On the Convergence Analysis of Aggregated Heavy-Ball Method*

Marina Danilova¹,²

¹Institute of Control Sciences of Russian Academy of Sciences, Russia
²Moscow Institute of Physics and Technology, Russia

Abstract

Momentum first-order optimization methods are the workhorses in various optimization tasks, e.g., in the training of deep neural networks. Recently, Lucas et al. (2019) [7] proposed a method called Aggregated Heavy-Ball (AggHB) that uses multiple momentum vectors corresponding to different momentum parameters and averages these vectors to compute the update direction at each iteration. Lucas et al. (2019) [7] show that AggHB is more stable than the classical Heavy-Ball method even with large momentum parameters and performs well in practice. However, the method was analyzed only for quadratic objectives and for online optimization tasks under uniformly bounded gradients assumption, which is not satisfied for many practically important problems. In this work, we address this issue and propose the first analysis of AggHB for smooth objective functions in non-convex, convex, and strongly convex cases without additional restrictive assumptions. Our complexity results match the best-known ones for the Heavy-Ball method. We also illustrate the efficiency of AggHB numerically on several non-convex and convex problems.

1 Introduction

Momentum [14] and acceleration [10] are popular techniques for speeding up first-order optimization methods both from practical and theoretical perspectives. Historically, one of the first examples of such methods is Heavy-Ball (HB) method proposed by B. Polyak in 1964 [14]. This method received a lot of attention from various research communities due to its efficiency in different convex and, more importantly, non-convex problems [2]. In particular, during the last few years a lot of variants of HB were proposed and analyzed by machine learning (ML) researchers, especially due to its efficiency in computer vision tasks [16].

Recently, another modification of HB called Aggregated Heavy-Ball (AggHB) method was proposed in [7]. In contrast to HB, AggHB has $m \geq 1$ different momentum parameters and $m$ corresponding momentum vectors. An average of these vectors is used as an update direction at each iteration. Such an averaging helps to make the method more stable via reducing the oscillations of the iterates, as the authors of [7] illustrated empirically. Moreover, the numerical results from [7] show the superiority of AggHB to HB at training several ML models.

*The research was supported by Russian Foundation for Basic Research (Theorem 2.1, project No. 20-31-90073) and by Russian Science Foundation (Theorem 2.5, project No. 21-71-30005).
1.1 Motivational Example

In this section, we consider the behavior of AggHB on Rosenbrock function, which is well-known non-convex test functions. The set of momentum parameters for AggHB were chosen as \([0.9, 0.95, 0.99, 0.999]\) (see Algorithm 2) and for HB a standard momentum parameter \(\beta = 0.95\) was taken (see Algorithm 1). Stepsize \(\gamma\) was tuned for each method. The results are presented in Figure 1. We observe much smaller oscillations for AggHB than for HB. Moreover, the trajectory of AggHB achieves better accuracy. This example motivates the detailed study of AggHB and, in particular, the theoretical study of its convergence.

1.2 Our Contributions

However, a little is known about theoretical convergence guarantees for AggHB. In particular, the authors of [7] analyzed AggHB for quadratic optimization problems, which is a very small class of problems, and for convex online optimization problems such that the gradients of the objective function are bounded on the whole domain. The former assumption is not satisfied for many practically important tasks. In this paper, we remove this limitation and derive new convergence results for AggHB for smooth non-convex and (strongly) convex problems.

Our main contributions can be summarized as follows.

- **First analysis of AggHB for non-convex problems.** For the problems with smooth but not necessary convex objective function \(f\), we prove that AggHB finds an \(\varepsilon\)-stationary point (point \(x\) such that \(\|\nabla f(x)\| \leq \varepsilon\)) after \(O(1/\varepsilon^2)\) iterations neglecting the dependence on momentum parameters, smoothness constant, and initial functional suboptimality. When \(m = 1\) we recover the complexity of HB and when \(m > 1\) our rate is better than the corresponding rate of HB with maximal momentum parameter (see Theorem 2.1 and Corollary 2.2 for the details).
Algorithm 1 Heavy-Ball method (HB)

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

1. for $k = 0, \ldots, N - 1$ do
   2. $V_k = \beta V_{k-1} + \nabla f(x_k)$
   3. $x_{k+1} = x_k - \gamma V_k$
4. end for

**Output:** $x_N$

Глава 1.3 Технические предпосылки

Мы рассмотрим неконstrained минимизационную проблему

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

где функция $f: \mathbb{R}^n \to \mathbb{R}$ является $L$-гладкой, т.е., для всех $x, y \in \mathbb{R}^n$

$$\|\nabla f(x) - \nabla f(y)\|_2 \leq L\|x - y\|_2.$$ (1)

Далее, мы предполагаем, что $f(x)$ является нижним ограниченным $f_{\inf} = \inf_{x \in \mathbb{R}^n} f(x) > -\infty$ или $\mu$-доменным

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

Формулы, которые мы используем, стандартны для оптимизационной литературы [13, 11], например, $x_*$ обозначает решение (1), а расстояние от начальной точки до решения обозначено как $R_0 = \|x_0 - x_*\|_2$.

