КОМПЬЮТЕРНЫЕ ИССЛЕДОВАНИЯ
И МОДЕЛИРОВАНИЕ 2022 Т. 14 № 2 С. 277–308
DOI: 10.20537/2076-7633-2022-14-2-277-308

Ки&М

МАТЕМАТИЧЕСКИЕ ОСНОВЫ И ЧИСЛЕННЫЕ МЕТОДЫ МОДЕЛИРОВАНИЯ

УДК: 519.6

Метод тяжелого шарика с усреднением

М. Ю. Данилова[1,2,a], Г. С. Малиновский[3,b]

[1]Институт проблем управления им. В. А. Трапезникова Российской академии наук,
Россия, 117997, г. Москва, ул. Профсоюзная, д. 65
[2]Московский физико-технический институт (национальный исследовательский университет),
Россия, 117303, г. Москва, ул. Керченская, д. 1а, корп. 1
[3]Научно-технологический университет имени короля Абдаллы,
Королевство Саудовская Аравия, 23955-6900, Тувал

E-mail: [a] danilovamarina15@gmail.com, [b] grigorii.malinovskii@kaust.edu.sa

Получено 18.11.2021.
Принято к публикации 13.02.2022.

Методы оптимизации первого порядка являются важным рабочим инструментов для широкого спектра современных приложений в разных областях, среди которых можно выделить экономику, физику, биологию, машинное обучение и управление. Среди методов первого порядка особого внимания заслуживают ускоренные (моментные) методы в силу их практической эффективности. Метод тяжелого шарика (heavy-ball method — HB) — один из первых ускоренных методов. Данный метод был разработан в 1964 г., и для него был проведен анализ сходимости для квадратичных сильно выпуклых функций. С тех пор были предложены и проанализированы разные варианты HB. В частности, HB известен своей простотой реализации и эффективностью при решении невыпуклых задач. Однако, как и другие моментные методы, он имеет немонотонное поведение; более того, при сходимости HB с оптимальными параметрами наблюдается нежелательное явление, называемое пик-эффектом. Чтобы решить эту проблему, в этой статье мы рассматриваем усредненную версию метода тяжелого шарика (averaged heavy-ball method — AHB). Мы показываем, что для квадратичных задач AHB имеет меньшее максимальное отклонение от решения, чем HB. Кроме того, для общих выпуклых и сильно выпуклых функций доказаны неускоренные скорости глобальной сходимости AHB, его версии WAHB со взвешенным усреднением, а также для AHB с рестартами R-AHB. Насколько нам известно, такие гарантии для HB с усреднением не были явно доказаны для сильно выпуклых задач в существующих работах. Наконец, мы проводим несколько численных экспериментов для минимизации квадратичных и неквадратичных функций, чтобы продемонстрировать преимущества использования усреднения для HB. Кроме того, мы также протестировали еще одну модификацию AHB, называемую методом tail-averaged heavy-ball (TAHB). В экспериментах мы наблюдали, что HB с правильно настроенной схемой усреднения сходится быстрее, чем HB без усреднения, и имеет меньшие осцилляции.

Ключевые слова: методы первого порядка, выпуклая оптимизация, ускоренные градиентные методы, глобальная сходимость

Исследование выполнено при финансовой поддержке РФФИ (теоремы 2 и 3 в рамках научного проекта № 20-31-90073) и РНФ (теоремы 5 и 4 в рамках научного проекта № 21-71-30005)

COMPUTER RESEARCH AND MODELING
2022 VOL. 14 NO. 2 P. 277–308
DOI: 10.20537/2076-7633-2022-14-2-277-308

Ки&М

MATHEMATICAL MODELING AND NUMERICAL SIMULATION

UDC: 519.6

Averaged heavy-ball method

M. Yu. Danilova[1,2,a], G. S. Malinovsky[3,b]

[1]V. A. Trapeznikov Institute of Control Sciences of Russian Academy of Sciences,
65 Profsoyuznaya st., Moscow, 117997, Russia
[2]Moscow Institute of Physics and Technology (National Research University),
1a Kerchenskaya st., Moscow, 117303, Russia
[3]King Abdullah University of Science and Technology,
Thuwal 23955-6900, Kingdom of Saudi Arabia

E-mail: [a] danilovamarina15@gmail.com, [b] grigorii.malinovskii@kaust.edu.sa

Received 18.11.2021.
Accepted for publication 13.02.2022.

First-order optimization methods are workhorses in a wide range of modern applications in economics, physics, biology, machine learning, control, and other fields. Among other first-order methods accelerated and momentum ones obtain special attention because of their practical efficiency. The heavy-ball method (HB) is one of the first momentum methods. The method was proposed in 1964 and the first analysis was conducted for quadratic strongly convex functions. Since then a number of variations of HB have been proposed and analyzed. In particular, HB is known for its simplicity in implementation and its performance on nonconvex problems. However, as other momentum methods, it has nonmonotone behavior, and for optimal parameters, the method suffers from the so-called *peak effect*. To address this issue, in this paper, we consider an averaged version of the heavy-ball method (AHB). We show that for quadratic problems AHB has a smaller maximal deviation from the solution than HB. Moreover, for general convex and strongly convex functions, we prove non-accelerated rates of global convergence of AHB, its weighted version WAHB, and for AHB with restarts R-AHB. To the best of our knowledge, such guarantees for HB with averaging were not explicitly proven for strongly convex problems in the existing works. Finally, we conduct several numerical experiments on minimizing quadratic and nonquadratic functions to demonstrate the advantages of using averaging for HB. Moreover, we also tested one more modification of AHB called the tail-averaged heavy-ball method (TAHB). In the experiments, we observed that HB with a properly adjusted averaging scheme converges faster than HB without averaging and has smaller oscillations.

Keywords: first-order methods, convex optimization, momentum methods, global convergence guarantees

Citation: *Computer Research and Modeling*, 2022, vol. 14, no. 2, pp. 277–308.

This work was supported by Russian Foundation for Basic Research (Theorems 2 and 3, project No. 20-31-90073) and by Russian Science Foundation (Theorems 5 and 4, project No. 21-71-30005)

Introduction

First-order optimization methods have good convergence guarantees and are simple to implement. Therefore, they are widely used in various applications. In particular, accelerated or first-order momentum methods such as Nesterov's method [Nesterov, 1983] and Heavy-Ball method [Polyak, 1964] and their various extensions are prevalent in some practically essential tasks, e. g., training of deep neural networks.

Due to its efficiency in solving nonconvex optimization problems [Danilova et al., 2020], heavy-ball method has gained significant attention in recent years. As a result, a number of its modifications were proposed, including stochastic [Yang, Lin, Li, 2016; Taylor, Bach, 2019; Defazio, 2020], zeroth-order [Gorbunov et al., 2019], and distributed variants [Yu, Jin, Yang, 2019; Mishchenko et al., 2019], to mention a few.

However, even for simple (strongly) convex problems, accelerated/momentum methods have nonmonotone behavior. For example, in the recent paper [Danilova, Kulakova, Polyak, 2018], the authors show that the heavy-ball method (HB) with optimal parameters has the so-called *peak-effect* even for simple quadratic minimization problems. This means that in this case the distance to the solution during the initial iterations of HB. Moreover, the maximal distance is proportional to $\sqrt{\varkappa}$ [Danilova, Kulakova, Polyak, 2018; Mohammadi, Samuelson, Jovanović, 2021], where $\varkappa$ is the condition number of the problem. Therefore, for ill-conditioned problems ($\varkappa \gg 1$) peak-effect can be significant.

Contributions

To address this issue, in this work, we consider an averaged version of the Heavy-Ball method called Averaged heavy-ball method (AHB). We study the maximal deviation of this method for quadratic functions and prove the global convergence guarantees in the convex and strongly convex (not necessarily quadratic) cases for AHB and its version based on the weighted averaging (WAHB). For quadratic functions with a specific property of the spectrum, our theoretical results show that there exists a choice of parameters for AHB such that the momentum parameter β is sufficiently large but the maximal deviation is significantly smaller than for HB with optimal parameters. We derive global complexity results for AHB and WAHB matching the best-known ones for HB. To the best of our knowledge, we prove the first global convergence results for HB with averaging in the strongly convex case (see the summary in Table 1). Moreover, our numerical experiments corroborate our theoretical observations and show that HB with a properly adjusted averaging scheme converges faster than HB without averaging and has smaller oscillations.

Preliminaries

We focus on the following minimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \tag{1}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a L-smooth and μ-strongly convex function.

Definition 1 (L-smoothness). A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called L-smooth for some constant $L > 0$ if its gradient is L-Lipschitz, i. e., for all $x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leqslant L\|x - y\|_2. \tag{2}$$

Definition 2 (μ-strong convexity). A differentiable function $f\colon \mathbb{R}^n \to \mathbb{R}$ is called μ-strongly convex for some constant $\mu \geqslant 0$ if for all $x, y \in \mathbb{R}^n$ the following inequality holds:

$$f(y) \geqslant f(x) + \langle \nabla f(x), y - x\rangle + \frac{\mu}{2}\|y - x\|_2^2. \tag{3}$$

Throughout the paper we use standard notation for the optimization literature [Polyak, 1987; Nesterov, 2018], e. g., x^* denotes the solution of (1), $R_0 = \|x_0 - x^*\|_2$ is the distance from the starting point to the solution, $\varkappa = \frac{L}{\mu}$ is the condition number of the problem.

Related work

Algorithm 1. Heavy-Ball method (HB)

Input: starting points x_0, x_1 (by default $x_0 = x_1$), number of iterations N, stepsize $\alpha > 0$, momentum parameter $\beta \in [0, 1]$

1: **for** $k = 0, \ldots, N-1$ **do**

2: $\quad x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k+1})$

3: **end for**

Output: x_k

Convergence guarantees for the heavy-ball method

The heavy-ball method [Polyak, 1964] (HB, Algorithm 1) is the first optimization method with momentum proposed in the literature. In [Polyak, 1964], the author proves the *local* $O\left(\sqrt{\frac{L}{\mu}}\log\left(\frac{1}{\varepsilon}\right)\right)$ convergence rate for twice continuously differentiable L-smooth and μ-strongly convex functions. The first global convergence results for HB are obtained in [Ghadimi, Feyzmahdavian, Johansson, 2015], where the authors derive *global* $O\left(\frac{LR_0^2}{\varepsilon}\right)$ convergence rate of HB and AHB for L-smooth convex ($\mu = 0$) functions and $O\left(\frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)\right)$ convergence rate of HB for L-smooth and μ-strongly convex functions. Although these results establish the global convergence of HB (and AHB in the convex case), the rates are non-accelerated, i. e., they are not optimal [Nemirovskij, Yudin, 1983] unlike the local convergence rate derived in [Polyak, 1964]. This issue is partially resolved in [Lessard, Recht, Packard, 2016], where the authors prove that HB converges with the asymptotically accelerated rate for strongly convex quadratic functions. Moreover, they also show that there exists a non-twice differentiable strongly convex function such that HB does not converge for this objective. Next, using Performance Estimation Problem tools [Taylor, Hendrickx, Glineur, 2017; Taylor, Van Scoy, Lessard, 2018; Taylor, Bach, 2019], one can show that for standard choices of parameters HB has the non-accelerated rate of convergence. However, the following question remains open: *does there exist a choice of parameters for* HB *such that the method converges globally with the accelerated rate for twice differentiable L-smooth and* (*strongly*) *convex functions?* Although we do not address this question in our work, we highlight it here due to its theoretical importance.

Nonmonotone behavior of the heavy-ball method

From the classical analysis of HB [Polyak, 1964], it is known that the following choice of parameters α and β ensures the best convergence rate for HB up to the numerical constant factors:

$$\alpha = \alpha^* = \frac{4}{\left(\sqrt{L} + \sqrt{\mu}\right)^2}, \quad \beta = \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}}\right)^2. \tag{4}$$

However, recently it was shown [Danilova, Kulakova, Polyak, 2018] that HB with optimal parameters suffers from the so-called *peak effect* at the beginning of the convergence. In particular, the maximal deviation can be of the order $\sqrt{\varkappa} = \sqrt{\frac{L}{\mu}}$. Similar results were also derived in [Mohammadi, Samuelson, Jovanović, 2021]. However, in practice, it is worth mentioning that the optimal parameters from (4) are rarely used and, as a result, the nonmonotonicity of HB is not that significant.

Maximal deviations on quadratic problems

In this section, we consider the instance of (1) with $f(x)$ being a quadratic function. That is, we assume that $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, where $\mathbf{A} \in \mathbb{S}^n_{++}$ is an $n \times n$ positive definite matrix. For this problem, we prove that the averaged heavy-ball method with a certain choice of parameters has a smaller deviation of the iterates from the optimum at initial iterations than the heavy-ball method with optimal parameters.

The heavy-ball method

Recently it was shown [Danilova, Kulakova, Polyak, 2018] that HB with optimal parameters (4) suffers from the so-called *peak effect* at the beginning of the convergence. In particular, according to the following theorem, the maximal deviation can be of the order $\sqrt{\varkappa}$.

Theorem 1 (Theorem 1 from [Danilova, Kulakova, Polyak, 2018]). *Consider* $f(x) = \frac{1}{2}x^\top \mathbf{A}x$, $\mathbf{A} = \operatorname{diag}\left(\mu, \lambda_2, \ldots, \lambda_{n-1}, L\right)$, *where* $\mu \leqslant \lambda_2 \leqslant \lambda_3 \leqslant \ldots \leqslant \lambda_{n-1} \leqslant L$. *Then for* $x^0 = x^1 = (1, 1, \ldots, 1)^\top$ *the iterates* $\{x_k\}_{k\geq 0}$ *produced by* HB *with* $\alpha = \alpha^*$, $\beta = \beta^*$ *satisfy*

$$\max_k \|x_k\|_\infty \geqslant \frac{\sqrt{\varkappa}}{2e}. \tag{5}$$

Algorithm 2. Averaged heavy-ball method (AHB)

Input: starting points x_0, x_1 (by default $x_0 = x_1$), number of iterations N, stepsize $\alpha > 0$, momentum parameter $\beta \in [0, 1]$

1: **for** $k = 1, \ldots, N-1$ **do**

2: $\quad x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k+1})$

3: $\quad \overline{x}_{k+1} = \frac{1}{k+2}\sum\limits_{i=0}^{k+1} x_i$ ▷ One can recurrently implement this step: $\overline{x}_{k+1} = \frac{k\overline{x}_k + x_{k+1}}{k+1}$

4: **end for**

Output: $\overline{x}_k$

The averaged heavy-ball method

In this subsection, we consider the modification of HB that returns the average of the iterates produced by HB. We call the resulting method averaged heavy-ball method (AHB, see Algorithm 2).

We start by showing that for the same initialization, AHB with $\alpha = \frac{1}{L}$ and not too large β has significantly more minor deviations than HB with optimal parameters when $\varkappa$ is sufficiently large under some assumptions on the spectrum of $\mathbf{A}$.

