCESS Linnaeus Center, Royal Institute of Technology - KTH, Stockholm, Sweden. Emails:{euhanna, hamidrez, mikaelj}@kth.se.

set of natural numbers including zero, respectively. The Euclidean norm is denoted by $\|\cdot\|$. We use I_n to denote the $n \times n$ identity matrix.

II. Background

We consider 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 $\mathcal{F}_L^{1,1}$, 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.*,

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

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

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 $f \in \mathcal{F}_L^{1,1}$, then $f(x_k) - f^\star$ associated to the sequence $\{x_k\}$ in (2) converges at the rate $\mathcal{O}(1/k)$. On the other hand, if $f \in \mathcal{S}_{\mu,L}^{1,1}$, 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 $f \in \mathcal{S}_{\mu,L}^{1,1}$ 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 derived 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.

Another first-order method that achieves faster convergence than the gradient method is Nesterov's fast gradient method [14]. In it's simplest form, Nesterov's algorithm with constant step-sizes takes the form

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

where $\beta > 0$. When $f \in \mathcal{S}_{\mu,L}^{1,1}$, the iterates produced by (4) with $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ converge globally at the linear rate towards the optimal point with a convergence factor $1 - \sqrt{\mu/L}$. This factor is smaller than that of the gradient, but larger than the local convergence factor of the Heavy-ball method. Later, in Section V, we compare three aforementioned algorithms for a test problem belonging to $\mathcal{S}_{\mu,L}^{1,1}$. For the class $\mathcal{F}_L^{1,1}$, however, to the best of our knowledge no convergence result has been established for Nesterov's algorithm with constant step-sizes.

III. Global analysis of Heavy-ball algorithm for the class $\mathcal{F}_L^{1,1}$

In this section, we consider the Heavy-ball iterates (3) for the class $\mathcal{F}_L^{1,1}$ and establish a new rate of convergence for it.

Theorem 1: Assume that $f \in \mathcal{F}_L^{1,1}$ and that

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

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

$$f(\overline{x}_T) - f^\star \leq \frac{1}{T+1}\Big(\frac{\beta}{1-\beta}(f(x_0) - f^\star) + \frac{1-\beta}{2\alpha}\|x_0 - x^\star\|^2\Big), \tag{5}$$

where $\overline{x}_T$ is the time averaged vector:

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

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

$$p_k = \frac{\beta}{1-\beta}(x_k - x_{k-1}), \quad k \in \mathbb{N}_0. \tag{6}$$

We have

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

which implies that

$$\begin{aligned}\|x_{k+1}+p_{k+1}-x^\star\|^2 =&\|x_k+p_k-x^\star\|^2\\ &-\frac{2\alpha}{1-\beta}\langle x_k+p_k-x^\star,\nabla f(x_k)\rangle\\ &+\left(\frac{\alpha}{1-\beta}\right)^2\|\nabla f(x_k)\|^2\\ \overset{(6)}{=}&\|x_k+p_k-x^\star\|^2\\ &-\frac{2\alpha}{1-\beta}\langle x_k-x^\star,\nabla f(x_k)\rangle\\ &-\frac{2\alpha\beta}{(1-\beta)^2}\langle x_k-x_{k-1},\nabla f(x_k)\rangle\\ &+\left(\frac{\alpha}{1-\beta}\right)^2\|\nabla f(x_k)\|^2. \qquad (7)\end{aligned}$$

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

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

Substituting the above inequalities into (7), after a few manipulations, it yields

$$\begin{aligned}&\frac{2\alpha}{(1-\beta)^2}\big(f(x_k)-f^\star\big)+\|x_{k+1}+p_{k+1}-x^\star\|^2\\ &\quad\le\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_{k-1})-f^\star\big)+\|x_k+p_k-x^\star\|^2\\ &\quad+\left(\frac{\alpha}{1-\beta}\right)\left(\frac{\alpha}{1-\beta}-\frac{1}{L}\right)\|\nabla f(x_k)\|^2. \qquad (8)\end{aligned}$$

Note that when $\alpha\in(0,(1-\beta)/L]$, the rightmost term of (8) becomes non-positive and, therefore, can be eliminated from the right-hand-side. Summing (8) over $k=0,\dots,T$ gives

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

which implies that

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

Note that as f is convex, we have

$$(T+1)f(\overline{x}_T)\le\sum_{k=0}^{T}f(x_k).$$

It now follows that

$$f(\overline{x}_T)-f^\star\le\frac{1}{T+1}\big(\frac{\beta}{1-\beta}(f(x_0)-f^\star)+\frac{1-\beta}{2\alpha}\|x_0-x^\star\|^2\big).$$