Глава 1.4 Релевантная работа

Теоретические гарантии сходимости для HB. Первое сходимость анализом Heavy-Ball метод (HB, Алгоритм 1) был сделан в оригинальной работе B. Polyak в 1964 [14], где локальная $O(\sqrt{L/\mu \log(1/\varepsilon))}$ сходимость скорость была показана для дважды непрерывно дифференцируемой $L$-гладкой и $\mu$-доменной функций. После 50 лет работы Ghadimi et al. (2015) [5] получил первый глобальный $O(L\mu \sqrt{\log(1/\varepsilon))}$ и $O(LR^2\mu \sqrt{\log(1/\varepsilon))}$ сложности границы для L-гладкой $\mu$-доменной и выпуклой функций соответственно. В контраст к локальным сходимостным гарантиям, эти скорости не ускорялись [9, 10]. Хотя один может улучшить анализ HB для квадратичных функций и получить асимптотически ускоренные скорость [6], это по-прежнему неясно, может ли этот результат быть распространен до общей не-квадратичной функции.
The non-triviality of this question is supported by the negative result from [17] showing that one cannot derive accelerated rate of HB for the standard choice of parameters using quadratic potentials in the analysis.

**HB with aggregation and averaging.** As we we already mentioned, Aggregated Heavy-Ball method (AggHB, Algorithm 2) was proposed in [7], where authors empirically shown that aggregation helps to stabilize the methods behavior, speeds up the method in practice, and they also derive some convergence guarantees under uniformly bounded gradients assumption in the stochastic case. Recently, in [3], another approach for stabilizing HB was considered. In particular, the authors of [3] considered several averaging techniques for HB and shown that they help to reduce the maximal deviation of the method and improve the performance of the method in practice.

### 2 Analysis of Aggregated Heavy-Ball Method

In this section we propose a new convergence analysis for Aggregated Heavy-Ball method (AggHB, Algorithm 2). The key difference between HB and AggHB is that instead of one direction determined by parameter $\beta$ the method uses to the vector of, o,entum parameters $\beta = [\beta_1, \ldots, \beta_m]$ and takes and average over $m$ corresponding directions. When $m = 1$ AggHB recovers HB. Moreover, we consider a slight generalization of the method proposed in [7], since we allow to use different stepsizes for different momentum parameters.

Following [8, 18] we consider perturbed/virtual iterates:

$$\tilde{x}_k = x_k - \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} V^{(i)}_{k-1}, \quad k \geq 0. \quad (4)$$

This representation is used for the analysis only and there is no need to compute this sequence when running the method. Virtual iterates satisfy the following useful recursion: for all $k \geq 0$

$$\tilde{x}_{k+1} = x_{k+1} - \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} V^{(i)}_k = x_k - \frac{1}{m} \sum_{i=1}^{m} \gamma_i V^{(i)}_k - \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} V^{(i)}_k$$

$$\quad = x_k - \frac{1}{m} \sum_{i=1}^{m} \frac{\gamma_i V^{(i)}_k}{1 - \beta_i} = x_k - \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} V^{(i)}_{k-1} - \frac{1}{m} \sum_{i=1}^{m} \frac{\gamma_i}{1 - \beta_i} \nabla f(x_k)$$

$$\quad = \tilde{x}_k - \frac{1}{m} \sum_{i=1}^{m} \frac{\gamma_i}{1 - \beta_i} \nabla f(x_k). \quad (5)$$

#### 2.1 Non-Convex Case

Below we present our main convergence result\(^1\) for non-convex problems.

**Theorem 2.1.** Let be $f$ is $L$-smooth and possibly non-convex function with values lower bounded

---

\(^1\)We defer all the proofs to the Appendix.
Algorithm 2 Aggregated Heavy-Ball method (AggHB)

**Input:** number of iterations \( N \), stepsize \( \gamma_i > 0 \), momentum parameters \( \{\beta_i\}_{i=1}^m \in [0, 1] \), starting points \( x_0, x_1 \) (by default \( x_1 = x_0 - \alpha \nabla f(x_0) \))

1: for \( k = 1, \ldots, N - 1 \) do
2: \( V_k^{(i)} = \beta_i V_{k-1}^{(i)} + \nabla f(x_k) \) for \( i = 1, \ldots, m \)
3: \( x_{k+1} = x_k - \frac{1}{m} \sum_{i=1}^m \gamma_i V_k^{(i)} \)
4: end for

**Output:** \( x_N \)

by \( f_{inf} \). Assume that

\[-\frac{A}{2} \left( 1 - \frac{CDEl^2}{2m^2} - LA \right) < 0, \tag{6}\]

where

\[A = \frac{1}{m} \sum_{i=1}^m \frac{\beta_i \gamma_i}{1 - \beta_i}, \quad C = \sum_{i=1}^m \frac{\gamma_i}{(1 - \beta_i)^2}, \quad D = \max_{i=1,m} \frac{\gamma_i}{1 - \beta_i}, \quad E = \sum_{i=1}^m \frac{1}{1 - \beta_i}. \tag{7}\]

Then, for all \( K \geq 1 \) we have

\[
\min_{k=1,K} \|\nabla f(x_k)\|_2^2 \leq \frac{2}{K} \frac{f(x_0) - f_{inf}}{A \left( 1 - \frac{CDEl^2}{m^2} - LA \right)}.
\tag{8}\]

The above result provides a convergence guarantee in the general non-convex case and allows to use different \( \gamma_i \) such that (6) holds. To illustrate this result and, in particular, condition (6) we derive the following corollary of Theorem 2.1.