Theorem 2. *Consider* $f(x) = \frac{1}{2}x^\top \mathbf{A}x$ *with* $\mathbf{A} = \operatorname{diag}\left(\mu, \lambda_2, \ldots, \lambda_{n-1}, L\right)$, *where* $\mu \leqslant \lambda_2 \leqslant \lambda_3 \leqslant \ldots \leqslant \lambda_{n-1} \leqslant L$ *and* $\lambda_2 \geqslant 10\mu$, $L \geqslant 100\mu$. *Then for* $x^0 = x^1 = (1, 1, \ldots, 1)^\top$ *and for all* $k \geqslant 0$ *the iterates* $\{\overline{x}_k\}_{k\geq 0}$ *generated by* AHB *with* $\alpha = \frac{1}{L}$, $\beta \in \left[\left(1 - 3\sqrt{\frac{\mu}{L}}\right)^2, \left(1 - 2\sqrt{\frac{\mu}{L}}\right)^2\right]$ *satisfy*

$$\max_k \|\overline{x}_k\|_\infty \leqslant 2. \tag{6}$$

That is, comparing bounds (5) and (6) for $\varkappa \gg 1$, we conclude AHB with the parameters from Theorem 2 has much smaller deviations than HB with parameters from (4). However, Theorem 2 works only for the particular initialization. The guarantees independent of x^0, x^1 are much more valuable and that is what we derive in the next subsection.

Maximal deviation of AHB *for arbitrary initialization*

Consider the matrix representation of HB update rule:

$$\begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} = \mathbf{T} \cdot \begin{bmatrix} x_k - x_* \\ x_{k-1} - x_* \end{bmatrix} = \ldots = \mathbf{T}^k \cdot \begin{bmatrix} x_1 - x_* \\ x_0 - x_* \end{bmatrix}, \tag{7}$$

where

$$\mathbf{T} = \left[\begin{array}{c|c} (1+\beta)\mathbf{I} - \alpha\mathbf{A} & -\beta\mathbf{I} \\ \hline \mathbf{I} & \mathbf{0} \end{array}\right] \in \mathbb{R}^{2n \times 2n}, \quad \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix} \in \mathbb{R}^{2n}. \tag{8}$$

Therefore, we have

$$x_k - x^* = \underbrace{\begin{bmatrix} \mathbf{0} & \mathbf{I} \end{bmatrix}}_{\mathbf{C}} \mathbf{T}^k \begin{bmatrix} x_1 - x_* \\ x_0 - x_* \end{bmatrix}. \tag{9}$$

For convenience, we also introduce the following notation:

$$z_k = \begin{bmatrix} x_{k+1} - x_* \\ x_k - x_* \end{bmatrix}.$$

Following [Mohammadi, Samuelson, Jovanović, 2021], we study the worst case deviation $\|x_k - x_*\|_2$ in the relation to $\|z_0\|_2$, i. e., we focus on the following quantity:

$$\max_{k \geqslant 0} \sup_{z_0 \neq 0} \frac{\|x_k - x_*\|_2}{\|z_0\|_2} \overset{(9)}{=} \max_{k \geqslant 0} \sup_{z_0 \neq 0} \frac{\|\mathbf{C}\mathbf{T}^k z_0\|_2}{\|z_0\|_2} = \max_{k \geqslant 0} \|\mathbf{C}\mathbf{T}^k\|_2$$

that is the largest spectral norm of the matrices $\mathbf{C}\mathbf{T}^k$ for $k \geqslant 0$. Clearly, one can choose z_0, i. e., starting points x_0 and x_1, in such a way that z_0 is in the direction of the principal right singular vector of $\mathbf{C}\mathbf{T}^k$ implying $\|x_k - x_*\|_2 = \|\mathbf{C}\mathbf{T}^k\|_2 \|z_0\|_2$. Therefore, $\|\mathbf{C}\mathbf{T}^k\|_2$ is a tight and natural measure of the worst case deviation of the iterates produced by HB. Since this quantity depends on the choice of α and β, we denote it as $\mathrm{dev}_{\mathsf{HB}}(\alpha, \beta) := \max\limits_{k \geqslant 0} \|\mathbf{C}\mathbf{T}^k(\alpha, \beta)\|_2$.

For AHB we know

$$\overline{x}_k - x_* = \frac{1}{k+1} \sum_{t=0}^{k} (x_k - x_*) = \frac{1}{k+1} \sum_{t=0}^{k} \mathbf{C}\mathbf{T}^t \begin{bmatrix} x_1 - x_* \\ x_0 - x_* \end{bmatrix}.$$

We introduce new notation:

$$\mathrm{dev}_{\mathsf{AHB}}(\alpha, \beta) := \max_{k \geqslant 0} \left\| \frac{1}{k+1} \sum_{t=0}^{k} \mathbf{C}\mathbf{T}^t(\alpha, \beta) \right\|_2.$$

As for HB, $\mathrm{dev}_{\mathsf{AHB}}(\alpha, \beta)$ is also a reasonable measure of the worst case deviation of the iterates produced by AHB. Moreover, due to Jensen's inequality and convexity of $\|\cdot\|_2$ we have $\mathrm{dev}_{\mathsf{AHB}}(\alpha, \beta) \leqslant \mathrm{dev}_{\mathsf{HB}}(\alpha, \beta)$.

Theorem 3. *Consider* $f(x) = \frac{1}{2} x^\top \mathbf{A} x$ *with* $\mathbf{A} = \mathbf{A}^\top \succ 0$ *with eigenvalues* $\lambda_1 \leqslant \ldots \leqslant \lambda_n$, $\lambda_2 \geqslant F^2 \lambda_1$, $F > 14$, $F \leqslant \sqrt{\frac{\lambda_n}{\lambda_1}}$, $\lambda_n \geqslant 10\,000 \lambda_1$. *Then the maximal deviation of* AHB *and* HB *with* $\alpha = \frac{1}{L}$ *and* $\left(1 - \sqrt{\frac{\lambda_2}{\lambda_n}}\right)^2 < \beta \leqslant \left(1 - F\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2$ *is at least* $\frac{\sqrt{F^2-1}}{2e\sqrt{6}}$ *times smaller than the maximal deviation of* HB *with* $\alpha = \alpha^*$ *and* $\beta = \beta^*$ *given in* (4):

$$\mathrm{dev}_{\mathsf{AHB}}(\alpha, \beta) \leqslant \mathrm{dev}_{\mathsf{HB}}(\alpha, \beta) \leqslant \frac{2e\sqrt{6}}{\sqrt{F^2 - 1}} \mathrm{dev}_{\mathsf{HB}}(\alpha^*, \beta^*). \tag{10}$$

The constant $\frac{2e\sqrt{6}}{\sqrt{F^2-1}}$ can be sufficiently small and β can be sufficiently large at the same time when the condition number $\varkappa$ is large enough. For example, for $\varkappa = 10^8$ and $F = 200$ one can choose $\beta = \left(1 - \frac{F}{\sqrt{\varkappa}}\right)^2 \approx 0{,}96$ and get $\frac{2e\sqrt{6}}{\sqrt{F^2-1}} \approx 0{,}067$.

Convergence guarantees for nonquadratics

In this section, we study the convergence of AHB for problems (1) with (strongly) convex and smooth objectives. The first global convergence guarantees for HB and AHB in the convex case were obtained in [Ghadimi, Feyzmahdavian, Johansson, 2015]. In the same paper, the authors derived the convergence rate for HB in the strongly convex case. See the summary of known results in Table 1.

In contrast, for HB with averaging, there are no convergence results in the strongly convex case. Below we consider two options to derive such results.

Table 1. Summary of known and new results on the maximal deviation and complexity bounds for HB and its variants with averaging. Column "Max. deviation" contains the results on the maximal deviation of the methods on quadratic minimization problems (see Section 2 for details), columns "Complexity, $\mu = 0$" and "Complexity, $\mu > 0$" show iteration complexity bounds for the methods applied to (1) with f being L-smooth and convex / μ-strongly convex but not necessarily quadratic, i. e., the number of iterations needed to guarantee that the output of the method $\widehat{x}$ satisfies $f(\widehat{x}) - f(x_*) \leqslant \varepsilon$ where x_* is the solution of (1). Our results are highlighted in green. Notation: $\varkappa = \frac{L}{\mu}$ (condition number), $\Delta_0 = f(x_0) - f(x_*)$, $R_0 = \|x_0 - x_*\|_2$

Method	Citation	Max. deviation	Complexity, $\mu = 0$	Complexity, $\mu > 0$
HB	[Danilova, Kulakova, Polyak, 2018; Ghadimi, Feyzmahdavian, Johansson, 2015]	$\frac{\sqrt{\varkappa}}{2e}$ (1)	$\frac{LR_0^2}{\varepsilon}$ (2)	$\frac{\varkappa}{1-\beta}\log\frac{\Delta_0}{\varepsilon}$ (3)
AHB	[Ghadimi, Feyzmahdavian, Johansson, 2015]	N/A	$\frac{LR_0^2}{\varepsilon} + \frac{\beta LR_0^2}{(1-\beta)\varepsilon}$	N/A
AHB	Thm. 3 & 4 & 5	$\frac{\sqrt{6\varkappa}}{\sqrt{F^2-1}}$ (4)	$\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta}}{(1-\beta)\varepsilon}$	$\left(\varkappa + \frac{\varkappa\sqrt{\beta}}{1-\beta}\right)\log\frac{\mu R_0^2}{\varepsilon}$ (5)
WAHB	Thm. 4	$\frac{\sqrt{6\varkappa}}{\sqrt{F^2-1}}$ (6)	$\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta}}{(1-\beta)\varepsilon}$	$\left(\varkappa + \frac{\varkappa\sqrt{\beta}}{1-\beta}\right)\log\frac{LR_0^2\left(1+\frac{\sqrt{\beta}}{(1-\beta)}\right)}{\varepsilon}$

(1) This result is obtained for HB with optimal parameters from (4) (see Theorem 1).
(2) The complexity bound is obtained for iteration-dependent parameters: $\beta_k = \frac{k}{k+2}$, $\alpha_k = \frac{1}{L(k+1)}$.
(3) This result holds for $\alpha \in \left(0, \frac{1}{L}\right)$, $\beta \in \left[0, \sqrt{(1-\alpha L)(1-\alpha\mu)}\right]$. When $\varkappa \gg 1$ this assumption implies that $\beta \leqslant 0{,}75$. In practical applications, e. g., training deep neural networks, much larger values for parameter β are usually used.
(4) The result holds for a special class of quadratic functions described in Theorem 3. Parameters α and β for AHB are given there as well. Here F is such that $\lambda_2 \geqslant F^2\mu$, $F > 14$, $F \leqslant \sqrt{\varkappa}$, where λ_2 is the second smallest eigenvalue of the Hessian matrix. For large enough $\varkappa$ and F one can guarantee that maximal deviation for AHB with parameters from Theorem 3 is much smaller than for HB with optimal parameters from (4).
(5) The complexity bound is proven Restarted version of AHB (R-AHB, Algorithm 4).
(6) See (4) and Remark 1.

Weighted averaged heavy-ball method

One way to obtain them is to change the averaging weights, see the weighted averaged heavy-ball method (WAHB, Algorithm 3). When $w_k = 1$ for all $k \geqslant 0$ WAHB recovers AHB. However, it is natural to choose larger w_k for larger k: for such a choice of w_k the method gradually "forgets" about the early iterates that should lead to faster convergence. Guided by this intuition, we provide a rigorous analysis of WAHB with gradually increasing w_k.

Algorithm 3. Weighted averaged heavy-ball method (WAHB)

Input: number of iterations N, stepsize $\alpha > 0$, momentum parameter $\beta \in [0, 1]$, starting points x_0, x_1 (by default $x_1 = x_0 - \alpha\nabla f(x_0)$), weights for the averaging $\{w_k\}_{k=0}^N > 0$

for $k = 1, \ldots, N-1$ **do**

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

$\overline{x}_{k+1} = \frac{1}{W_{k+1}} \sum\limits_{i=0}^{k+1} w_i x_i$, where $W_{k+1} = \sum\limits_{i=0}^{k+1} w_i$ ▷ Recurrent analog: $\overline{x}_{k+1} = \frac{W_k \overline{x}_k + w_{k+1} x_{k+1}}{W_{k+1}}$

end for

Output: $\overline{x}_N$

REMARK 1. We emphasize that the proof of Theorem 3 holds for non-uniform averaging as well. That is, under assumptions of Theorem 3 we have

$$\mathrm{dev}_{\mathsf{WAHB}}(\alpha, \beta) \leqslant \mathrm{dev}_{\mathsf{HB}}(\alpha, \beta) \leqslant \frac{2e\sqrt{6}}{\sqrt{F^2 - 1}} \mathrm{dev}_{\mathsf{HB}}(\alpha^*, \beta^*),$$

where

$$\mathrm{dev}_{\mathsf{WAHB}}(\alpha, \beta) := \max_{k \geqslant 0} \left\| \frac{1}{W_k} \sum_{t=0}^{k} w_t \mathbf{C}\mathbf{T}^t(\alpha, \beta) \right\|_2.$$

In our derivations, we rely on the following representation of the update rule of HB with $x_1 = x_0 - \alpha\nabla f(x_0)$:

$$x_{k+1} = x_k - m_k, \quad m_k = \beta m_{k-1} + \alpha\nabla f(x_k), \quad m_{-1} = 0. \tag{11}$$

Indeed, since $m_{k-1} = x_{k-1} - x_k$ for all $k \geqslant 0$ (for convenience, we use the notation $x_{-1} = x_0$) we have

$$x_{k+1} = x_k - m_k = x_k - \alpha\nabla f(x_k) - \beta m_{k-1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k+1}).$$

Next, following [Mania et al., 2015; Yang, Lin, Li, 2016], we consider *perturbed* or *virtual* iterates:

$$\widetilde{x}_k = x_k - \frac{\beta}{1-\beta} m_{k-1}, \quad k \geqslant 0. \tag{12}$$

We notice that these iterates are not computed explicitly in the method. However, they turn out to be useful in the analysis because of the following relation: for all $k \geqslant 0$

$$\widetilde{x}_{k+1} = x_{k+1} - \frac{\beta}{1-\beta} m_k = x_k - \frac{1}{1-\beta} m_k =$$
$$= \widetilde{x}_k + \frac{\beta}{1-\beta} m_{k-1} - \frac{1}{1-\beta}\left(\beta m_{k-1} + \alpha\nabla f(x_k)\right) = \widetilde{x}_k - \frac{\alpha}{1-\beta}\nabla f(x_k). \tag{13}$$

Using this notation, we derive the following lemma measuring one iteration progress of HB.

Lemma 1. *Assume that f is L-smooth and μ-strongly convex. Let α and β satisfy*

$$0 < \alpha \leqslant \frac{1-\beta}{4L}, \quad \beta \in [0, 1). \tag{14}$$

Then, for all $k \geqslant 0$,

$$\frac{\alpha}{2(1-\beta)}\left(f(x_k) - f(x_*)\right) \leqslant \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)\|\widetilde{x}_k - x_*\|_2^2 - \|\widetilde{x}_{k+1} - x_*\|_2^2 + \frac{3L\alpha\beta^2}{(1-\beta)^3}\|m_{k-1}\|_2^2. \tag{15}$$

As the next step, it is natural to sum up inequalities (15) for $k = 0, 1, 2, \ldots, K$ with weights $w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k+1)}$ to get the bound on $f(\overline{x}_K) - f(x_*)$. However, in this case, we obtain

$$\frac{3L\alpha\beta^2}{(1-\beta)^3}\sum_{k=0}^{K} w_k\|m_{k-1}\|^2$$

in the upper bound for $f(\overline{x}_K) - f(x_*)$. Therefore, we need to estimate this sum and this is exactly what the next lemma is about.

Lemma 2. *Assume that f is L-smooth and μ-strongly convex. Let α and β satisfy*

$$0 < \alpha \leqslant \frac{(1-\beta)^2}{4L\sqrt{3\beta}}, \quad \beta \in [0, 1). \tag{16}$$

Then, for all $k \geqslant 0$,

$$\frac{3L\alpha\beta^2}{(1-\beta)^3}\sum_{k=0}^{K} w_k\|m_{k-1}\|^2 \leqslant \frac{\alpha}{4(1-\beta)}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big), \tag{17}$$

where $w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k+1)}$.