This completes the proof. ■

Based on this result, the following comments 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

$$\begin{aligned}\min_{0\le k\le T}f(x_k)-f^\star\le\frac{1}{T+1}\bigg(&\frac{\beta}{1-\beta}(f(x_0)-f^\star)\\ &+\frac{1-\beta}{2\alpha}\|x_0-x^\star\|^2\bigg),\end{aligned}$$

for all $T\in\mathbb{N}_0$. Second, according to [14, Lemma 1.2.3], one can obtain $f(x_0)-f^\star\le(L/2)\|x_0-x^\star\|^2$, which in combination to (5) yields the best Heavy-ball guaranteed upper bound

$$f(\overline{x}_T)-f^\star\le\frac{L}{2(T+1)}\frac{\|x_0-x^\star\|^2}{1-\beta}, \qquad (9)$$

for $\alpha=(1-\beta)/L$. Note also that by setting $\beta=0$ in the preceding upper bound, we can find the smallest convergence factor that coincides with the best convergence factor of the gradient descent method reported in [7].

Later, in section V, we numerically compare the performance of the Heavy-ball method to the gradient descent algorithm for a test problem.

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

In this section, we restrict our attention to the class $\mathcal{S}_{\mu,L}^{1,1}$and derive a global linear rate of convergence for the Heavy-ball algorithm.

Theorem 2: Assume that $f\in\mathcal{S}_{\mu,L}^{1,1}$and that

$$\beta\in[0,1),\quad \alpha\in\left(0,\frac{2(1-\beta)}{L+\mu}\right). \qquad (10)$$

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

$$\|x_k-x^\star\|^2\le cq^k,\quad k\in\mathbb{N}_0,$$

where $q\in(0,1)$ and $c>0$.

Proof: Since $f\in\mathcal{S}_{\mu,L}^{1,1}$, it follows from [14, Section 2.1.3] that

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

These inequalities together with (7) imply that

$$\begin{aligned}&\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_k)-f^\star\big)+\|x_{k+1}+p_{k+1}-x^\star\|^2\\ &\le\frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_{k-1})-f^\star\big)+\|x_k+p_k-x^\star\|^2\\ &\quad-\frac{2\alpha\mu L}{(1-\beta)(L+\mu)}\|x_k-x^\star\|^2-\frac{\alpha\beta\mu}{(1-\beta)^2}\|x_k-x_{k-1}\|^2\\ &\quad+\left(\frac{\alpha}{1-\beta}\right)\left(\frac{\alpha}{1-\beta}-\frac{2}{L+\mu}\right)\|\nabla f(x_k)\|^2, \qquad (11)\end{aligned}$$

where p_k, $k \in \mathbb{N}_0$, is defined in (6). According to [20, Theorem 2], as $f \in \mathcal{S}^{1,1}_{\mu,L}$, we have

$$2\mu\big(f(x_k) - f^\star\big) \leq \big\|\nabla f(x_k)\big\|^2. \tag{12}$$

Note also that when

$$\beta \in [0,1), \quad \alpha \in \left(0, \frac{2(1-\beta)}{L+\mu}\right), \tag{13}$$

the last term of (11) becomes negative. It now follows from (11) and (12) that if (13) holds, then

$$\begin{aligned}
&\left(\frac{2\alpha}{1-\beta}\right)\left(\frac{\beta-\alpha\mu}{1-\beta} + \frac{2\mu}{L+\mu}\right)\big(f(x_k) - f^\star\big) \\
&\quad + \|x_{k+1} + p_{k+1} - x^\star\|^2 \\
&\leq \frac{2\alpha\beta}{(1-\beta)^2}\big(f(x_{k-1}) - f^\star\big) + \|x_k + p_k - x^\star\|^2 \\
&\quad - \frac{2\alpha\mu L}{(1-\beta)(L+\mu)}\|x_k - x^\star\|^2 - \frac{\alpha\beta\mu}{(1-\beta)^2}\|x_k - x_{k-1}\|^2.
\end{aligned} \tag{14}$$

From (6), it is easy to show that $\|x_k + p_k - x^\star\|^2$ can be written as $z_k^\top M z_k$, where

$$z_k = [x_k - x^\star, x_k - x_{k-1}]^\top,$$

and

$$M = \begin{bmatrix} I_n & \dfrac{\beta}{1-\beta} I_n \\ \dfrac{\beta}{1-\beta} I_n & \left(\dfrac{\beta}{1-\beta}\right)^2 I_n \end{bmatrix}.$$