**Corollary 2.2.** Let the assumptions of Theorem 2.1 hold. Assume that the stepsize is constant \( \gamma_i \equiv \gamma \) for \( i = 1, \ldots, m \) and consider new constants \( \tilde{\beta} \) and \( \hat{\beta} \) satisfying the following conditions:

\[
\frac{1}{m} \sum_{i=1}^m \frac{\beta_i}{1 - \beta_i} = \tilde{\beta}, \quad \frac{1}{m} \sum_{i=1}^m \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta}.
\]

Let

\[
\gamma = \frac{1}{L \left( \frac{2\tilde{\beta}}{1 - \beta} + \sqrt{2 \left( \frac{\tilde{\beta}}{(1 - \beta)^2} + \frac{1}{1 - \beta} \right) \left( \frac{1}{1 - \max_{i=1,m} \beta_i} (1 - \hat{\beta}) \right)} \right)}.
\]

Then, to achieve \( \min_{k=1,K} \|\nabla f(x_k)\|_2^2 \leq \varepsilon^2 \) for \( \varepsilon > 0 \) AggHB requires

\[
\mathcal{O} \left( \frac{L(f(x_0) - f_{inf})}{\varepsilon^2} + \frac{L(f(x_0) - f_{inf})}{\varepsilon^2} \sqrt{\frac{(1 - \tilde{\beta})}{(1 - \beta)^2} + 1} \frac{1}{\max_{i=1,m} \beta_i \hat{\beta}^2} \right). \tag{9}\]
First of all, when \( m = 1 \), we have \( \beta = \tilde{\beta} = \hat{\beta} = \max_{i=1,m} \beta_i \) and the above convergence rate can be simplified to
\[
O \left( \frac{L(f(x_0) - f_{\text{int}})}{\varepsilon^2} + \frac{L(f(x_0) - f_{\text{int}})}{\varepsilon^2 (1 - \beta)} \right)
\]
that matches the rate of HB in the non-convex case (e.g., see \cite{4}). Next, constants \( \tilde{\beta} \) and \( \hat{\beta} \) can be viewed as special “averaged” momentum parameters. Indeed, we know that
\[
\min_{i=1,m} \beta_i \leq \frac{1}{m} \sum_{i=1}^{m} \beta_i \leq \max_{i=1,m} \beta_i \leq \frac{1}{(1 - \min_{i=1,m} \beta_i)^2}.
\]
This allows to use larger stepsize than maximal possible stepsize for HB with \( \beta = \max_{i=1,m} \beta_i \), i.e., the rate of AggHB is better than the one of HB with \( \beta = \max_{i=1,m} \beta_i \).

### 2.2 Convex and Strongly-Convex Cases

**Lemma 2.3.** Let be \( f \) is \( L \)-smooth and \( \mu \)-strongly convex. Let \( \gamma_i \) and \( \beta_i \) satisfy \( \gamma_i > 0, \beta_i \in [0, 1) \), and
\[
F = \frac{1}{m} \sum_{i=1}^{m} \gamma_i \frac{1}{1 - \beta_i} \leq \frac{1}{4L}.
\]
Then, for all \( k \geq 0 \)
\[
\frac{F}{2} (f(x_k) - f(x_*)) \leq \left( 1 - \frac{F \mu}{2} \right) \|x_k - x_*\|_2 - \|x_{k+1} - x_*\|_2^2 + 3LF \|x_k - \bar{x}_k\|_2^2.
\]

Next, it is sufficient to sum up (11) for \( k = 0, 1, \ldots, K \) with weights \( w_k = (1 - \mu F/2)^{-k+1} \), \( W_k = \sum_{k=0}^{K} w_k \) to get the bound on \( f(\bar{x}_K) - f(x_*) \), where \( \bar{x}_K = \frac{1}{W_K} \sum_{k=1}^{K} w_k (f(x_k) - f(x_*)). \) To get final result one needs to upper bound the sum \( 3LF \sum_{k=0}^{K} w_k \|x_k - \bar{x}_k\|_2^2 \). For this we consider the following lemma.