Combining Lemmas 1 and 2, we obtain the following result.

Theorem 4. *Assume that f is L-smooth and μ-strongly convex. Let α and β satisfy*

$$0 < \alpha \leqslant \min\left\{\frac{1-\beta}{4L}, \frac{(1-\beta)^2}{4L\sqrt{3\beta}}\right\}, \quad \beta \in [0, 1). \tag{18}$$

Then, after $K \geqslant 0$ iterations of WAHB*, we have*

$$f(\overline{x}_K) - f(x_*) \leqslant \frac{4(1-\beta)\|x_0 - x_*\|_2^2}{\alpha W_K}, \tag{19}$$

where $w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k+1)}$. That is, if $\mu > 0$, then

$$f(\overline{x}_K) - f(x_*) \leqslant \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{K} \frac{4(1-\beta)\|x_0 - x_*\|_2^2}{\alpha}, \tag{20}$$

and if $\mu = 0$, we have

$$f(\overline{x}_K) - f(x_*) \leqslant \frac{4(1-\beta)\|x_0 - x_*\|_2^2}{\alpha K}. \tag{21}$$

The following complexity results trivially follow from this theorem.

Corollar 1. *Let the assumptions of Theorem* 4 *hold and*

$$\alpha = \min\left\{\frac{1-\beta}{4L}, \frac{(1-\beta)^2}{4L\sqrt{3\beta}}\right\}.$$

Then, to achieve $f(\overline{x}_K) - f(x_*) \leqslant \varepsilon$ for $\varepsilon > 0$ WAHB *requires*

$$O\left(\left(\frac{L}{\mu} + \frac{L\sqrt{\beta}}{\mu(1-\beta)}\right)\log\frac{LR_0^2\left(1 + \frac{\sqrt{\beta}}{(1-\beta)}\right)}{\varepsilon}\right) \tag{22}$$

iterations when $\mu > 0$*, and*

$$O\left(\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta}}{(1-\beta)\varepsilon}\right) \tag{23}$$

iterations when $\mu = 0$*, where* $R_0 \geqslant \|x_0 - x_*\|_2$.

When $\mu = 0$ WAHB recovers AHB since $w_k = 1$ by definition. Therefore, in the convex case, this result establishes the complexity of AHB.

The restarted averaged heavy-ball method

An alternative way to achieve linear convergence in the strongly convex case for the heavy-ball method with averaging is to use the restarts technique. That is, consider the restarted averaged heavy-ball method (R-AHB, Algorithm 4). The work of the method is split into stages. Each stage is the run of AHB from the point obtained at the previous stage, the first stage initializes at the given point.

Algorithm 4. Restarted averaged heavy-ball method (R-AHB)

Input: number of restarts τ, numbers of iterations $\{N_t\}_{t=1}^{\tau}$, stepsizes $\{\alpha_t\}_{t=1}^{\tau} > 0$, momentum parameters $\{\beta_t\}_{t=1}^{\tau} \in [0, 1]$, starting point x_0

1: $\widehat{x_0} = x_0$
2: **for** $t = 1, \ldots, \tau$ **do**
3: Run AHB (Algorithm 2) for N_t iterations with stepsize α_t, momentum parameter β_t, and starting points $\widehat{x}_{t-1}$, $\widehat{x}_{t-1} - \alpha_t \nabla f(\widehat{x}_{t-1})$. Define the output of AHB by $\widehat{x}_t$.
4: **end for**

Output: $\widehat{x}_\tau$

Based on the convergence result for AHB in the convex case, one can get the convergence rate of R-AHB in the strongly convex case.

Theorem 5. *Assume that* f *is* L*-smooth and* μ*-strongly convex. Let* $\alpha_t = \alpha$, $\beta_t = \beta$, $N_t = N$ *for all* $t = 1, \ldots, \tau$ *and*

$$0 < \alpha \leqslant \min\left\{\frac{1-\beta}{4L}, \frac{(1-\beta)^2}{4L\sqrt{3\beta}}\right\}, \quad \beta \in [0, 1), \quad N = \left\lceil \frac{16(1-\beta)}{\alpha\mu} \right\rceil. \tag{24}$$

Then, after $\tau = \max\left\{\left\lceil \log_2\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil - 1, 1\right\}$ *iterations with* $R_0 \geqslant \|x_0 - x_*\|_2$ R-AHB *produces such a point* $\widehat{x}_\tau$ *that* $f(\widehat{x}_\tau) - f(x_*) \leqslant \varepsilon$*. Furthermore, if*

$$\alpha = \min\left\{\frac{1-\beta}{4L}, \frac{(1-\beta)^2}{4L\sqrt{3\beta}}\right\},$$

then the total number of AHB *iterations equals*

$$O\left(\left(\frac{L}{\mu} + \frac{L\sqrt{\beta}}{\mu(1-\beta)}\right)\log\frac{\mu R_0^2}{\varepsilon}\right). \tag{25}$$

Numerical Experiments

We conducted several numerical experiments to compare the behavior of HB with and without averaging applied to minimize quadratic functions and solve the logistic regression problem. The code was written in Python 3.7 using standard libraries.

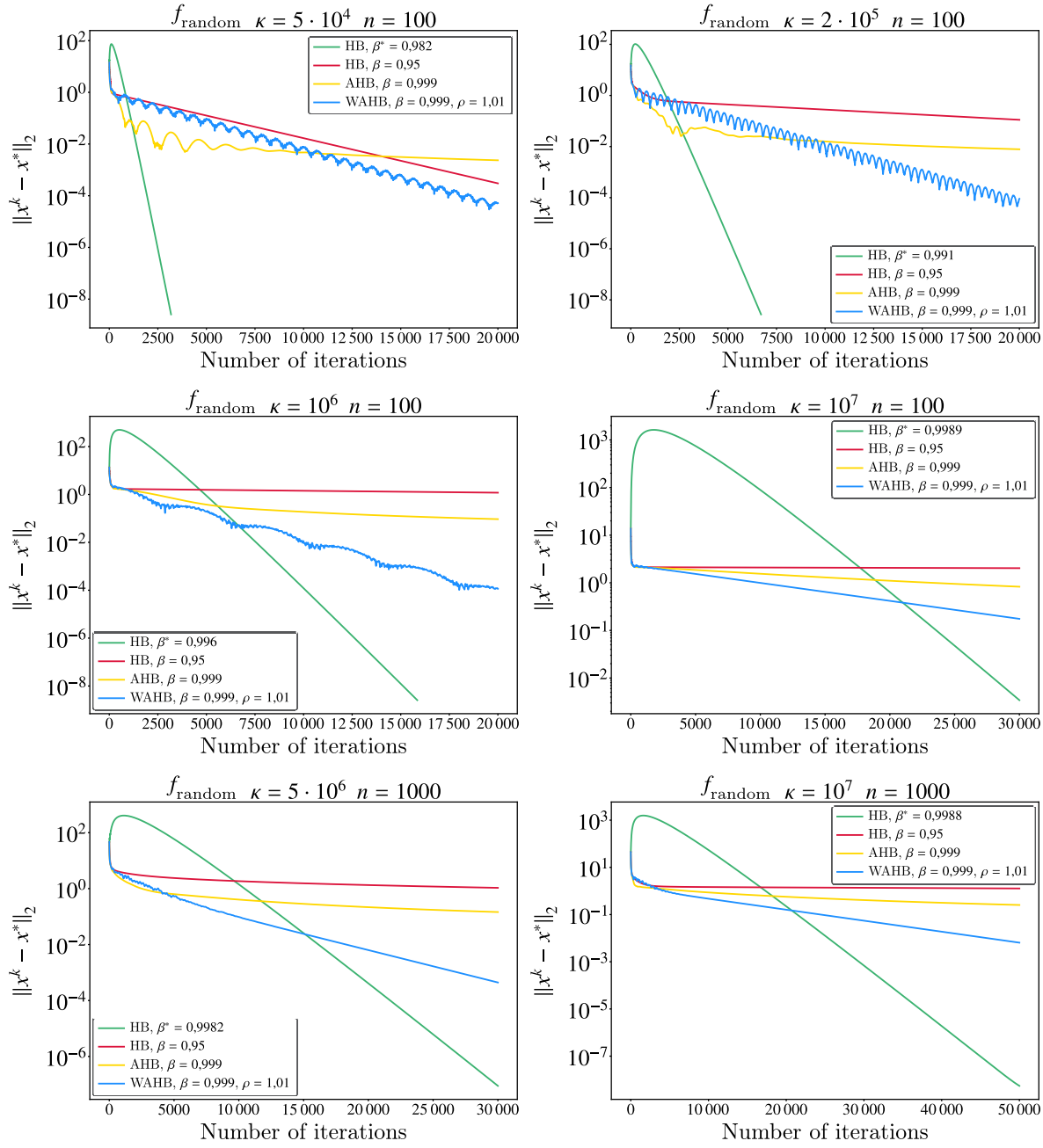

Figure 1. Trajectories of HB, AHB, and WAHB applied to minimize a quadratic function from (26) with different condition numbers $\varkappa$ and dimension n

Quadratic Functions

In this section, we consider three quadratic functions:

$$f_{\text{random}}(x) = \frac{1}{2} x^\top \mathbf{A}_{\text{rand}} x - (x^*)^\top \mathbf{A}_{\text{rand}} x, \tag{26}$$

$$f_{\text{Nesterov}}(x) = \frac{L-\mu}{8}\left(x_1^2 + \sum_{i=1}^{n-1}(x_i - x_{i+1})^2 - 2x_1\right) + \frac{\mu}{2}\|x\|^2, \tag{27}$$

$$f_{\text{Toeplitz}}(x) = \frac{1}{2} x^\top \mathbf{A}_{\text{Toeplitz}} x, \tag{28}$$

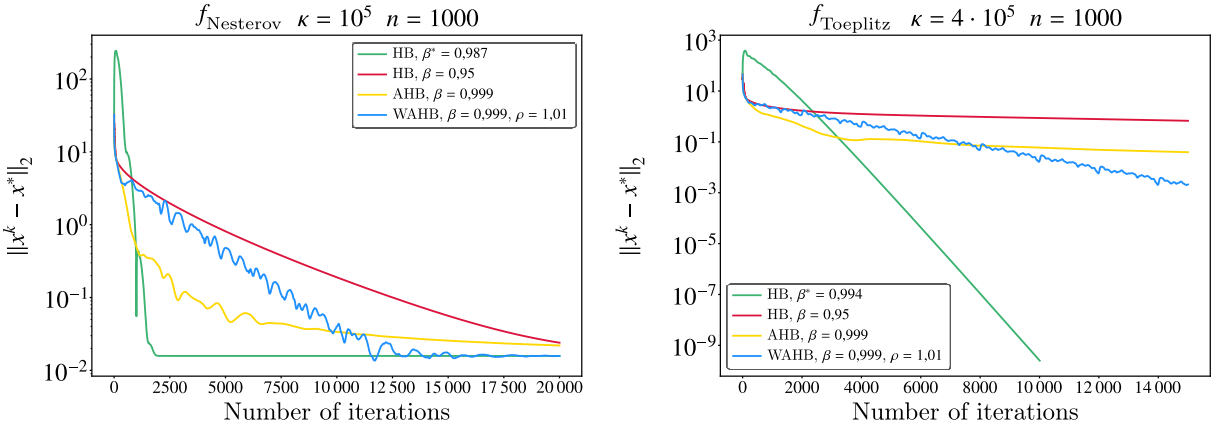

Figure 2. Trajectories of HB, AHB, and WAHB applied to minimize a quadratic functions from (27) and (28) with condition numbers $\varkappa \sim 10^5$ and dimension $n = 1000$

where matrix $\mathbf{A}_{\text{rand}} = \widehat{\mathbf{A}}^\top \widehat{\mathbf{A}}$, the elements of matrix $\widehat{\mathbf{A}} \in \mathbb{R}^{n \times n}$ are independently sampled from the standard Gaussian distribution, and $\mathbf{A}_{\text{Toeplitz}} \in \mathbb{R}^{n \times n}$ is a Toeplitz with a first row $(2, -1, 1, 0, \ldots, 0)$. The function from (27) is a classical function used to derive lower bounds for the complexity of first-order methods applied to minimize smooth strongly convex functions [Nesterov, 2018].

We run HB with $\beta = 0{,}95$ (standard choice of β), AHB and WAHB with $\beta = 0{,}999$ (large β) to minimize each of these functions. For these methods we used stepsize $\alpha = \frac{1}{L}$. The weights for WAHB were chosen as $w_k = \rho^k$ for $\rho = 1{,}01$. Moreover, we also tested HB with optimal parameters from (4). One can find the results in Figures 1 and 2.

These results show that methods with averaging (AHB and WAHB) converge reasonably well during the first iterations of the method even with large $\beta = 0{,}999$, which was larger than the optimal β^* in all our experiments. Moreover, unlike HB with optimal parameters, AHB and WAHB do not suffer from the peak effect. The absence of peak effect allows us to use HB with averaging for the first iterates and then restart the method. Finally, we emphasize that HB with $\beta = 0{,}95$ converges slower than WAHB with $\beta = 0{,}999$ in all our experiments and slower than AHB with $\beta = 0{,}999$ in almost all experiments (except the first one shown in Figure 1). We also tested HB with $\beta = 0{,}999$ and observed very slow convergence for the method in this case.

To conclude, our experiments on quadratic functions highlight the benefits of using AHB and WAHB with large β and standard $\alpha = \frac{1}{L}$.

Logistic regression with ℓ_2-regularization

Next, we also consider logistic regression with ℓ_2-regularization:

$$\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{m} \sum_{i=1}^{m} \log\left(1 + \exp\left(-y_i \cdot (\mathbf{A}x)_i\right)\right) + \frac{\ell_2}{2}\|x\|_2^2 \right\}, \tag{29}$$

where m is the total number of data points/samples, $y_i \in \{-1, 1\}$ is a label of the ith datapoint, and $\mathbf{A} \in \mathbb{R}^{m \times d}$ is a feature matrix. This function is known to be ℓ_2-strongly convex and $(L + \ell_2)$-smooth with $L = \frac{\sigma_{\max}^2(\mathbf{A})}{4m}$, where $\sigma_{\max}(A)$ is the maximal singular value of matrix $\mathbf{A}$. We take the datasets, i. e., pairs of $(\mathbf{A}, \{y_i\}_{i=1}^m)$, from LIBSVM library [Chang, Lin, 2011], see the summary of the considered datasets in Table 2.

We run HB, AHB, and WAHB with different momentum parameters β solve this problem. Moreover, we also tested a modification of AHB called the tail-averaged heavy-ball method (TAHB, see

Table 2. Summary of the considered datasets for the logistic regression

	a9a	phishing	w8a
m (# of data points)	32 561	11 055	49 749
d (# of features)	123	68	300

Algorithm 5. Tail-averaged heavy-ball method (TAHB)

Input: starting points x_0, x_1 (by default $x_0 = x_1$), number of iterations N, stepsize $\alpha > 0$, momentum parameter $\beta \in [0, 1]$, tail size $s \geqslant 0$

1: **for** $k = 1, \ldots, N-1$ **do**

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

3: $\overline{x}_{k+1} = \begin{cases} \dfrac{1}{k+2}\displaystyle\sum_{i=0}^{k+1} x_i, & \text{if } k+1 < s, \\ \dfrac{1}{s}\displaystyle\sum_{i=0}^{s-1} x_{k+1-i}, & \text{if } k+1 \geqslant s \end{cases}$ ▷ It is required to store the last s iterates

4: **end for**

Output: $\overline{x}_k$

Algorithm 5) with $s \in \{10, 50\}$.[1] The weights for WAHB were chosen as $w_k = \rho^k$ for $\rho \in \{1{,}1, 1{,}01\}$. Next, we chose parameter β from the set {0,9, 0,95, 0,99, 0,999}, and tuned stepsize parameter $\alpha \in \{2^{-4}, 2^{-3}, 2^{-2}, 2^{-1}, 1, 2, 4, 8, 16, 32, 64, 128, 256\} \cdot \frac{1}{L}$ for each method separately for given β (and for given ρ in the case of WAHB, for given s for TAHB). The result are shown in Figures 3–6.