Thus, (14) can be written in the matrix form as

$$m\big(f(x_k) - f^\star\big) + z_{k+1}^\top M z_{k+1} \leq n\big(f(x_{k-1}) - f^\star\big) + z_k^\top N z_k, \tag{15}$$

where

$$\begin{aligned}
m &= \frac{2\alpha}{1-\beta}\left(\frac{\beta-\alpha\mu}{1-\beta} + \frac{2\mu}{L+\mu}\right), \\
n &= \frac{2\alpha\beta}{(1-\beta)^2},
\end{aligned}$$

and

$$N = \begin{bmatrix} \big(1 - \dfrac{2\alpha\mu L}{(1-\beta)(L+\mu)}\big) I_n & \dfrac{\beta}{1-\beta} I_n \\ \dfrac{\beta}{1-\beta} I_n & \dfrac{\beta(\beta-\alpha\mu)}{(1-\beta)^2} I_n \end{bmatrix}.$$

One can verify that for any α and β satisfying (13), there exist constants $q_1, q_2 \in (0,1)$ such that $n \leq q_1 m$ and that $q_2 M - N$ is a positive semidefinite matrix, see Appendix. Let $q = \max\{q_1, q_2\}$. Then, from (15), we obtain

$$f(x_k) - f^\star + z_{k+1}^\top M z_{k+1} \leq q\big(f(x_{k-1}) - f^\star + z_k^\top M z_k\big),$$

for all $k \in \mathbb{N}_0$. Since $q \in (0,1)$ and M is a positive semidefinite matrix, it follows that

$$f(x_k) - f^\star \leq q^k\big(f(x_0) - f^\star + \|x_0 - x^\star\|^2\big),$$

which by strong convexity of f implies that

$$\frac{\mu}{2}\|x_k - x^\star\|^2 \leq q^k\big(f(x_0) - f^\star + \|x_0 - x^\star\|^2\big),$$

for all $k \in \mathbb{N}_0$. The proof is complete. ∎

This result closes the analysis gap between the $\mathcal{S}^{1,1}_{\mu,L}$and $\mathcal{S}^{2,1}_{L,\mu}$ classes for the Heavy-ball method and shows that the method enjoys a similar rate of convergence as the one for the standard gradient and the Nesterov's algorithms. In section V, we compare the performance of the Heavy-ball method with two other methods.

By comparing (10) with the following α and β:

$$\beta \in [0,1), \quad \alpha \in \left(0, \frac{2(1+\beta)}{L}\right), \tag{16}$$

which guarantee the local convergence of Heavy-ball method for twice differentiable strongly convex functions [12], our stability criteria, at first look, may appear restrictive. However, in Section V, we will present an example in $\mathcal{S}^{1,1}_{\mu,L}$where the corresponding Heavy-ball iterates with parameters satisfying (16) does not converge to the optimum.

V. Numerical examples

In this section, we give numerical examples to validate our theoretical results.

The first example, considers a counter example presented in [19] where the local stability criteria (16) do not hold globally for the class $\mathcal{S}^{1,1}_{\mu,L}$. In particular, consider

$$\nabla f(x) = \begin{cases} 16x + 45 & x < -3, \\ x & -3 \leq x < 0, \\ 16x & x \geq 0. \end{cases} \tag{17}$$

It is easy to check that ∇f is continuous and $f \in \mathcal{S}^{1,1}_{\mu,L}$with $\mu = 1$ and $L = 16$. As shown in [19], selecting an initial condition in the interval $x_0 \in [1.9, 2.4]$, the Heavy-ball method with the parameters $\alpha^\star = 4/(\sqrt{L} + \sqrt{\mu})^2$ and $\beta = \sqrt{\beta^\star} = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ ($\alpha^\star$ and $\beta^\star$ are reported to be the optimal choices for the class $\mathcal{S}^{2,1}_{L,\mu}$in [12]) produces a limit cycle with oscillations that never die out. However, for the same β, Fig. 1 shows that our global stability result offers a maximum allowable value of $\alpha < 2(1-\beta)/(L+\mu)$ which stabilizes the Heavy-ball iterates.