**Lemma 2.4.** Assume that \( f \) is \( L \)-smooth and \( \mu \)-strongly convex. Let \( \gamma_i \) and \( \beta_i \) satisfy
\[
0 < \gamma_i \leq \frac{(1 - \max_{i=1,m} \beta_i) (1 - \beta_i)}{2 \mu}, \quad \beta_i \in [0, 1),
\]
\[
F = \frac{1}{m} \sum_{i=1}^{m} \gamma_i \frac{1}{1 - \beta_i} \leq \frac{1}{4L}, \quad BF \leq \frac{1}{4L^2} \leq \frac{1}{48L^2}, \quad \text{BF} \leq \frac{1}{48L^2},
\]
where \( B = \frac{1}{m} \sum_{i=1}^{m} \beta_i \gamma_i (1 - \beta_i^{K+1}) (1 - \beta_i)^2 \). Then, for all \( k \geq 0 \) and \( w_k = (1 - \mu F/2)^{-(k+1)} \)
\[
3LF \sum_{k=0}^{K} w_k \|x_k - \bar{x}_k\|_2^2 \leq \frac{F}{4} \sum_{k=0}^{K} w_k (f(x_k) - f(x_*)).
\]

Combining these lemmas, we get the main result in (strongly) convex case.
**Theorem 2.5.** Assume that \( f \) is \( L \)-smooth and \( \mu \)-strongly convex. Let \( \gamma_i \) and \( \beta_i \) satisfy conditions from (12) and (13). Then, after \( K \geq 0 \) iterations of AggHB we have

\[
f(x_K) - f(x_*) \leq \frac{4\|x_0 - x_*\|^2}{FW_K}, \quad x_K = \frac{1}{W_K} \sum_{i=1}^{K} i^K w_k(f(x_k) - f(x_*))
\]  

(15)

where \( w_k = (1 - \mu F/2)^{(k+1)} \), \( W_k = \sum_{k=0}^{K} w_k \), i.e.,

\[
f(x_K) - f(x_*) \leq \left(1 - \frac{\mu F}{2}\right)^{K} \frac{4\|x_0 - x_*\|^2}{F}, \quad \text{if } \mu > 0,
\]

(16)

\[
f(x_K) - f(x_*) \leq \frac{4\|x_0 - x_*\|^2}{FK}, \quad \text{if } \mu = 0.
\]

(17)

As in the non-convex case, the above result gives convergence guarantees in the general convex and strongly convex cases and allows to use different \( \gamma_i \) such that (12) and (13) hold. To illustrate this result and, in particular, conditions (12) and (13) we derive the following corollary of Theorem 2.5.

**Corollary 2.6.** Let the assumptions of Theorem 2.5 hold. Assume that the stepsize is constant \( \gamma_i \equiv \gamma \) for \( i = 1, \ldots, m \) and consider constants \( \tilde{\beta} \) and \( \hat{\beta} \) satisfying the following conditions:

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1 - \beta_i)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta}.
\]

Let

\[
\gamma = \min \left\{ \frac{\left(1 - \max_{i=1,m} \beta_i\right)^2}{2\mu}, \frac{1 - \beta}{4L}, \frac{(1 - \beta)\sqrt{(1 - \hat{\beta})\left(1 - \max_{i=1,m} \beta_i\right)}}{4\sqrt{3}L\sqrt{\beta}} \right\}.
\]

Then, to achieve \( f(x_K) - f(x_*) \leq \varepsilon \) for \( \varepsilon > 0 \) AggHB requires

\[
\mathcal{O}\left(\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta(1 - \beta)}}{\varepsilon(1 - \beta)\sqrt{1 - \max_{i=1,m} \beta_i}}\right)
\]  

(18)

iterations when \( \mu > 0 \), and

\[
\mathcal{O}\left(\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta(1 - \beta)}}{\varepsilon(1 - \beta)\sqrt{1 - \max_{i=1,m} \beta_i}}\right)
\]  

(19)

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

First of all, when \( m = 1 \), we have \( \beta = \tilde{\beta} = \hat{\beta} = \max_{i=1,m} \beta_i \) and the above convergence rates
Figure 2: Trajectories of HB and AHB with different momentum parameters $\beta$ applied to solve logistic regression problem with $\ell_2$-regularization (the first two rows) and non-convex regularization (the third row) for a9a, madelon, and australian datasets. Stepsize $\gamma$ was tuned for each method.

can be simplified to

$$O\left(\frac{L}{\mu} + \frac{L\sqrt{\beta}}{\mu(1 - \beta)}\right) \log \left(\frac{R_0^2}{\varepsilon} \cdot \left(L + \frac{L\sqrt{\beta}}{1 - \beta}\right)\right), \text{ when } \mu > 0,$$

$$O\left(\frac{LR_0^2}{\varepsilon} + \frac{LR_0^2\sqrt{\beta}}{\varepsilon(1 - \beta)}\right), \text{ when } \mu = 0$$

that matches the rate of HB in the strongly convex and convex cases (e.g., see [5]). Next, as we already mentioned before, constants $\beta$ and $\tilde{\beta}$ can be viewed as special “averaged” momentum parameters. This allows to use larger stepsize than maximal possible stepsize for HB with $\beta = \max_{i=1,m} \beta_i$, i.e., the rate of AggHB is better than the one of HB with $\beta = \max_{i=1,m} \beta_i$. 
3 Numerical Experiments