Figures 3–5. The plots show that for small β, i. e., $\beta = 0{,}9, 0{,}95$, HB does not have significant oscillations and WAHB and TAHB have comparable performance. However, for larger β, i. e., $\beta = 0{,}99, 0{,}999$, the behavior of HB is signigicantly nonmonotone and oscillations are quite large. In contrast, WAHB and TAHB have much smaller oscillations and converge faster than HB. These facts illustrate the advantages of using proper averaging scheme for HB (either in the form of WAHB or TAHB).

Figure 6. In these plots, we highlight the effect of averaging for large β. That is, we compare HB with standard and commonly used choice of β ($\beta = 0{,}95$) and TAHB with $\beta = \{0{,}95, 0{,}99\}$. Moreover, for $\ell_2 > 0$ we also tested HB with optimal parameters from (4). The results for all considered datasets show that TAHB with $\beta = 0{,}95$ has comparable performance with HB and oscillates smaller, while TAHB with $\beta = 0{,}99$ is always slower than TAHB with $\beta = 0{,}95$. Next, when $\ell_2 = \frac{L}{100000}$ (ill-conditioned problems), TAHB with $\beta = 0{,}99$ is as fast as HB with optimal parameters but has smaller oscillations. Finally, when $\ell_2 = \frac{L}{1000}$ (well-conditioned problems), HB with optimal parameters has negligible oscillations and shows the best performance. Such behavior is natural since for the well-conditioned problems HB does not suffer significantly from the nonmonotone behavior and peak-effect.

Conclusion

This paper shows the advantages of using averaging for the heavy-ball method both in theory and practice. That is, our theory and experiments imply that averaging helps to reduce the oscillations of HB. Although the derived theoretical convergence guarantees for HB with averaging are not better than the existing ones for HB, in our experiments, we observe that HB with a properly adjusted averaging scheme can converge faster than HB without averaging. In particular, we observe this phenomenon

[1] In our experiments, TAHB with $s \geqslant 100$ performed significantly worse than TAHB with $s = 50$. Therefore, we report only the resuts for $s \in \{10, 50\}$.

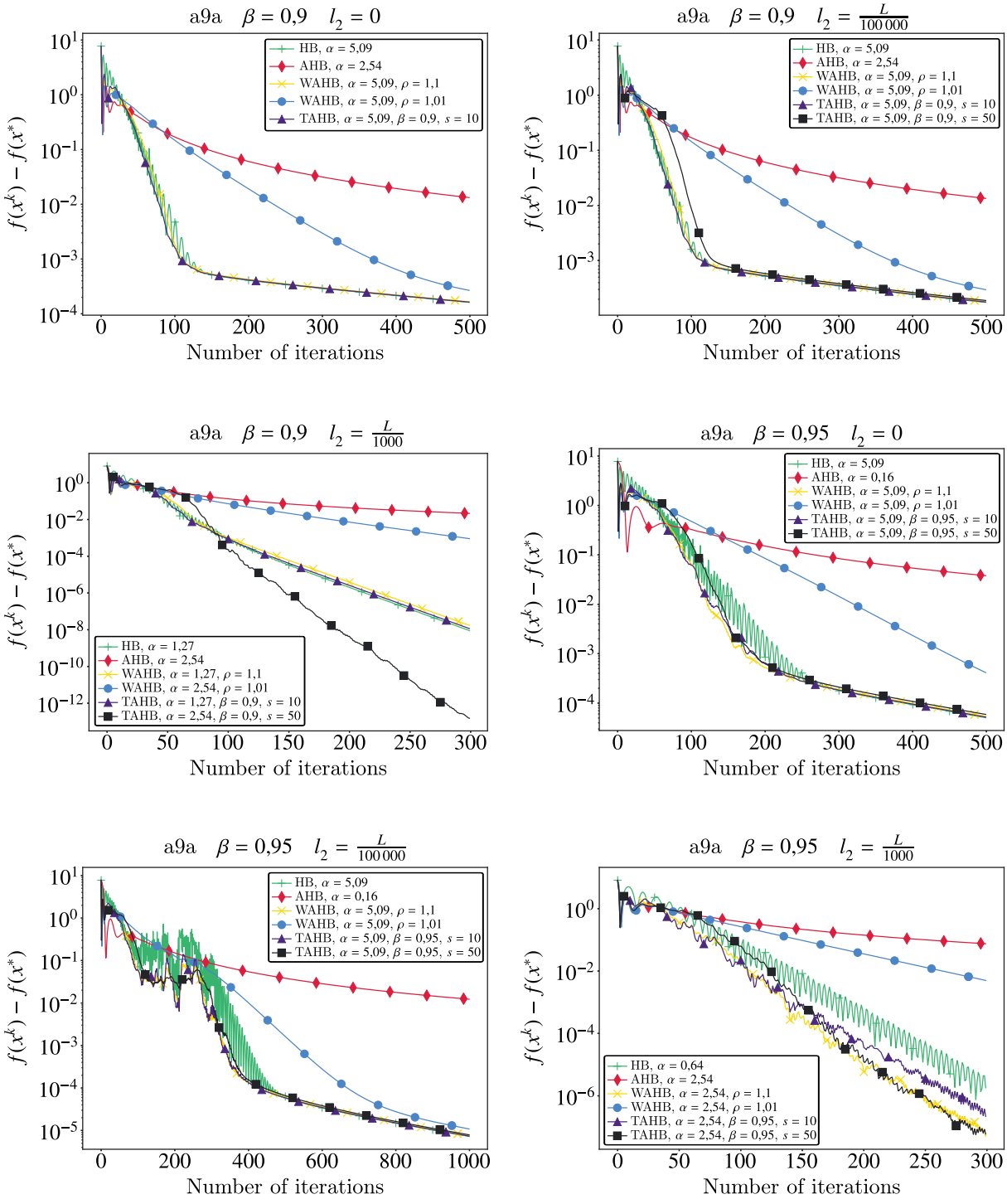

Figure 3. Trajectories of HB, AHB, WAHB, and TAHB with different momentum parameters β applied to solve the logistic regression problem with ℓ_2-regularization for a9a dataset. Stepsize α was tuned for each method and each choice of β (and ρ, s) separately

when the momentum parameter β for averaged versions of HB is chosen to be large enough, e. g., larger than the standard choice of $\beta = 0{,}95$ and sometimes larger than the optimal choice of β from (4).

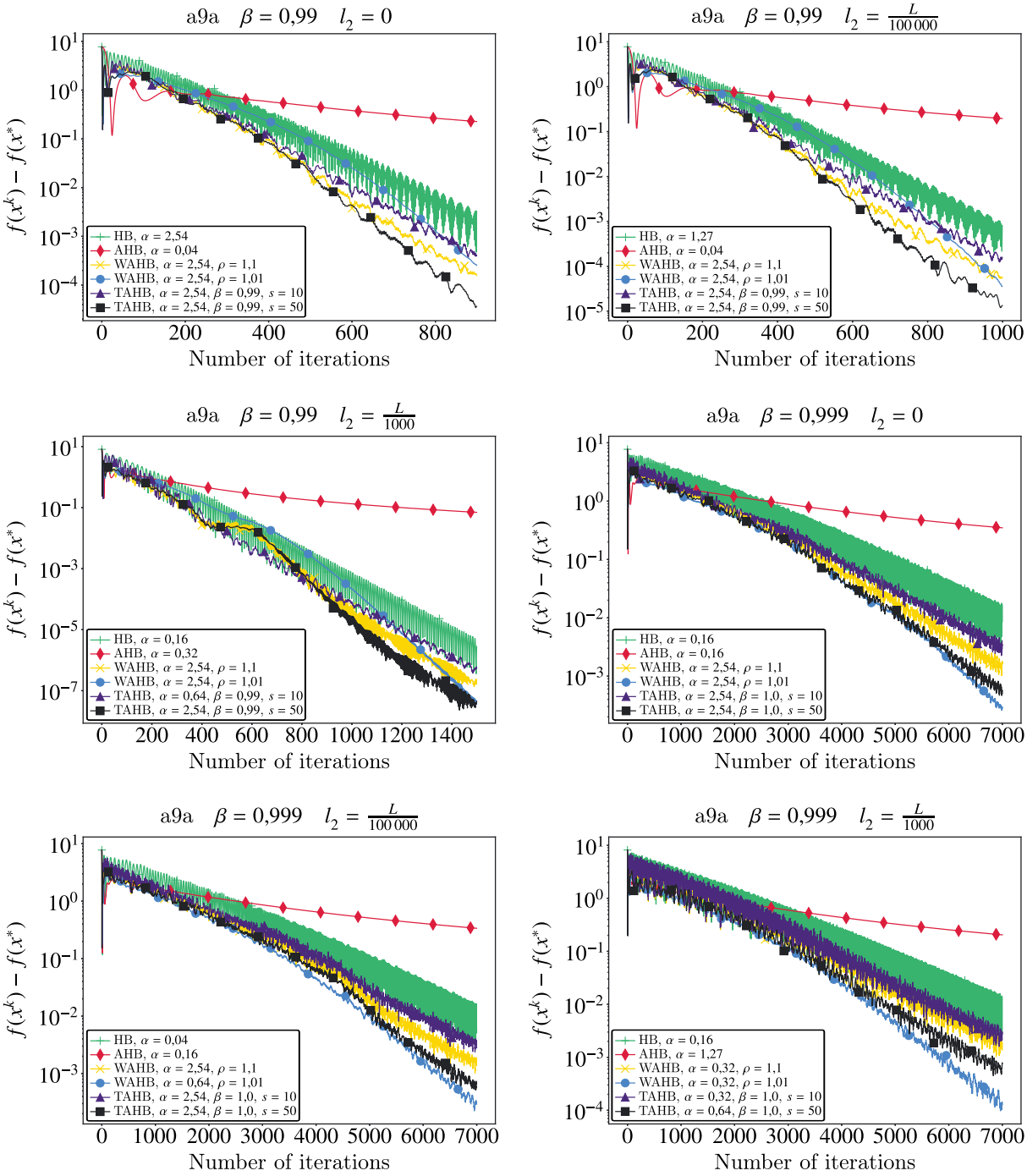

Figure 3 (ending)

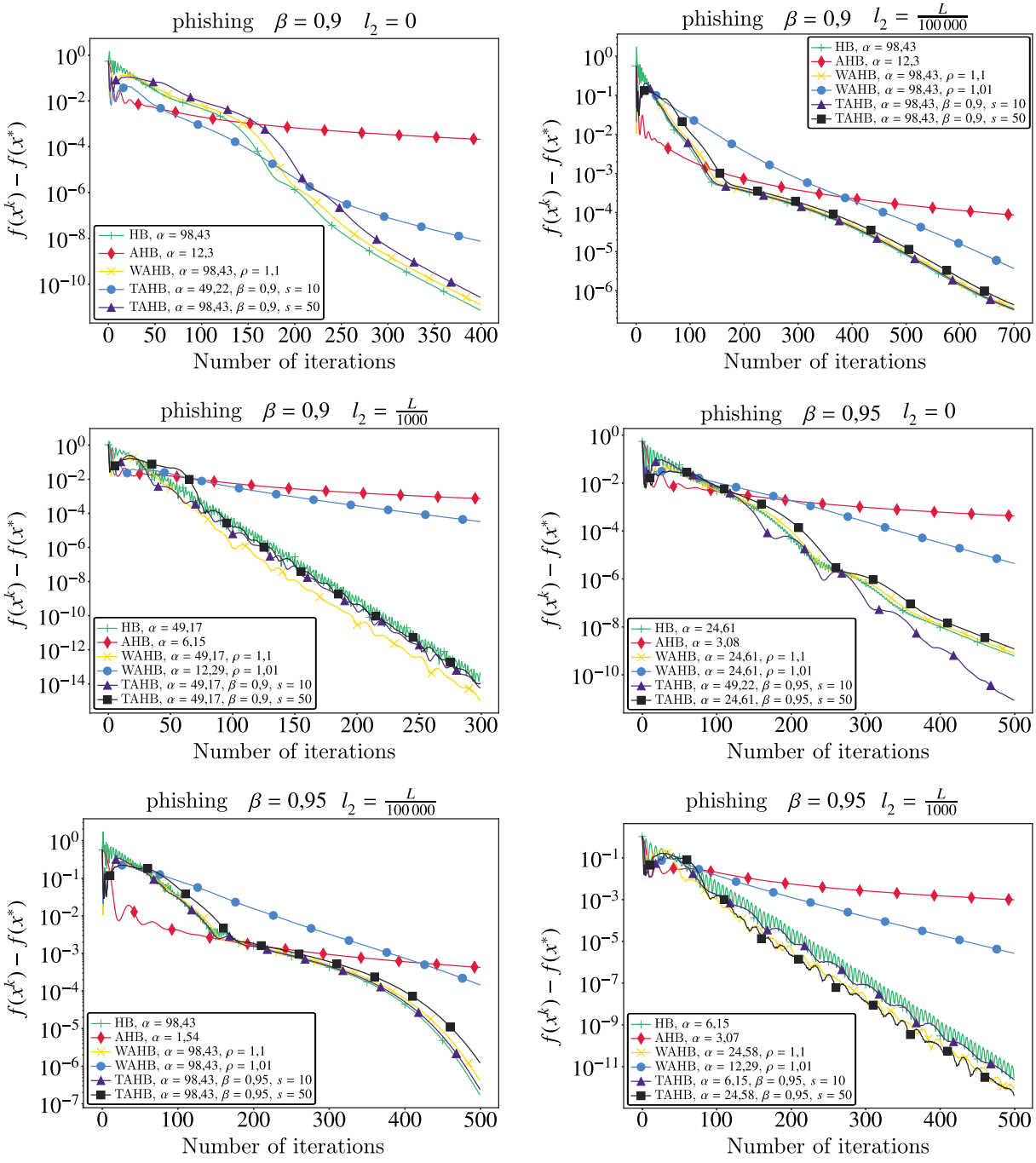

Figure 4. Trajectories of HB, AHB, WAHB, and TAHB with different momentum parameters β applied to solve the logistic regression problem with ℓ_2-regularization for `phishing` dataset. Stepsize α was tuned for each method and each choice of β (and ρ, s) separately