As the second experiment, we compare the performance of the gradient and Heavy-ball algorithms for the class $\mathcal{F}^{1,1}_L$. The objective function is the Moreau proximal envelope of the function $f(x) = (1/c)\|x\|$. That is,

$$f(x) = \begin{cases} \dfrac{1}{c}\|x\| - \dfrac{1}{2c^2} & \|x\| \geq \dfrac{1}{c}, \\ \dfrac{1}{2}\|x\|^2 & \|x\| \leq \dfrac{1}{c}, \end{cases} \tag{18}$$

where $c \in \mathbb{R}_+$. Note that (18) is convex and continuously differentiable with Lipschitz constant $L = 1$ [21]. First-order methods designed to find the minimum of this cost function are expected to pertain very poor convergence behavior [6]. We picked $x \in \mathbb{R}^{50}$ and $c = 5$ to conduct the numerical study. Fig. 2 shows the progress of the objective values evaluated at the mean of the primal variables towards the

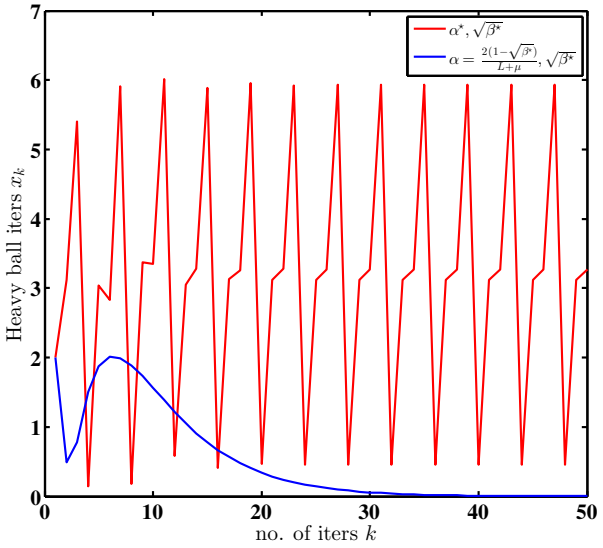

Fig. 1. Heavy-ball iterates with the original step-size parameters does not converge for the example in (17). However, our new global stability bounds stabilize the Heavy-ball iterates.

optimal solution. For the Heavy-ball algorithm we picked $\beta = 0.8$ and $\alpha = 0.2/L$ whereas the gradient algorithm is implemented with the step-size $\alpha = 1/L$. The plot shows that the gradient method outperforms the Heavy-ball algorithm. This observation agrees with the analytical bound (9) and the discussion afterwards.

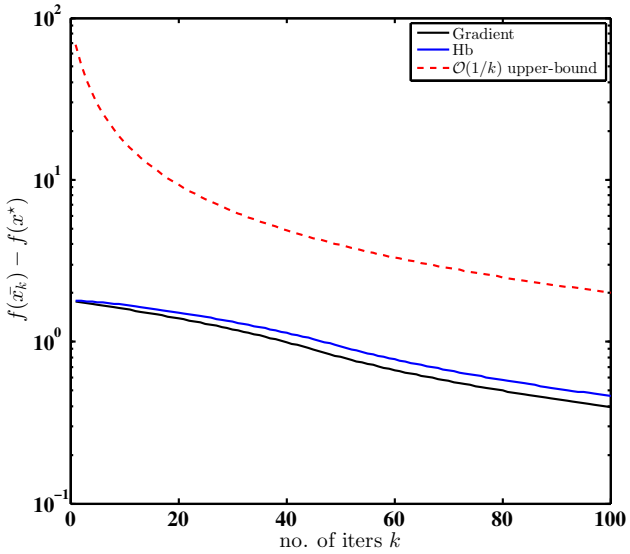

Fig. 2. Comparison of the progress of the objective values evaluated at the mean of primal variable for the gradient descent and Heavy-ball methods.

Our last experiment is devoted to the class $\mathcal{S}^{1,1}_{\mu,L}$. The objective function is the regularized logistic regression

$$f(x) = \log(1 + \exp(-a^\top x)) + (1/2)a^\top x + (\mu/2)\|x\|^2,$$

with random parameters $a \in \mathbb{R}^{100}$ and $\mu \in \mathbb{R}_+$. The problem is widely used in machine learning applications [8]. Fig. 3 compares the per-iterate progress of the gradient descent, Nesterov's and Heavy-ball algorithms. For the gradient descent method, we set the optimal step-size $\alpha = 2/(L+\mu)$; for the Nesterov's algorithm we set $\alpha = 1/L$ and $\beta = (\sqrt{L} - \sqrt{\mu})/(\sqrt{L} + \sqrt{\mu})$ according to [14], and finally for the Heavy-ball method we picked $\beta = 0.4$ and $\alpha = 2(1-\beta)/(L+\mu)$. Our extensive simulations indicate that it is possible to tune the heavy-ball algorithm so that it outperforms the gradient method in this class. Finding the Heavy-ball optimal step-sizes and the corresponding convergence factor for the $\mathcal{S}^{1,1}_{\mu,L}$class is an attractive future direction.