We compare the behavior of HB and AggHB on solving logistic regression problem with $\ell_2$-regularization and with special non-convex regularization:

$$
\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{M} \sum_{i=1}^{M} \log (1 + \exp (-y_i \cdot [Ax]_i)) + \frac{l_2}{2} \|x\|_2^2 \right\}, \tag{20}
$$

$$
\min_{x \in \mathbb{R}^n} \left\{ f(x) = \frac{1}{M} \sum_{i=1}^{M} \log (1 + \exp (-y_i \cdot [Ax]_i)) + \lambda \sum_{j=1}^{n} \frac{x_j^2}{1 + x_j^2} \right\}, \tag{21}
$$

where $M$ denotes the number of samples in the dataset, $A \in \mathbb{R}^{M \times n}$ is a “feature matrix”, $y_1, \ldots, y_M \in \{-1, 1\}$ are labels, and $l_2, \lambda \geq 0$ are the regularization parameters. One can show that $f(x)$ is $L$-smooth and $\mu$-strongly convex with $L = \frac{1}{M} \lambda_{\max}(A^TA) + l_2$ and $\mu = l_2$ in the first case, and $L$-smooth and non-convex with $L = \frac{1}{M} \lambda_{\max}(A^TA) + 2\lambda$. To construct the problems we use the following datasets from LIBSVM [1]: a9a ($M = 32561$, $n = 123$), madelon ($M = 2000$, $n = 500$), and australian ($M = 690$, $n = 14$). Regularization parameter $l_2$ is either 0 (convex problem) or $\frac{L}{10000}$ (strongly convex problem) and $\lambda$ is chosen as $\lambda = \frac{l_2}{1000}$. We run HB with standard momentum parameters $\beta = 0.9, 0.95$ for both problems. AggHB was tested with $m = 3$, $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\beta_3 = 0.99$, and $\gamma_1 = \gamma_2 = \gamma_3 = \gamma$ for $\ell_2$-regularized problem and with $m = 2$, $\beta_1 = 0.9$, $\beta_2 = 0.95$, and $\gamma_1 = \gamma_2 = \gamma$. For each method we tune stepsize parameter $\gamma$ as follows: we choose $\gamma = \frac{a}{t}$ with the best $a \in \{2^{-6}, 2^{-5}, 2^{-4}, \ldots, 2^8\}$, i.e., the method achieves the best accuracy with the chosen $a$ from the considered set.

The results are shown in the Figure 2. We observe that AggHB outperforms HB in all cases. In particular, for $\ell_2$-regularized problem the large value of $\beta_3$ does not slow down the convergence of AggHB. In contrast, we observed that HB performs relatively bad with $\beta = 0.99$. Next, in the experiments with non-convex regularization, AggHB takes the best from two choices of momentum parameters.

4 Conclusion

In this paper, we obtain the first convergence guarantees for AggHB without assuming that the gradients of the objective function are uniformly bounded. In the special case when $m = 1$, our results recover the known ones for HB and outperform the corresponding guarantees for HB with $\beta = \max_{i=1,m} \beta_i$ when $m > 1$. Our numerical results show the superiority of AggHB to HB. Together with the results from [7] they indicate high practical potential of AggHB.

References

[1] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. ACM transactions on intelligent systems and technology (TIST), 2(3):1–27, 2011.

[2] Marina Danilova, Pavel Dvurechensky, Alexander Gasnikov, Eduard Gorbunov, Sergey Guminov, Dmitry Kamzolov, and Innokentiy Shibaev. Recent theoretical advances in non-convex optimization. arXiv preprint arXiv:2012.06188, 2020.
[3] Marina Danilova and Grigory Malinovsky. Averaged heavy-ball method. arXiv preprint arXiv:2111.05430, 2021.

[4] Aaron Defazio. Momentum via primal averaging: Theoretical insights and learning rate schedules for non-convex optimization. arXiv preprint arXiv:2010.00406, 2020.

[5] Euhanna Ghadimi, Hamid Reza Feyzmahdavian, and Mikael Johansson. Global convergence of the heavy-ball method for convex optimization. In 2015 European control conference (ECC), pages 310–315. IEEE, 2015.

[6] Laurent Lessard, Benjamin Recht, and Andrew Packard. Analysis and design of optimization algorithms via integral quadratic constraints. SIAM Journal on Optimization, 26(1):57–95, 2016.

[7] James Lucas, Shengyang Sun, Richard Zemel, and Roger Grosse. Aggregated momentum: Stability through passive damping. In International Conference on Learning Representations, 2019.

[8] Horia Mania, Xinghao Pan, Dimitris Papailiopoulos, Benjamin Recht, Kannan Ramchandran, and Michael I Jordan. Perturbed iterate analysis for asynchronous stochastic optimization. SIAM Journal on Optimization, 27(4):2202–2229, 2017.

[9] A.S. Nemirovsky and D.B. Yudin. Problem Complexity and Method Efficiency in Optimization. J. Wiley & Sons, New York, 1983.