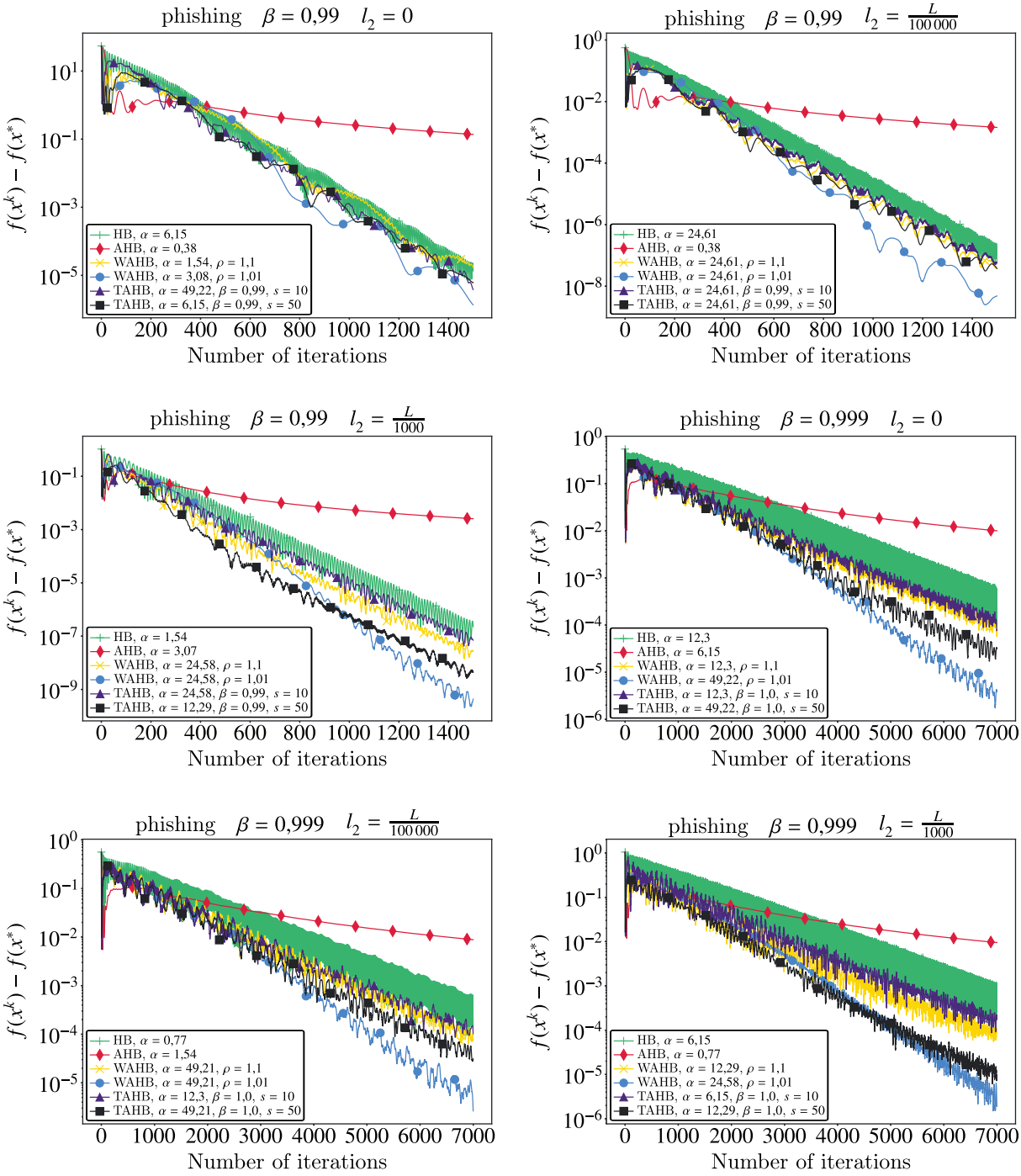

Figure 4 (ending)

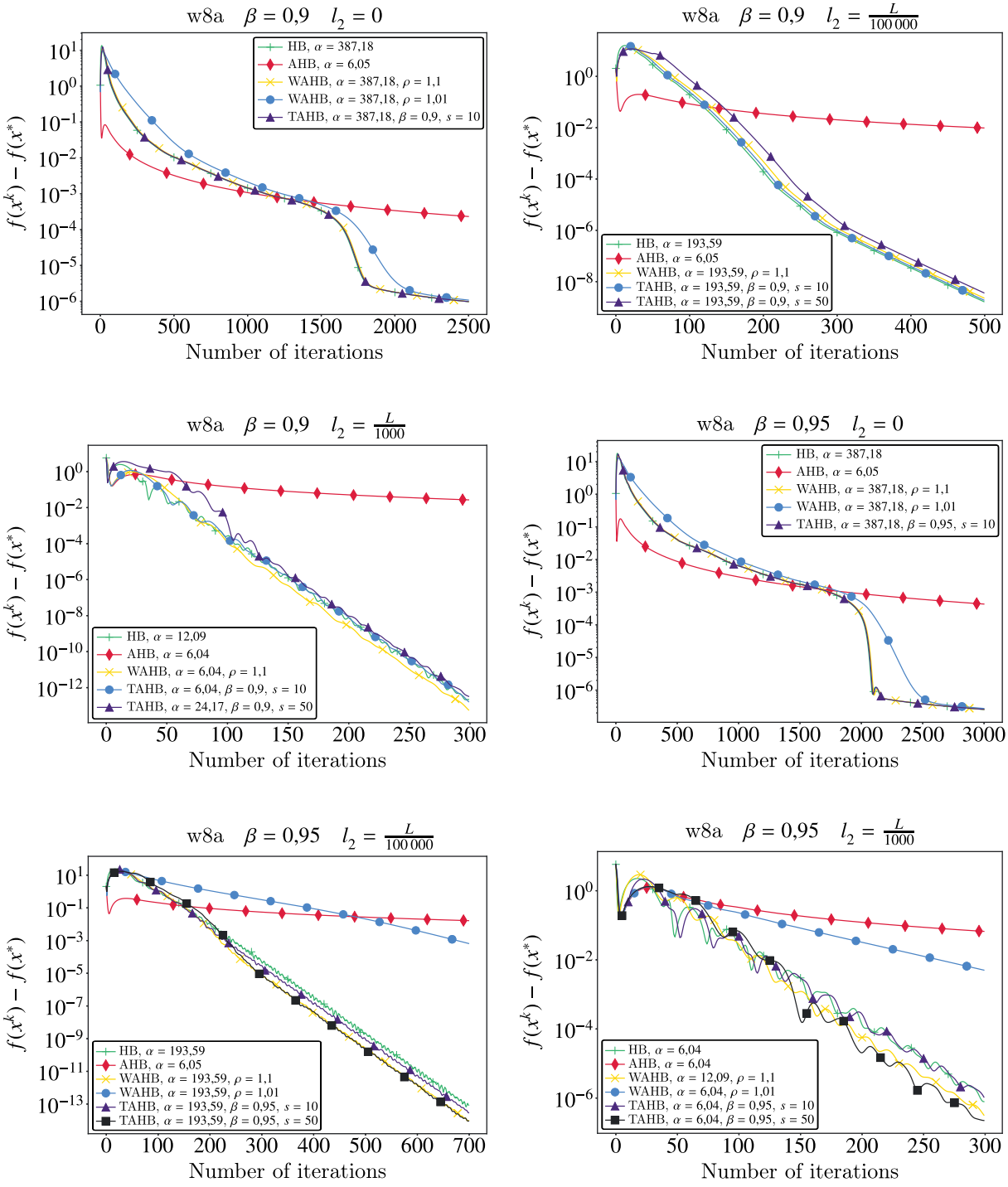

Figure 5. Trajectories of HB, AHB, WAHB, and TAHB with different momentum parameters β applied to solve the logistic regression problem with ℓ_2-regularization for `phishing` dataset. Stepsize α was tuned for each method and each choice of β (and ρ, s) separately

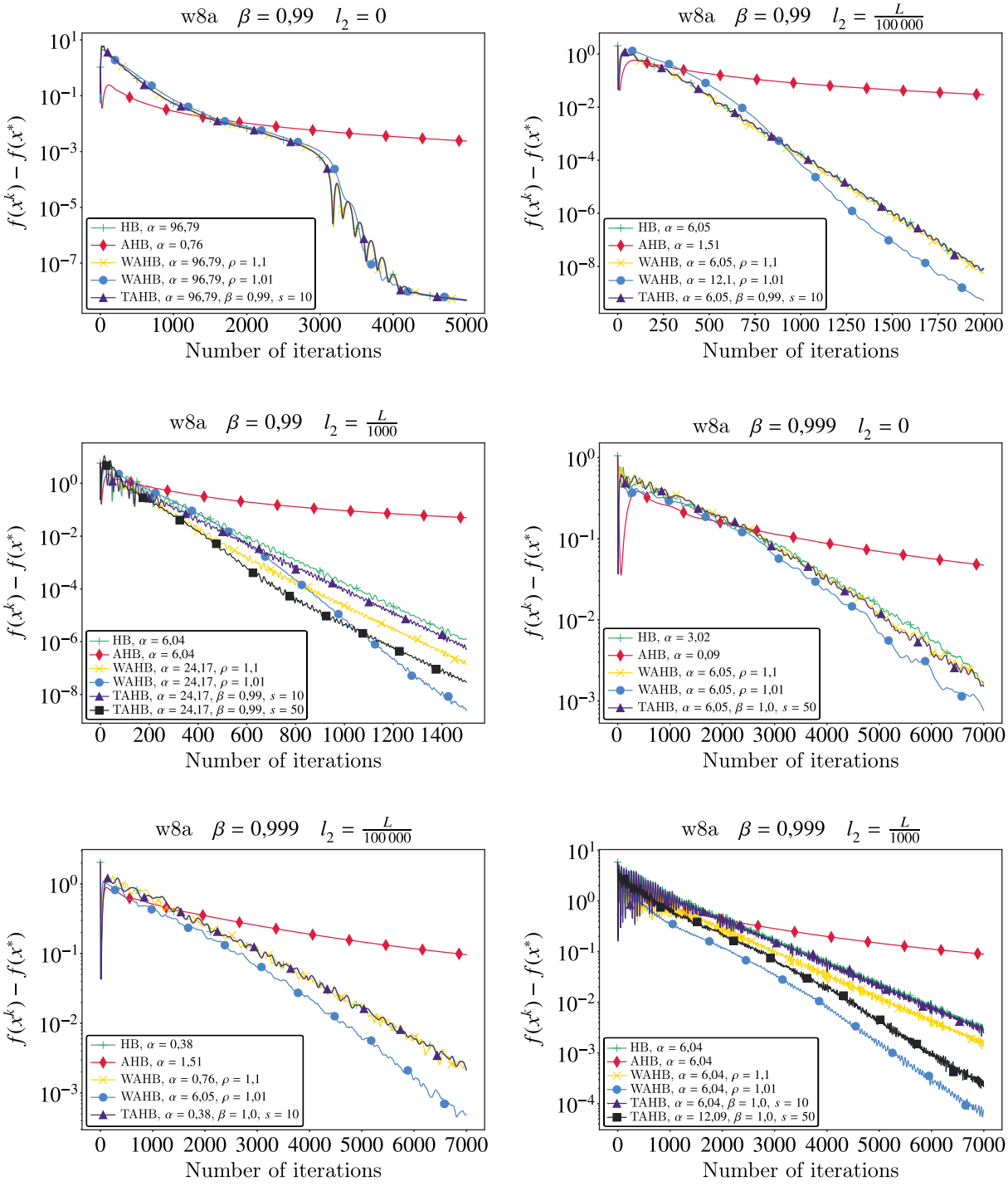

Figure 5 (ending)

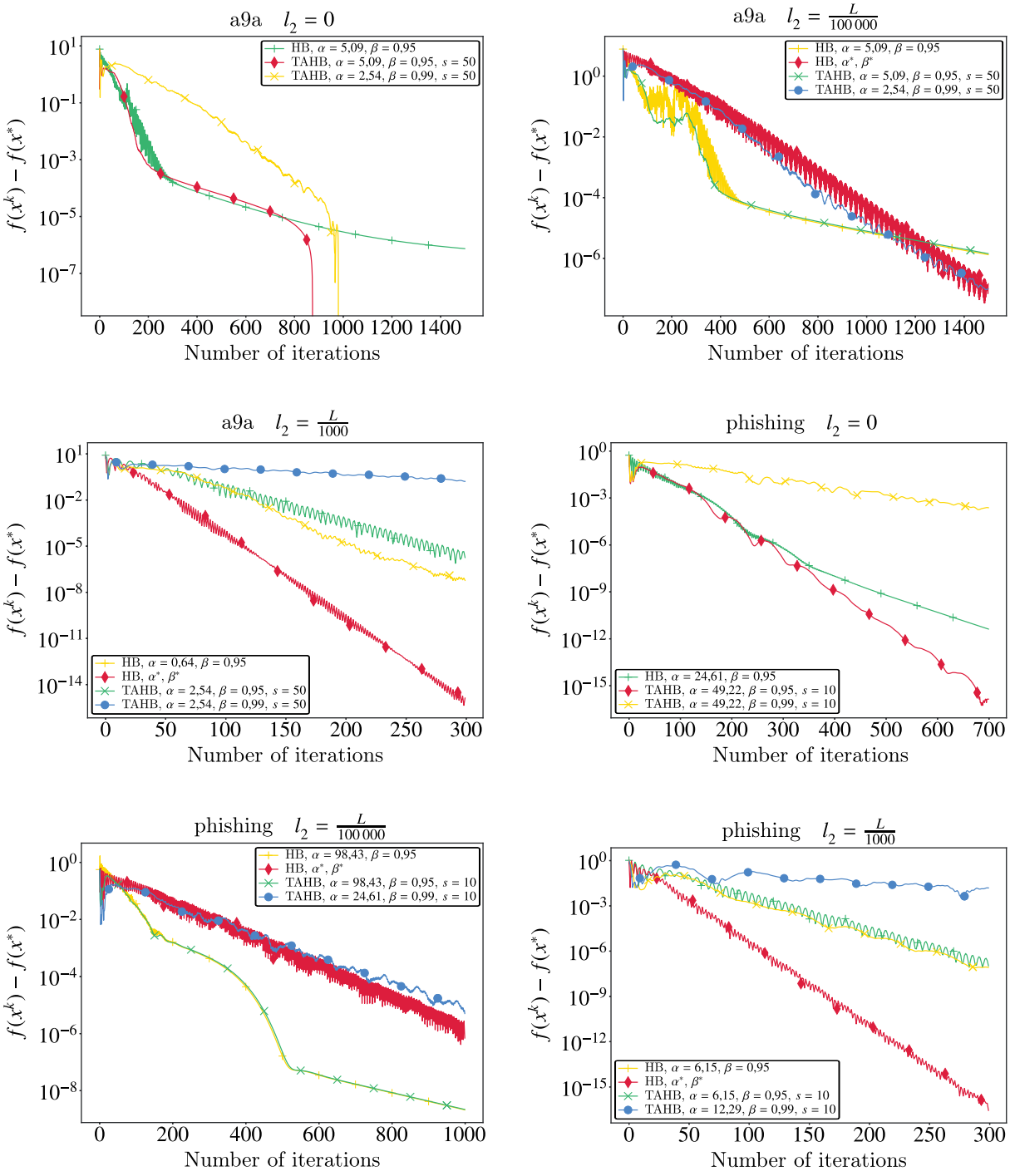

Figure 6. Trajectories of HB with $\beta = 0{,}95$ (standard choice of β) and TAHB with $\beta = 0{,}95$ and $\beta = 0{,}99$ (large β) applied to solve the logistic regression problem with ℓ_2-regularization for dataset from Table 2. Stepsize parameter α was tuned for each method separately. For $\ell_2 > 0$ we also show the trajectories of HB with optimal parameters $\alpha = \alpha^*$ and $\beta = \beta^*$ from (4)