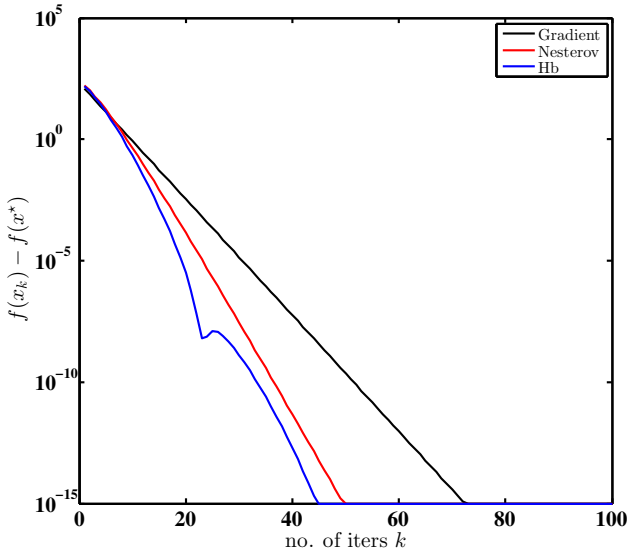

Fig. 3. Comparison of the progress of the objective values for the gradient descent, Nesterov's and Heavy-ball methods. For this particular example, both Nesterov's and Heavy-ball algorithms out-perform the gradient method.

VI. 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 $\mathcal{O}(1/k)$, 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. Numerical examples confirmed our theoretical findings.

In our future work, we hope to better understand the properties of the iterates themselves when $f \in \mathcal{F}^{1,1}_L$, and to derive sharper bounds on the guaranteed convergence factor when $f \in \mathcal{S}^{1,1}_{\mu,L}$.

APPENDIX

We will show that if (13) holds, then there are constants $q_1, q_2 \in (0,1)$ such that $n \leq q_1 m$ and that $q_2 M - N$ is positive semidefinite, where m, n, M and N are defined in the proof of Theorem 2.

Let $\widehat{q} = n/m$. We have

$$\widehat{q} = \frac{\beta}{\beta + \mu \underbrace{\left(\frac{2(1-\beta)}{L+\mu} - \alpha\right)}_{\overset{(13)}{>} 0}},$$

which shows that $\widehat{q} \in (0,1)$. Therefore, $n \leq q_1 m$ holds for any $q_1 \in [\widehat{q}, 1)$.

Now, define $S_s = sM - N$ with $s > 0$. According to [22, Theorem 7.7.6], the symmetric matrix S_s, given by

$$\begin{bmatrix} \left(s - 1 + \frac{2\alpha\mu L}{(1-\beta)(L+\mu)}\right) I_n & \frac{(s-1)\beta}{1-\beta} I_n \\ \frac{(s-1)\beta}{1-\beta} I_n & \frac{(s-1)\beta^2 + \alpha\beta\mu}{(1-\beta)^2} I_n \end{bmatrix},$$

is positive semidefinite if and only if

$$s - 1 + \frac{2\alpha\mu L}{(1-\beta)(L+\mu)} \geq 0,$$

and

$$\left(s - 1 + \frac{2\alpha\mu L}{(1-\beta)(L+\mu)}\right)\left(\frac{(s-1)\beta^2 + \alpha\beta\mu}{(1-\beta)^2}\right) \geq \left(\frac{(s-1)\beta}{1-\beta}\right)^2.$$

It is easy to show that these inequalities hold if and only if

$$\begin{cases} s & \geq 1 - \dfrac{2\alpha\mu L}{(1-\beta)(L+\mu)}, \\ s & \geq 1 - \dfrac{2\alpha\mu L}{(1-\beta)(L+\mu) + 2L\beta}. \end{cases} \tag{19}$$

Let

$$\widetilde{q} = \max\left\{0, 1 - \frac{2\alpha\mu L}{(1-\beta)(L+\mu) + 2L\beta}\right\}.$$

Since $\widetilde{q} \in [0,1)$, it follows from (19) that for any $q_2 \in [\widetilde{q}, 1)$, $S_{q_2} = q_2 M - N$ is positive semidefinite.

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] Y. Nesterov, "Efficiency of coordinate descent methods on huge-scale optimization problems." *SIAM Journal on Optimization*, vol. 22, no. 2, pp. 341–362, 2012.

[21] R. T. Rockafellar and R. J.-B. Wets, *Variational analysis*. Springer, 1998, vol. 317.

[22] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge University Press, 1999.