[10] Yurii Nesterov. A method for unconstrained convex minimization problem with the rate of convergence O(1/k^2). In Doklady an ussr, volume 269, pages 543–547, 1983.

[11] Yurii Nesterov. Lectures on convex optimization, volume 137. Springer, 2018.

[12] Mykola Novik. torch-optimizer – collection of optimization algorithms for PyTorch. github repository, 2020.

[13] Boris Polyak. Introduction to Optimization. New York, Optimization Software, 1987.

[14] Boris T Polyak. Some methods of speeding up the convergence of iteration methods. USSR Computational Mathematics and Mathematical Physics, 4(5):1–17, 1964.

[15] Peter Richtárik, Igor Sokolov, and Ilyas Fatkhullin. Ef21: A new, simpler, theoretically better, and practically faster error feedback. arXiv preprint arXiv:2106.05203, 2021.

[16] Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In International conference on machine learning, pages 1139–1147. PMLR, 2013.

[17] Adrien Taylor and Francis Bach. Stochastic first-order methods: non-asymptotic and computer-aided analyses via potential functions. In Conference on Learning Theory, pages 2934–2992. PMLR, 2019.

[18] Tianbao Yang, Qihang Lin, and Zhe Li. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. arXiv preprint arXiv:1604.03257, 2016.
A Missing Proofs from Section 2

A.1 Proof of Theorem 2.1

From $L$-smoothness of $f$ we have

$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - A \langle \nabla f(\bar{x}_k), \nabla f(x_k) \rangle + \frac{LA^2}{2} \|\nabla f(x_k)\|_2^2,$$  \hspace{1cm} (22)

where $A = \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{1-\beta_i} \gamma_i$. Next, we estimate a second term $-A \langle \nabla f(\bar{x}_k), \nabla f(x_k) \rangle$ in the previous expression:

$$-A \langle \nabla f(\bar{x}_k), \nabla f(x_k) \rangle = A \frac{1}{2} \left( \|\nabla f(\bar{x}_k) - \nabla f(x_k)\|_2^2 - \|\nabla f(\bar{x}_k)\|_2^2 - \|\nabla f(x_k)\|_2^2 \right)$$

$$\leq A \left( L^2 \|\bar{x}_k - x_k\|_2^2 - \|\nabla f(x_k)\|_2^2 \right) \tag{22}$$

$$= \frac{AL^2}{2m^2} \left( \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1-\beta_i} V_k^{(i)} \right) \left( \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1-\beta_i} V_k^{(i)} \right) \tag{4}$$

From AggHB update rule we know that $V_k^{(i)}$ is linear combination of gradients: $V_k^{(i)} = \sum_{l=0}^{k} (\beta_i)^l \nabla f(x_{k-l})$. Applying this to (23) we have

$$AL^2 \frac{2m^2}{2m^2} \left( \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1-\beta_i} V_k^{(i)} \right) \leq AL^2 \frac{2m^2}{2m^2} \sum_{l=0}^{k-1} \|\nabla f(x_{k-l})\|_2^2 \sum_{i=1}^{m} (\beta_i)^l \gamma_i,$$ \hspace{1cm} (24)

where $B = \sum_{l=0}^{k-1} \sum_{i=1}^{m} \frac{(\beta_i)^l \gamma_i}{1-\beta_i} \leq \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2}$. Combining (23), (24), we continue the derivation from (22):

$$f(\bar{x}_{k+1}) \leq f(\bar{x}_k) - A \frac{1}{2} \left( 1 - LA \right) \|\nabla f(x_k)\|_2^2 + \frac{AL^2 B}{2m^2} \sum_{l=0}^{k-1} \|\nabla f(x_{k-l})\|_2^2 \sum_{i=1}^{m} (\beta_i)^l \gamma_i \frac{1}{1-\beta_i}$$

$$+ \frac{AL^2}{2m^2} \left( \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2} \right) \left( \sum_{l=0}^{k-1} \|\nabla f(x_l)\|_2^2 \sum_{i=1}^{m} (\beta_i)^{k-1-l} \gamma_i \right)$$

$$\leq f(\bar{x}_k) - A \frac{1}{2} \left( 1 - LA \right) \|\nabla f(x_k)\|_2^2$$

$$+ AL^2 \left( \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2} \right) \left( \max_{i=1,m} \frac{\gamma_i}{1-\beta_i} \right) \sum_{l=0}^{k-1} \sum_{i=1}^{m} (\beta_i)^{k-1-l} \|\nabla f(x_l)\|_2^2.$$