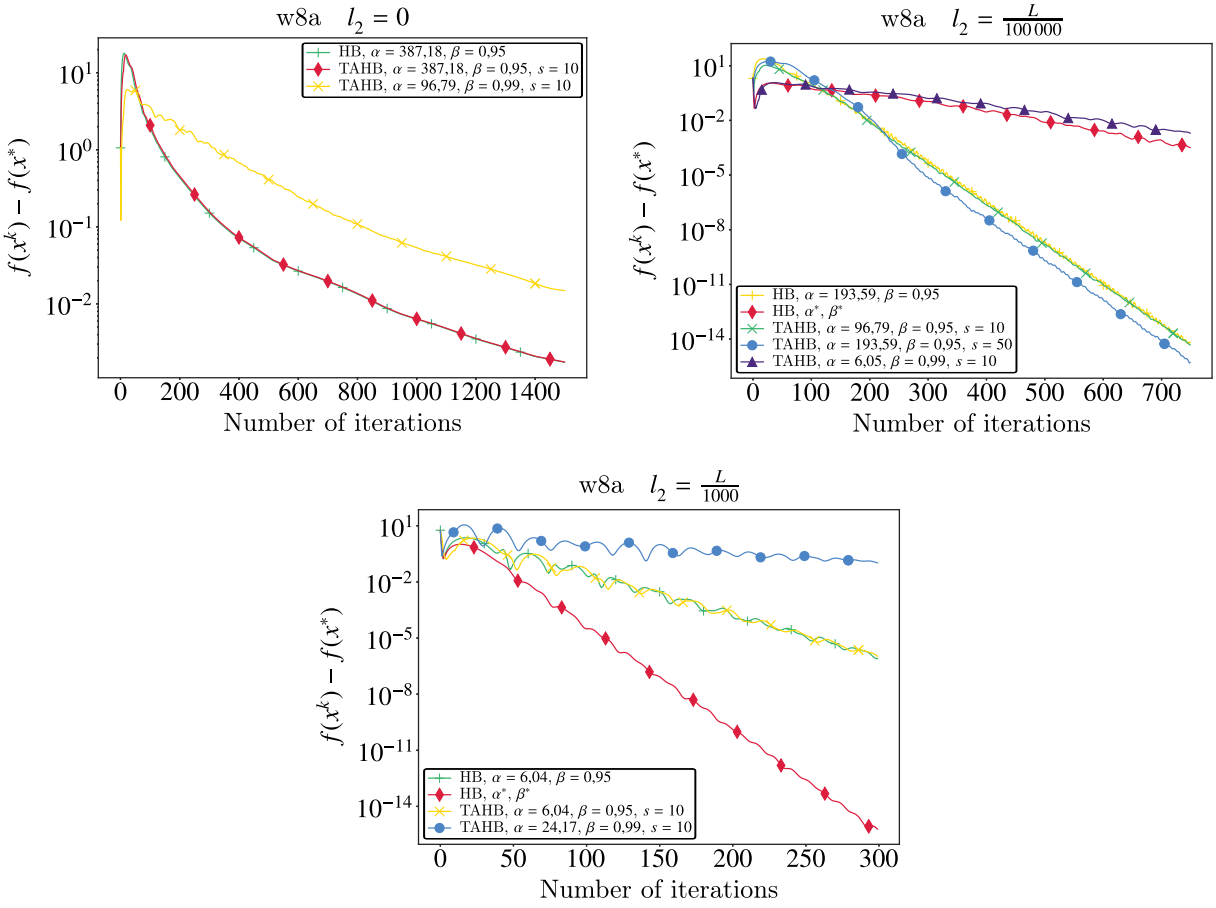

Figure 6 (ending)

References

Chang C. C., Lin C. J. LIBSVM: a library for support vector machines // ACM transactions on intelligent systems and technology (TIST). — 2011. — Vol. 2, No. 3. — P. 1–27.

Danilova M., Dvurechensky P., Gasnikov A., Gorbunov E., Guminov S., Kamzolov D., Shibaev I. Recent theoretical advances in non-convex optimization // arXiv preprint. — 2020. — https://arxiv.org/pdf/2012.06188

Danilova M., Kulakova A., Polyak B. Non-monotone behavior of the heavy ball method // International Conference on Difference Equations and Applications. — 2018. — P. 213–230.

Defazio A. Understanding the role of momentum in non-convex optimization: Practical insights from a Lyapunov analysis // 2020. — https://onikle.com/articles/317396

Ghadimi E., Feyzmahdavian H. R., Johansson M. Global convergence of the heavy-ball method for convex optimization // European control conference (ECC). — 2015. — P. 310–315.

Gorbunov E., Bibi A., Sener O., Bergou E. H., Richtárik P. A stochastic derivative free optimization method with momentum // arXiv preprint. — 2019. — https://arxiv.org/pdf/1905.13278

Lessard L., Recht B., Packard A. Analysis and design of optimization algorithms via integral quadratic constraints // SIAM Journal on Optimization. — 2016. — Vol. 26, No. 1. — P. 57–95.

Mania H., Pan X., Papailiopoulos D., Recht B., Ramchandran K., Jordan M. I. Perturbed iterate analysis for asynchronous stochastic optimization // SIAM Journal on Optimization. — 2017. — Vol. 27, No. 4. — P. 2202–2229.

Mishchenko K., Gorbunov E., Takáč M., Richtárik P. Distributed learning with compressed gradient differences // arXiv preprint. — 2019. — https://arxiv.org/pdf/1901.09269

Mohammadi H., Samuelson S., Jovanović M. R. Transient growth of accelerated first-order methods for strongly convex optimization problems // arXiv preprint. 2021. — https://arxiv.org/pdf/2103.08017

Nemirovskij A. S., Yudin D. B. Problem complexity and method efficiency in optimization // New York: J. Wiley & Sons, 1983.

Nesterov Yu. A method for unconstrained convex minimization problem with the rate of convergence $O(1/k^2)$ // Doklady an ussr. — 1983. — Vol. 269. — P. 543–547.

Nesterov Yu. Lectures on convex optimization. — Berlin, Germany: Springer International Publishing, 2018. — Vol. 137.

Polyak B. T. Some methods of speeding up the convergence of iteration methods // Ussr computational mathematics and mathematical physics. — 1964. — Vol. 4, No. 5. — P. 1–7.

Polyak B. Introduction to Optimization. — New York: Optimization Software, 1987.

Taylor A., Bach F. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions // Conference on Learning Theory. — 2019. — P. 2934–2992.

Taylor A. B., Hendrickx J. M., Glineur F. Performance estimation toolbox (PESTO): automated worst-case analysis of first-order optimization methods // IEEE 56th Annual Conference on Decision and Control (CDC). — 2017. — P. 1278–1283.

Taylor A., Van Scoy B., Lessard L. Lyapunov functions for first-order methods: Tight automated convergence guarantees // International Conference on Machine Learning. — 2018. — P. 4897–4906.

Yang T., Lin Q., Li Z. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization // arXiv preprint. — 2016. — https://arxiv.org/pdf/1604.03257

Yu H., Jin R., Yang S. On the linear speedup analysis of communication efficient momentum SGD for distributed non-convex optimization // International Conference on Machine Learning. — 2019. — P. 7184–7193.

Appendices

Basic inequalities

For all $a, b \in \mathbb{R}^n$ and $\lambda > 0$, $q \in (0, 1]$

$$|\langle a, b\rangle| \leqslant \frac{\|a\|_2^2}{2\lambda} + \frac{\lambda\|b\|_2^2}{2}, \tag{30}$$

$$\|a + b\|_2^2 \leqslant 2\|a\|_2^2 + 2\|b\|_2^2, \tag{31}$$

$$\|a + b\|^2 \leqslant (1 + \lambda)\|a\|^2 + \left(1 + \frac{1}{\lambda}\right)\|b\|^2, \tag{32}$$

$$\langle a, b\rangle = \frac{1}{2}\left(\|a + b\|_2^2 - \|a\|_2^2 - \|b\|_2^2\right), \tag{33}$$

$$\left(1 - \frac{q}{2}\right)^{-1} \leqslant 1 + q, \tag{34}$$

$$\left(1 + \frac{q}{2}\right)(1 - q) \leqslant 1 - \frac{q}{2}. \tag{35}$$

Auxiliary results

Lemma 3 (Lemma 1 from [Mohammadi, Samuelson, Jovanović, 2021]). *Let ρ_1 and ρ_2 be the eigenvalues of the matrix $\mathbf{M} = \left[\begin{smallmatrix} a & b \\ 1 & 0 \end{smallmatrix}\right]$ and let k be a positive integer. If $\rho_1 \neq \rho_2$, then we have*

$$\mathbf{M}^k = \frac{1}{\rho_2 - \rho_1} \begin{bmatrix} \rho_2^{k+1} - \rho_1^{k+1} & \rho_1\rho_2(\rho_1^k - \rho_2^k) \\ \rho_2^k - \rho_1^k & \rho_1\rho_2(\rho_1^{k-1} - \rho_2^{k-1}) \end{bmatrix}.$$

Moreover, if $\rho_1 = \rho_2 = \rho$, the matrix $\mathbf{M}^k$ satisfies

$$\mathbf{M}^k = \begin{bmatrix} (k+1)\rho^k & -k\rho^{k+1} \\ k\rho^{k-1} & (1-k)\rho^k \end{bmatrix}.$$

Missing proofs from Section 2

In this section, for x, we use the upper index for an iteration counter, and the lower index denotes the component of the vector.

Proof of Theorem 2

Rewriting the update rule of HB for $f(x) = \frac{1}{2}x^\top \mathbf{A} x$ with $\mathbf{A} = \operatorname{diag}\left(\mu, \lambda_2, \ldots, \lambda_{n-1}, L\right)$ with $\alpha = \frac{1}{L}$, we get

$$\begin{aligned} x_1^{k+1} &= \left(1 - \frac{\mu}{L} + \beta\right) x_1^k - \beta x_1^{k-1}, \\ x_2^{k+1} &= \left(1 - \frac{\lambda_2}{L} + \beta\right) x_2^k - \beta x_2^{k-1}, \\ &\vdots \\ x_{n-1}^{k+1} &= \left(1 - \frac{\lambda_{n-1}}{L} + \beta\right) x_{n-1}^k - \beta x_{n-1}^{k-1}, \\ x_n^{k+1} &= \beta x_n^k - \beta x_n^{k-1}. \end{aligned}$$

To solve these recurrences, we consider the corresponding characteristic equations:

$$\begin{aligned} \rho^2 &= \left(1 - \frac{\mu}{L} + \beta\right)\rho - \beta, \\ \rho^2 &= \left(1 - \frac{\lambda_2}{L} + \beta\right)\rho - \beta, \\ &\vdots \\ \rho^2 &= \left(1 - \frac{\lambda_{n-1}}{L} + \beta\right)\rho - \beta, \\ \rho^2 &= \beta\rho - \beta. \end{aligned}$$

Since $\beta \leqslant \left(1 - 2\sqrt{\frac{\mu}{L}}\right)^2 < \left(1 - \sqrt{\frac{\mu}{L}}\right)^2$ the roots of the first equation are

$$\begin{aligned} \rho_1(\mu) &= \frac{1 + \beta - \frac{\mu}{L} + \sqrt{\left(1 + \beta - \frac{\mu}{L}\right)^2 - 4\beta}}{2}, \\ \rho_2(\mu) &= \frac{1 + \beta - \frac{\mu}{L} - \sqrt{\left(1 + \beta - \frac{\mu}{L}\right)^2 - 4\beta}}{2}. \end{aligned}$$

Moreover, we have $\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta} \leqslant 1-\beta+\frac{\mu}{L}$, and, as a consequence, $0 < \rho_2(\mu) < \rho_1(\mu) < 1$. Next, the first components of iterates produced by HB satisfy

$$x_1^k = C_1\rho_1^k(\mu) + C_2\rho_2^k(\mu)$$

with some constants $C_1, C_2 \in \mathbb{R}$. This equation and the choice of the starting points $x^0 = x^1 = (1, 1, \ldots, 1)^\top$ imply

$$\begin{cases} C_1 + C_2 = 1, \\ C_1\rho_1(\mu) + C_2\rho_2(\mu) = 1, \end{cases}$$

whence

$$C_1 = \frac{1-\rho_2(\mu)}{\rho_1(\mu)-\rho_2(\mu)}, \quad C_2 = 1 - C_1 = \frac{\rho_1(\mu)-1}{\rho_1(\mu)-\rho_2(\mu)}.$$

Using the formula for C_1 and $\beta \in \left[\left(1-3\sqrt{\frac{\mu}{L}}\right)^2, \left(1-2\sqrt{\frac{\mu}{L}}\right)^2\right]$ we derive that $C_1 > 0$ and

$$\begin{aligned} C_1 &= \left(1 - \frac{1+\beta-\frac{\mu}{L}-\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta}}{2}\right)\frac{1}{\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta}} = \\ &= \frac{1-\beta+\frac{\mu}{L}+\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta}}{2\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta}} = \frac{1}{2} + \frac{1-\beta+\frac{\mu}{L}}{2\sqrt{\left(1+\beta-\frac{\mu}{L}\right)^2-4\beta}} \leqslant \\ &\leqslant \frac{1}{2} + \frac{1-\left(1-3\sqrt{\frac{\mu}{L}}\right)^2+\frac{\mu}{L}}{2\sqrt{\left(1+\left(1-2\sqrt{\frac{\mu}{L}}\right)^2-\frac{\mu}{L}\right)^2-4\left(1-2\sqrt{\frac{\mu}{L}}\right)^2}} = \\ &= \frac{1}{2} + \frac{3\sqrt{\frac{\mu}{L}}-4\frac{\mu}{L}}{\sqrt{\left(2-4\sqrt{\frac{\mu}{L}}+3\frac{\mu}{L}\right)^2-\left(2-4\sqrt{\frac{\mu}{L}}\right)^2}} = \frac{1}{2} + \frac{3\sqrt{\frac{\mu}{L}}-4\frac{\mu}{L}}{\sqrt{3\frac{\mu}{L}\left(4-8\sqrt{\frac{\mu}{L}}+3\frac{\mu}{L}\right)}}. \end{aligned}$$

Since $L \geqslant 100\mu$ we can further upper bound the right-hand side of the previous inequality and get

$$C_1 \leqslant \frac{1}{2} + \frac{3\sqrt{\frac{\mu}{L}}}{\sqrt{3\frac{\mu}{L}\left(4-8\sqrt{\frac{\mu}{L}}\right)}} \leqslant \frac{1}{2} + \frac{\sqrt{3}}{\sqrt{4-\frac{4}{5}}} = \frac{1}{2} + \frac{1}{2}\sqrt{\frac{15}{4}} \leqslant \frac{3}{2}.$$

Taking into account that $C_1 > 0$ and $C_2 = 1 - C_1$ we derive that $|C_2| = \max\{1-C_1, C_1-1\} \leqslant \frac{1}{2}$. Putting all together, we obtain

$$|x_1^k| = |C_1\rho_1^k(\mu) + C_2\rho_2^k(\mu)| \leqslant |C_1| + |C_2| \leqslant 2 \quad \forall k \geqslant 0.$$

In the remaining part of the proof, we handle the characteristic equations

$$\begin{aligned} \rho^2 &= \left(1 - \frac{\lambda_2}{L} + \beta\right)\rho - \beta, \\ &\vdots \\ \rho^2 &= \left(1 - \frac{\lambda_{n-1}}{L} + \beta\right)\rho - \beta, \\ \rho^2 &= \beta\rho - \beta. \end{aligned}$$