Summing up (25) for $k = 0, 1, \ldots, K$ we get
\[ f(\bar{x}_{k+1}) \leq f(\bar{x}_0) + \sum_{k=1}^{K} \left( \frac{LA^2 - A}{2} \right) \| \nabla f(x_k) \|_2^2 \]
\[ + \sum_{k=1}^{K} \left( \frac{AL^2}{2m^2} \left( \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2} \right) \left( \max_{i=1,m} \frac{\gamma_i}{1-\beta_i} \right) \sum_{i=1}^{m} \sum_{l=k+1}^{K-1} \beta_i^{l-1-k} \right) \| \nabla f(x_k) \|_2^2 \]
\[ \leq f(\bar{x}_0) + \sum_{k=1}^{K} \left( \frac{LA^2 - A}{2} \right) \| \nabla f(x_k) \|_2^2 \]
\[ + \sum_{k=1}^{K} \left( \frac{AL^2}{2m^2} \left( \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2} \right) \left( \max_{i=1,m} \frac{\gamma_i}{1-\beta_i} \right) \sum_{i=1}^{m} \frac{1}{1-\beta_i} \right) \| \nabla f(x_k) \|_2^2 \]
\[ = f(\bar{x}_0) + \sum_{k=1}^{K} \left( \left( \frac{LA^2 - A}{2} \right) \frac{ACDEl^2}{2m^2} \right) \| \nabla f(x_k) \|_2^2, \]

where \( A = \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1-\beta_i}, \ C = \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2}, \ D = \max_{i=1,m} \frac{\gamma_i}{1-\beta_i}, \ E = \sum_{i=1}^{m} \frac{1}{1-\beta_i}. \)

Finally, by choosing sufficiently small \( \gamma_i \) one can ensure that \(-A \left( 1 - \frac{CDEl^2}{2m^2} - LA \right) \leq 0\) and get (8)

### A.2 Proof of Corollary 2.2

From Theorem 2.1 we have \( \min_{k=1,K} \| \nabla f(x_k) \|_2^2 \leq \frac{2}{K} \frac{f(x_0) - f_{\inf}}{A \left( 1 - \frac{CDEl^2}{m^2} - LA \right) \varepsilon^2} \) for all \( K \geq 1. \) This upper bound implies that to achieve \( \min_{k=1,K} \| \nabla f(x_k) \|_2^2 \leq \varepsilon^2, \) the method requires

\[ K = O \left( \frac{\Delta_0}{A \left( 1 - \frac{CDEl^2}{m^2} - LA \right) \varepsilon^2} \right) \] iterations, \( (26) \)

where \( \Delta_0 = f(x_0) - f_{\inf}. \) It remains to estimate the denominator in the above complexity bound.

To do that, we introduce new constants \( \hat{\beta} \) and \( \tilde{\beta} \) satisfying the following conditions

\[ \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1-\beta_i)^2} = \frac{\tilde{\beta}}{(1-\beta)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1-\beta_i} = \frac{1}{1-\tilde{\beta}}. \]

Next, using the formulas for \( \tilde{\beta}, \hat{\beta} \) and assuming \( \gamma_i = \gamma, \) we get new expressions for constants \( A, C, D, E: \)

\[ A = \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1-\beta_i} = \frac{\gamma \hat{\beta}}{1-\beta}, \ C = \sum_{i=1}^{m} \frac{\gamma_i}{(1-\beta_i)^2} = \gamma m \left( \frac{\tilde{\beta}}{(1-\beta)^2} + \frac{1}{1-\beta} \right), \]
\[ D = \max_{i=1,m} \frac{\gamma_i}{1-\beta_i} = \frac{\gamma}{1-\max \beta_i}, \ E = \sum_{i=1}^{m} \frac{1}{1-\beta_i} = \frac{m}{1-\beta}. \]
Then, condition (6), which is equivalent to \(1 - \frac{CDE}{m^2} - LA > 0\), can be written as

\[
1 - \gamma^2 L^2 \left( \frac{\tilde{\beta}}{(1 - \beta)^2} + \frac{1}{1 - \beta} \right) \cdot \frac{1}{\left( 1 - \max_{i=1,m} \beta_i \right) (1 - \tilde{\beta})} - \gamma L \frac{\tilde{\beta}}{1 - \beta} > 0.
\]

To derive the complexity stated in the corollary, we choose \(\gamma\) such that

\[
\frac{1}{2} - \gamma^2 L^2 \left( \frac{\tilde{\beta}}{(1 - \beta)^2} + \frac{1}{1 - \beta} \right) \cdot \frac{1}{\left( 1 - \max_{i=1,m} \beta_i \right) (1 - \tilde{\beta})} - \gamma L \frac{\tilde{\beta}}{1 - \beta} \geq 0,
\]

implying (6). One can show (see Lemma 5 from [15]) that this condition holds for

\[
\gamma = \frac{1}{L \left( \frac{2 \tilde{\beta}}{1 - \beta} + \sqrt{2 \left( \frac{\tilde{\beta}}{(1 - \beta)^2} + \frac{1}{1 - \beta} \right)} \frac{1}{\left( 1 - \max_{i=1,m} \beta_i \right) (1 - \tilde{\beta})} \right)}
\]

Plugging this value of \(\gamma\) in (26) and using \(1 - \frac{CDE}{m^2} - LA \geq \frac{1}{2}\), we finally obtain

\[
K = O \left( \frac{L \Delta_0}{\varepsilon^2} + \frac{L \Delta_0 \left( (1 - \tilde{\beta}) \left( \frac{\beta}{1 - \beta} + 1 \right) \frac{1}{\left( 1 - \max_{i=1,m} \beta_i \right) \beta^2} \right)}{\varepsilon^2} \right).
\]