Without loss of generality, we consider the equation

$$\rho^2 = \left(1 - \frac{\lambda}{L} + \beta\right)\rho - \beta \tag{36}$$

with $\lambda \in [\lambda_2, L]$. This equation serves as a characteristic equation for the sequence $\{y_k\}_{k\geqslant 0} \subseteq \mathbb{R}$ satisfying

$$y_{k+1} = \left(1 - \frac{\lambda}{L} + \beta\right)y_k - \beta y_{k-1}.$$

Since $\lambda \geqslant \lambda_2 \geqslant 10\mu$ and $\beta \geqslant \left(1 - 3\sqrt{\frac{\mu}{L}}\right)^2$, we conclude that $\beta \geqslant \left(1 - \sqrt{\frac{\lambda}{L}}\right)^2$ and the characteristic equation has the complex roots with nonzero imaginary parts:

$$\rho_1(\lambda) = \frac{1 + \beta - \frac{\lambda}{L} + i\sqrt{4\beta - \left(1 + \beta - \frac{\mu}{L}\right)^2}}{2},$$
$$\rho_2(\lambda) = \frac{1 + \beta - \frac{\mu}{L} - i\sqrt{4\beta - \left(1 + \beta - \frac{\mu}{L}\right)^2}}{2}.$$

This implies that $|\rho_1(\lambda)| = |\rho_1(\lambda)| = \sqrt{\beta} < 1$ and

$$y_k = C_1\rho_1^k(\lambda) + C_2\rho_2^k(\lambda)$$

for some complex numbers C_1, C_2. Let $y_0 = y_1 = 1$. Then,

$$\begin{cases} C_1 + C_2 = 1, \\ C_1\rho_1(\lambda) + C_2\rho_2(\lambda) = 1, \end{cases}$$

whence

$$C_1 = \frac{1 - \rho_2(\lambda)}{\rho_1(\lambda) - \rho_2(\lambda)}, \quad C_2 = 1 - C_1 = \frac{\rho_1(\lambda) - 1}{\rho_1(\lambda) - \rho_2(\lambda)}.$$

Using the formula for C_1 and $\beta \in \left[\left(1 - 3\sqrt{\frac{\mu}{L}}\right)^2, \left(1 - 2\sqrt{\frac{\mu}{L}}\right)^2\right]$, we derive

$$C_1 = \left(1 - \frac{1 + \beta - \frac{\lambda}{L} - i\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}}{2}\right)\frac{1}{i\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}} =$$
$$= \frac{1 - \beta + \frac{\lambda}{L} + i\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}}{2i\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}} = \frac{1}{2} - i\frac{1 - \beta + \frac{\lambda}{L}}{2\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}}.$$

Then, for the absolute value of C_1 we have

$$|C_1| = \frac{1}{2}\sqrt{1 + \frac{(1 + \frac{\lambda}{L} - \beta)^2}{4\beta - (1 - \frac{\lambda}{L} + \beta)^2}} \leqslant \frac{1}{2}\sqrt{1 + \frac{\left(1 + \frac{\lambda}{L} - \left(1 - 3\sqrt{\frac{\mu}{L}}\right)^2\right)^2}{4\left(1 - 2\sqrt{\frac{\mu}{L}}\right)^2 - \left(1 - \frac{\lambda}{L} + \left(1 - 2\sqrt{\frac{\mu}{L}}\right)^2\right)^2}} =$$

$$= \frac{1}{2}\sqrt{1 + \frac{\left(\frac{\lambda}{L} - 9\frac{\mu}{L} + 6\sqrt{\frac{\mu}{L}}\right)^2}{\left(2 - 4\sqrt{\frac{\mu}{L}}\right)^2 - \left(2 - \frac{\lambda}{L} - 4\sqrt{\frac{\mu}{L}} + 4\frac{\mu}{L}\right)^2}} = \frac{1}{2}\sqrt{1 + \frac{\left(\frac{\lambda}{L} - 9\frac{\mu}{L} + 6\sqrt{\frac{\mu}{L}}\right)^2}{\left(4 - \frac{\lambda}{L} - 8\sqrt{\frac{\mu}{L}} + 4\frac{\mu}{L}\right)\left(\frac{\lambda}{L} - 4\frac{\mu}{L}\right)}} \leqslant$$

$$\leqslant \frac{1}{2}\sqrt{1 + \frac{\left(\frac{\lambda}{L} + 6\sqrt{\frac{\mu}{L}}\right)^2}{\left(3 - \frac{4}{5}\right)\left(\frac{\lambda}{L} - \frac{2\lambda}{5L}\right)}} = \frac{1}{2}\sqrt{1 + \frac{25}{33}\frac{\frac{\lambda^2}{L^2} + 12\frac{(\lambda\sqrt{\mu})}{(L\sqrt{L})} + 36\frac{\mu}{L}}{\frac{\lambda}{L}}} =$$

$$= \frac{1}{2}\sqrt{1 + \frac{25}{33}\left(\frac{\lambda}{L} + 12\sqrt{\frac{\mu}{L}} + 36\frac{\mu}{\lambda}\right)} \leqslant \frac{1}{2}\sqrt{1 + \frac{25}{33}\left(1 + \frac{6}{5} + \frac{8}{25}\right)} \leqslant 1.$$

Since

$$C_2 = 1 - C_1 = \frac{1}{2} + i\frac{1 - \beta + \frac{\lambda}{L}}{2\sqrt{4\beta - \left(1 + \beta - \frac{\lambda}{L}\right)^2}}$$

we also have $|C_2| = |C_1| \leqslant 1$, and, as a consequence,

$$|y_k| = |C_1\rho_1^k(\lambda) + C_2\rho_2^k(\lambda)| \leqslant |C_1| + |C_2| \leqslant 2 \quad \forall k \geqslant 0.$$

This result implies that $|x_i^k| \leqslant 2$ for all $k \geqslant 0$ and $i = 2, \ldots, n$.

Finally, since $\overline{x}^k = \frac{1}{k+1}\sum\limits_{t=0}^{k} x^t$ we conclude that

$$|\overline{x}_i^k| \leqslant \frac{1}{k+1}\sum_{t=0}^{k}|x_i^t| \leqslant 2 \quad \forall k \geqslant 0,\ i = 1, \ldots, n,$$

which is equivalent to (6). □

Proof of Theorem 3

To estimate $\mathrm{dev}_{\mathsf{HB}}(\alpha, \beta)$ we consider the spectral decomposition of matrix $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top \succ 0$, where $\mathbf{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ is a diagonal matrix of the eigenvalues of $\mathbf{A}$, $\lambda_1 \leqslant \ldots \leqslant \lambda_n$, and U is a unitary matrix of the eigenvectors of $\mathbf{A}$. Next, without loss of the generality we assume that $x^* = 0$. Applying the unitary transformation $\mathbf{U}^\top$ to x^k, we obtain $\widehat{x}^k = \mathbf{U}^\top x^k$ and

$$\widehat{z}^k := \begin{bmatrix}\widehat{x}^{k+1}\\ \widehat{x}^k\end{bmatrix} = \widehat{\mathbf{T}}\begin{bmatrix}\widehat{x}^k\\ \widehat{x}^{k-1}\end{bmatrix} = \ldots = \widehat{\mathbf{T}}^k\begin{bmatrix}\widehat{x}^1\\ \widehat{x}^0\end{bmatrix},$$

where

$$\widehat{\mathbf{T}} = \left[\begin{array}{c|c}(1+\beta)\mathbf{I} - \alpha\mathbf{\Lambda} & -\beta\mathbf{I}\\ \hline \mathbf{I} & \mathbf{0}\end{array}\right] = \left[\begin{array}{c|c}\mathbf{U}^\top & \mathbf{0}\\ \hline \mathbf{0} & \mathbf{U}^\top\end{array}\right]\mathbf{T}.$$

In particular, these formulas imply

$$\begin{bmatrix} \widehat{x}_j^{k+1} \\ \widehat{x}_j^{k} \end{bmatrix} = \widehat{\mathbf{T}}_j \begin{bmatrix} \widehat{x}_j^{k} \\ \widehat{x}_j^{k-1} \end{bmatrix} = \ldots = \widehat{\mathbf{T}}_j^k \begin{bmatrix} \widehat{x}_j^{1} \\ \widehat{x}_j^{0} \end{bmatrix},$$

where

$$\widehat{\mathbf{T}}_j = \begin{bmatrix} 1+\beta-\alpha\lambda_j & -\beta \\ 1 & 0 \end{bmatrix}$$

for all $j = 1, \ldots, n$. Moreover, $\|\mathbf{C}\mathbf{T}^k\|_2 = \max\limits_{j=1,\ldots,n} \|\mathbf{C}_j \widehat{\mathbf{T}}_j^k\|_2$, where $\mathbf{C}_j = [\,0\ 1\,]$.

It is easy to see that the eigenvalues of $\widehat{\mathbf{T}}_j$ are

$$\rho_{j,1} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} + \sqrt{\left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2 - 4\beta}}{2}, \quad \rho_{j,2} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} - \sqrt{\left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2 - 4\beta}}{2}$$

for all λ_j such that $\left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2 - 4\beta > 0$ and

$$\rho_{j,1} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} + i\sqrt{4\beta - \left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2}}{2}, \quad \rho_{j,2} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} - i\sqrt{4\beta - \left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2}}{2}$$

for all λ_j such that $\left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2 - 4\beta < 0$. Taking into account the assumptions of the theorem, we derive

$$\rho_{1,1} = \frac{1+\beta-\frac{\lambda_1}{\lambda_n} + \sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2 - 4\beta}}{2}, \quad \rho_{1,2} = \frac{1+\beta-\frac{\lambda_1}{\lambda_n} - \sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2 - 4\beta}}{2}$$

and

$$\rho_{j,1} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} + i\sqrt{4\beta - \left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2}}{2}, \quad \rho_{j,2} = \frac{1+\beta-\frac{\lambda_j}{\lambda_n} - i\sqrt{4\beta - \left(1+\beta-\frac{\lambda_j}{\lambda_n}\right)^2}}{2}$$

for all $j = 2, \ldots, n$. Moreover, $|\rho_{j,1}| = |\rho_{j,2}| = \sqrt{\beta}$.

Next, using Lemma 3 we get

$$\begin{aligned} \|\mathbf{C}_j \widehat{\mathbf{T}}_j^k\|_2 &= \left\| \frac{1}{\rho_{j,2}-\rho_{j,1}} \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} \rho_{j,2}^{k+1} - \rho_{j,1}^{k+1} & \rho_{j,1}\rho_{j,2}(\rho_{j,1}^{k} - \rho_{j,2}^{k}) \\ \rho_{j,2}^{k} - \rho_{j,1}^{k} & \rho_{j,1}\rho_{j,2}(\rho_{j,1}^{k-1} - \rho_{j,2}^{k-1}) \end{bmatrix} \right\|_2 = \\ &= \sqrt{\left| \frac{\rho_{j,2}^k - \rho_{j,1}^k}{\rho_{j,2}-\rho_{j,1}} \right|^2 + \left| \frac{\rho_{j,1}\rho_{j,2}\left(\rho_{j,2}^{k-1} - \rho_{j,1}^{k-1}\right)}{\rho_{j,2}-\rho_{j,1}} \right|^2} \leqslant \\ &\leqslant \sqrt{\left(\sum_{t=0}^{k-1} |\rho_{j,1}|^{k-1-t} |\rho_{j,2}|^t \right)^2 + \left(|\rho_{j,1}||\rho_{j,2}| \sum_{t=0}^{k-2} |\rho_{j,1}|^{k-2-t} |\rho_{j,2}|^t \right)^2}. \end{aligned} \quad (37)$$

Consider the expression above for $j = 1$. To bound the sums appearing on the right-hand side of the previous inequality, we derive:

$$\frac{|\rho_{1,2}|}{|\rho_{1,1}|} = \frac{1+\beta-\frac{\lambda_1}{\lambda_n}-\sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2-4\beta}}{1+\beta-\frac{\lambda_1}{\lambda_n}+\sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2-4\beta}} = 1-\frac{2\sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2-4\beta}}{1+\beta-\frac{\lambda_1}{\lambda_n}+\sqrt{\left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2-4\beta}} \leqslant$$

$$\leqslant 1-\frac{2\sqrt{\left(1+\left(1-F\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2-\frac{\lambda_1}{\lambda_n}\right)^2-4\left(1-F\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2}}{2-\frac{\lambda_1}{\lambda_n}+\sqrt{\left(1-\frac{\lambda_1}{\lambda_n}\right)^2}} =$$

$$= 1-\frac{2\sqrt{\left(2-2F\sqrt{\frac{\lambda_1}{\lambda_n}}+(F^2-1)\frac{\lambda_1}{\lambda_n}\right)^2-4\left(1-F\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2}}{3-2\frac{\lambda_1}{\lambda_n}} \leqslant$$

$$\leqslant 1-\frac{2\sqrt{(F^2-1)\frac{\lambda_1}{\lambda_n}\left(4-4F\sqrt{\frac{\lambda_1}{\lambda_n}}+(F^2-1)\frac{\lambda_1}{\lambda_n}\right)}}{3} =$$

$$= 1-\frac{2\sqrt{(F^2-1)\left(\left(2-F\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2-\frac{\lambda_1}{\lambda_n}\right)}}{3\sqrt{\varkappa}} \leqslant 1-\frac{\sqrt{F^2-1}}{\sqrt{3\varkappa}},$$

where the first inequality follows from the fact the function $g(\beta) = \left(1+\beta-\frac{\lambda_1}{\lambda_n}\right)^2 - 4\beta$ is decreasing for $\beta \leqslant \left(1-\sqrt{\frac{\lambda_1}{\lambda_n}}\right)^2$, and in the last inequality we apply $1 - F\sqrt{\frac{\lambda_1}{\lambda_n}} \geqslant 0$, $\frac{\lambda_1}{\lambda_n} \leqslant \frac{1}{10\,000} < \frac{1}{4}$, and $\varkappa = \frac{\lambda_n}{\lambda_1}$. Therefore,

$$\sum_{t=0}^{k-1} |\rho_{1,1}|^{k-1-t}|\rho_{1,2}|^t = |\rho_{1,1}|^{k-1}\sum_{t=0}^{k-1}\left(\frac{|\rho_{1,2}|}{|\rho_{1,1}|}\right)^t \leqslant \sum_{t=0}^{\infty}\left(1-\frac{\sqrt{F^2-1}}{\sqrt{3\varkappa}}\right)^t = \frac{\sqrt{3\varkappa}}{\sqrt{F^2-1}}$$

and, similarly,

$$|\rho_{j,1}||\rho_{j,2}|\sum_{t=0}^{k-2}|\rho_{j,1}|^{k-2-t}|\rho_{j,2}|^t \leqslant \sum_{t=0}^{k-2}\left(\frac{|\rho_{1,2}|}{|\rho_{1,1}|}\right)^t \leqslant \sum_{t=0}^{\infty}\left(1-\frac{\sqrt{F^2-1}}{\sqrt{3\varkappa}}\right)^t = \frac{\sqrt{3\varkappa}}{\sqrt{F^2-1}}.$$

Plugging these upper bounds in (37), we derive

$$\|\mathbf{C}_j\widehat{\mathbf{T}}_j^k\|_2 \leqslant \frac{\sqrt{6\varkappa}}{\sqrt{F^2-1}}. \tag{38}$$

Next, we consider the right-hand side of (37) for $j = 2, \ldots, n$. In this case, $|\rho_{j,1}| = |\rho_{j,2}| = \sqrt{\beta} \leqslant 1 - \frac{F}{\sqrt{\varkappa}}$. Therefore,

$$\sum_{t=0}^{k-1}|\rho_{j,1}|^{k-1-t}|\rho_{j,2}|^t = k\left(\sqrt{\beta}\right)^{k-1} \leqslant k\left(1-\frac{F}{\sqrt{\varkappa}}\right)^{k-1} \leqslant (k-1)\exp\left(-(k-1)\frac{F}{\sqrt{\varkappa}}\right)+1$$

and, similarly,

$$|\rho_{j,1}||\rho_{j,2}|\sum_{t=0}^{k-2}|\rho_{j,1}|^{k-2-t}|\rho_{j,2}|^t = (k-1)\left(\sqrt{\beta}\right)^k \leqslant (k-1)\exp\left(-(k-1)\frac{F}{\sqrt{\varkappa}}\right).$$

Since the maximal value of the function $g(x) = xa^x$ for $x \geqslant 0$ equals $-\frac{1}{(e\ln(a))}$, we have

$$(k-1)\exp\left(-(k-1)\frac{F}{\sqrt{\varkappa}}\right) \leqslant -\frac{1}{e\ln\left(\exp\left(-\frac{F}{\sqrt{\varkappa}}\right)\right)} = \frac{\sqrt{\varkappa}}{eF}.$$

Putting all together, we obtain for all $j = 2, \ldots, n$

$$\|\mathbf{C}_j\widehat{\mathbf{T}}_j^k\|_2 \overset{(37)}{\leqslant} \sqrt{\left(\frac{\sqrt{\varkappa}}{eF}+1\right)^2 + \frac{\varkappa}{e^2F^2}} \leqslant \frac{\sqrt{5\varkappa}}{eF}, \tag{39}$$

where we use $F \leqslant \sqrt{\varkappa}$.

Finally, with (38) and (39) in hand we derive

$$\mathrm{dev}_{\mathsf{AHB}}(\alpha,\beta) \leqslant \mathrm{dev}_{\mathsf{HB}}(\alpha,\beta) = \|\mathbf{C}\mathbf{T}^k\|_2 = \max_{j=1,\ldots,n}\|\mathbf{C}_j\widehat{\mathbf{T}}_j^k\|_2 \leqslant \frac{\sqrt{6\varkappa}}{\sqrt{F^2-1}}.$$

Theorem 1 from [Danilova, Kulakova, Polyak, 2018] implies that

$$\mathrm{dev}_{\mathsf{HB}}(\alpha^*,\beta^*) \geqslant \frac{\sqrt{\varkappa}}{2e},$$

where α^* and β^* are given in (4). Therefore,

$$\mathrm{dev}_{\mathsf{AHB}}(\alpha,\beta) \leqslant \mathrm{dev}_{\mathsf{HB}}(\alpha,\beta) \leqslant \frac{2e\sqrt{6}}{\sqrt{F^2-1}}\,\mathrm{dev}_{\mathsf{HB}}(\alpha^*,\beta^*).$$

□

Missing proofs from Section 3

Proof of Lemma 1

Using recursion (13) for the virtual iterates defined in (12), we derive

$$\begin{aligned}\|\widetilde{x}_{k+1}-x_*\|^2 &= \|\widetilde{x}_k-x_*\|^2 - \frac{2\alpha}{1-\beta}\langle\widetilde{x}_k-x_*,\,\nabla f(x_k)\rangle + \frac{\alpha^2}{(1-\beta)^2}\|\nabla f(x_k)\|^2 = \\ &= \|\widetilde{x}_k-x_*\|^2 - \frac{2\alpha}{1-\beta}\langle x_k-x_*,\,\nabla f(x_k)\rangle + \frac{2\alpha}{1-\beta}\langle x_k-\widetilde{x}_k,\,\nabla f(x_k)\rangle + \frac{\alpha^2}{(1-\beta)^2}\|\nabla f(x_k)\|^2.\end{aligned} \tag{40}$$

From μ-strong convexity and L-smoothness of f we have (e. g., see [Nesterov, 2018])

$$\begin{gathered}\langle x_k-x_*,\,\nabla f(x_k)\rangle \geqslant f(x_k)-f(x_*)+\frac{\mu}{2}\|x_k-x_*\|^2,\\ \|\nabla f(x_k)\|^2 \leqslant 2L\left(f(x_k)-f(x_*)\right).\end{gathered} \tag{41}$$

Together with (40) these relations give

$$\|\widetilde{x}_{k+1} - x_*\|^2 \leqslant \|\widetilde{x}_k - x_*\|^2 - \frac{\alpha\mu}{1-\beta}\|x_k - x_*\|^2 - \frac{2\alpha}{1-\beta}\left(1 - \frac{\alpha L}{1-\beta}\right)\Big(f(x_k) - f(x_*)\Big) + \\ + \frac{2\alpha}{1-\beta}\langle x_k - \widetilde{x}_k,\, \nabla f(x_k)\rangle.$$

Next, we estimate the second and the fourth terms in the inequality above. Since $\|a+b\|^2 \geqslant \frac{1}{2}\|a\|^2 - \|b\|^2$ for all $a,\, b \in \mathbb{R}^n$ (see also (31)), we can estimate the second term as

$$-\frac{\alpha\mu}{1-\beta}\|x_k - x_*\|^2 \leqslant -\frac{\alpha\mu}{2(1-\beta)}\|\widetilde{x}_k - x_*\|^2 + \frac{\alpha\mu}{1-\beta}\|x_k - \widetilde{x}_k\|^2.$$

Using the Fenchel – Young inequality (30), we derive

$$\frac{2\alpha}{1-\beta}\langle x_k - \widetilde{x}_k,\, \nabla f(x_k)\rangle \leqslant \frac{2\alpha L}{1-\beta}\|x_k - \widetilde{x}_k\|^2 + \frac{2\alpha}{4L(1-\beta)}\|\nabla f(x_k)\|^2 \overset{(41)}{\leqslant} \\ \overset{(41)}{\leqslant} \frac{2\alpha L}{1-\beta}\|x_k - \widetilde{x}_k\|^2 + \frac{\alpha}{1-\beta}\Big(f(x_k) - f(x_*)\Big).$$

Putting all together, we obtain

$$\|\widetilde{x}_{k+1} - x_*\|^2 \leqslant \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)\|\widetilde{x}_k - x_*\|^2 - \frac{2\alpha}{1-\beta}\left(\frac{1}{2} - \frac{\alpha L}{1-\beta}\right)\Big(f(x_k) - f(x_*)\Big) + \\ + \frac{\alpha}{1-\beta}(2L+\mu)\|x_k - \widetilde{x}_k\|^2 \overset{(12),(14)}{\leqslant} \\ \overset{(12),(14)}{\leqslant} \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)\|\widetilde{x}_k - x_*\|^2 - \frac{\alpha}{1-\beta}\Big(f(x_k) - f(x_*)\Big) + \frac{3L\alpha\beta^2}{(1-\beta)^3}\|m_{k-1}\|^2.$$

This completes the proof. □

Proof of Lemma 2

Using the update rule for m_k, we get

$$\|m_k\|^2 = \|\beta m_{k-1} + \alpha\nabla f(x_k)\|^2 \overset{(32)}{\leqslant} \beta^2\left(1 + \frac{1-\beta}{\beta}\right)\|m_{k-1}\|^2 + \alpha^2\left(1 + \frac{\beta}{1-\beta}\right)\|\nabla f(x_k)\|^2 \overset{(41)}{\leqslant} \\ \overset{(41)}{\leqslant} \beta\|m_{k-1}\|^2 + \frac{2L\alpha^2}{1-\beta}\Big(f(x_k) - f(x_*)\Big),$$

implying

$$\|m_{k-1}\|^2 \leqslant \frac{2L\alpha^2}{1-\beta}\sum_{l=0}^{k-1}\beta^{k-1-l}\,(f(x_l) - f(x_*)).$$

Summing up these inequalities for $k = 0,\, 1,\, \ldots,\, K$ with weights $w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k+1)}$, we derive

$$\frac{3L\alpha\beta^2}{(1-\beta)^3}\sum_{k=0}^{K} w_k\|m_{k-1}\|^2 \leqslant \frac{6L^2\alpha^3\beta^2}{(1-\beta)^4}\sum_{k=0}^{K}\sum_{l=0}^{k-1} w_k\Big(f(x_l) - f(x_*)\Big)\beta^{k-1-l} \leqslant \\ \leqslant \frac{6L^2\alpha^3\beta}{(1-\beta)^4}\sum_{k=0}^{K}\sum_{l=0}^{k} w_k\Big(f(x_l) - f(x_*)\Big)\beta^{k-l}. \quad (42)$$

Next, we upper bound w_k in the following way: for all $l = 0, 1, \ldots, k$

$$w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k-l)} w_l \overset{(34)}{\leqslant} \left(1 + \frac{\alpha\mu}{1-\beta}\right)^{k-l} w_l \overset{(16)}{\leqslant} \left(1 + \frac{1-\beta}{2}\right)^{k-l} w_l.$$

Plugging this inequality into (42), we get

$$\frac{3L\alpha\beta^2}{(1-\beta)^3}\sum_{k=0}^{K} w_k\|m_{k-1}\|^2 \leqslant \frac{6L^2\alpha^3\beta}{(1-\beta)^4}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l\Big(f(x_l) - f(x_*)\Big)\left(1 + \frac{1-\beta}{2}\right)^{k-l}\beta^{k-l} \overset{(34)}{\leqslant}$$

$$\overset{(34)}{\leqslant} \frac{6L^2\alpha^3\beta}{(1-\beta)^4}\sum_{k=0}^{K}\sum_{l=0}^{k} w_l\Big(f(x_l) - f(x_*)\Big)\left(1 - \frac{1-\beta}{2}\right)^{k-l} \leqslant$$

$$\leqslant \frac{6L^2\alpha^3\beta}{(1-\beta)^4}\left(\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big)\right)\left(\sum_{k=0}^{\infty}\left(1 - \frac{1-\beta}{2}\right)^k\right) = \frac{12L^2\alpha^3\beta}{(1-\beta)^5}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big).$$

Note that our choice of α (16) implies

$$\frac{12L^2\alpha^3\beta}{(1-\beta)^5} \leqslant \frac{\alpha}{4(1-\beta)}.$$

Together with the previous inequality it gives (17). □

Proof of Theorem 4

From Lemma 1 we have

$$\frac{\alpha}{2(1-\beta)}\Big(f(x_k) - f(x_*)\Big) \leqslant \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)\|\widetilde{x}_k - x_*\|_2^2 - \|\widetilde{x}_{k+1} - x_*\|_2^2 + \frac{3L\alpha\beta^2}{(1-\beta)^3}\|m_{k-1}\|_2^2.$$

Summing up these inequalities for $k = 0, 1, \ldots, K$ with weights $w_k = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-(k+1)}$, we get

$$\frac{\alpha}{2(1-\beta)}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big) \leqslant$$

$$\leqslant \sum_{k=0}^{K}\left(w_k\left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)\|\widetilde{x}_k - x_*\|_2^2 - w_k\|\widetilde{x}_{k+1} - x_*\|_2^2\right) + \frac{3L\alpha\beta^2}{(1-\beta)^3}\sum_{k=0}^{K} w_k\|m_{k-1}\|_2^2 \overset{(17)}{\leqslant}$$

$$\overset{(17)}{\leqslant} \sum_{k=0}^{K}\Big(w_{k-1}\|\widetilde{x}_k - x_*\|_2^2 - w_k\|\widetilde{x}_{k+1} - x_*\|_2^2\Big) + \frac{\alpha}{4(1-\beta)}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big) \leqslant$$

$$\leqslant \|x_0 - x_*\|_2^2 + \frac{\alpha}{4(1-\beta)}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big).$$

Rearranging the terms and dividing both sides of the inequality by $W_K = \sum\limits_{k=0}^{K} w_k$, we derive

$$\frac{1}{W_K}\sum_{k=0}^{K} w_k\Big(f(x_k) - f(x_*)\Big) \leqslant \frac{4(1-\beta)\|x_0 - x_*\|_2^2}{\alpha W_K}.$$

Using Jensen's inequality, we obtain

$$f(\overline{x}_K) \leqslant \frac{1}{W_K}\sum_{k=0}^{K} w_k f(x_k),$$

which implies (19). Next, when $\mu > 0$ we have $W_K \geqslant w_{K-1} = \left(1 - \frac{\alpha\mu}{2(1-\beta)}\right)^{-K}$, which gives (20). Finally, when $\mu = 0$ we have $W_K = K + 1 > K$, implying (21).

Proof of Theorem 5

Theorem 4 for $\mu = 0$ implies that for $t = 1, 2, \ldots, \tau$

$$f(\widehat{x}_t) - f(x_*) \leqslant \frac{4(1-\beta)\widehat{R}_{t-1}^2}{\alpha N}, \tag{43}$$

where $\widehat{R}_t = \|\widehat{x}_t - x_*\|_2$ for $t = 0, 1, \ldots, \tau$. In the remaining part of the proof, we derive via induction that for $t = 1, 2, \ldots, \tau$

$$f(\widehat{x}_t) - f(x_*) \leqslant \frac{\mu R_0^2}{2^{t+1}}, \quad \widehat{R}_t \leqslant \frac{R_0^2}{2^t}, \tag{44}$$

where $R_0 \geqslant \|x_0 - x_*\|_2 = \|\widehat{x}_0 - x_*\|_2$. First of all, for $t = 1$ we have

$$f(\widehat{x}_1) - f(x_*) \overset{(24),(43)}{\leqslant} \frac{\mu R_0^2}{4}.$$

From μ-strong convexity of f we derive

$$\frac{\mu\widehat{R}_1^2}{2} \leqslant f(\widehat{x}_1) - f(x_*) \quad \Longrightarrow \quad \widehat{R}_1^2 \leqslant \frac{R_0^2}{2}.$$

Next, assume that (44) holds for all $t = 1, 2, \ldots, k < \tau$ and prove it for $t = k + 1$. From (43) we have

$$f(\widehat{x}_{k+1}) - f(x_*) \leqslant \frac{4(1-\beta)\widehat{R}_k^2}{\alpha N} \overset{(43)}{\leqslant} \frac{(1-\beta)R_0^2}{2^{k-2}\alpha N} \overset{(24)}{\leqslant} \frac{\mu R_0^2}{2^{k+2}}.$$

Again, applying μ-strong convexity of f we derive

$$\frac{\mu\widehat{R}_{k+1}^2}{2} \leqslant f(\widehat{x}_{k+1}) - f(x_*) \quad \Longrightarrow \quad \widehat{R}_{k+1}^2 \leqslant \frac{R_0^2}{2^{k+1}},$$

which finishes the proof of (44). Therefore, after $\tau = \max\left\{\left\lceil \log_2\left(\frac{\mu R_0^2}{\varepsilon}\right)\right\rceil - 1,\ 1\right\}$ iterations R-AHB finds such a point $\widehat{x}_\tau$ that

$$f(\widehat{x}_\tau) - f(x_*) \leqslant \frac{\mu R_0^2}{2^{\tau+1}} \leqslant \frac{\mu R_0^2}{2^{\log_2(\mu R_0^2/\varepsilon)}} = \varepsilon.$$

Finally, if

$$\alpha = \min\left\{\frac{1-\beta}{4L},\ \frac{(1-\beta)^2}{4L\sqrt{3\beta}}\right\},$$

then the total number of AHB iterations equals

$$N\tau = O\left(\left(\frac{L}{\mu} + \frac{L\sqrt{\beta}}{\mu(1-\beta)}\right)\log\frac{\mu R_0^2}{\varepsilon}\right).$$

□