### A.3 Proof of Lemma 2.3

Applying the virtual iterates determined in (4), we obtain

\[
\|x_{k+1} - x_*\|^2 = \|x_k - x_*\|^2 - 2F(x_k - x_*, \nabla f(x_k)) + F^2 \|\nabla f(x_k)\|^2
\]

\[
= \|x_k - x_*\|^2 - 2F(x_k - x_*, \nabla f(x_k)) - 2F(x_k - x_k, \nabla f(x_k))
\]

\[
+ F^2 \|\nabla f(x_k)\|^2.
\]

(27)

From \(\mu\)-strong convexity and \(L\)-smoothness of \(f\) we have (e.g., see [11])

\[
\langle x_k - x_*, \nabla f(x_k) \rangle \geq f(x_k) - f(x_*) + \frac{\mu}{2} \|x_k - x_*\|^2
\]

\[
\|\nabla f(x_k)\|^2 \leq 2L (f(x_k) - f(x_*)).
\]

Using these inequalities for (27) we get

\[
\|x_{k+1} - x_*\|^2 \leq \|x_k - x_*\|^2 - \mu F \|x_k - x_*\|^2 - 2F(f(x_k) - f(x_*))
\]

\[
- 2F(x_k - x_k, \nabla f(x_k)) + F^2 \|\nabla f(x_k)\|^2.
\]

Firstly, we evaluate the second term \(-\mu F \|x_k - x_*\|^2\) using that \(\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2\) for all \(a, b \in \mathbb{R}^n\) as follows

\[
-\mu F \|x_k - x_*\|^2 \leq -\frac{\mu F}{2} \|x_k - x_*\|^2 + \mu F \|x_k - x_k\|^2.
\]

13
Secondly, we estimate the fourth term $-2F\langle \bar{x}_k - x_k, \nabla f(x_k) \rangle$ using Fenchel-Young inequality\(^2\) and get
\[
-2F\langle \bar{x}_k - x_k, \nabla f(x_k) \rangle \leq -2LF\|\bar{x}_k - x_k\|^2 + \frac{F}{2L}\|\nabla f(x_k)\|^2 
\]
\[
\overset{(28)}{\leq} -2LF\|\bar{x}_k - x_k\|^2 + F\|f(x_k) - f(x_*)\|. 
\]
Combining the results above, we finish the proof
\[
\|x_{k+1} - x_*\|^2 \overset{(4),(11)}{\leq} \left(1 - \frac{\mu F}{2}\right)\|x_k - x_*\|^2 - \frac{F}{2} (f(x_k) - f(x_*)) + 3LF\|x_k - \bar{x}_k\|^2. 
\]

A.4 Proof of Lemma 2.4

From AggHB update rule we know that $V_k^{(i)}$ is linear combination of gradients: $V_k^{(i)} = \sum_{t=0}^{k} \beta^t_i \nabla f(x_{k-t})$

Next, by the definition of $\bar{x}_k$ we have
\[
\|x_{k+1} - \bar{x}_{k+1}\|^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} V_k^{(i)} \right\|^2 = \left\| \frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i \gamma_i}{1 - \beta_i} \sum_{t=0}^{k} \beta^t_i \nabla f(x_{k-t}) \right\|^2 
\]
\[
= \left\| \sum_{t=0}^{k} \left( \frac{1}{m} \sum_{i=1}^{m} \frac{\beta^t_{i+1} \gamma_i}{1 - \beta_i} \right) \nabla f(x_{k-t}) \right\|^2. 
\]
\[
(29) 
\]
Define constant $B_k$ as following
\[
B_k = \sum_{t=0}^{k} \frac{1}{m} \sum_{i=1}^{m} \frac{\beta^t_{i+1} \gamma_i}{1 - \beta_i} = \frac{1}{m} \sum_{i=1}^{m} \beta_i \gamma_i \sum_{t=0}^{k} \frac{\beta^t_k}{1 - \beta_i} = \frac{1}{m} \sum_{i=1}^{m} \beta_i \gamma_i \frac{(1 - \beta_{k+1})}{(1 - \beta_i)^2}. 
\]
Using this, we continue the derivation from (29)
\[
\|x_{k+1} - \bar{x}_{k+1}\|^2 = \min \left\{ B_k^2 \left\| \sum_{t=0}^{k} \frac{1}{B} \left( \frac{1}{m} \sum_{i=1}^{m} \frac{\beta^t_{i+1} \gamma_i}{1 - \beta_i} \right) \nabla f(x_t) \right\|^2 \right. 
\]
\[
\overset{\text{Jensen’s inequality}}{\leq} B_k^2 \sum_{t=0}^{k} \left( \frac{1}{m} \sum_{i=1}^{m} \frac{\beta^t_{i+1} \gamma_i}{1 - \beta_i} \right) \|\nabla f(x_t)\|^2 
\]
\[
= B_k \cdot \sum_{t=0}^{k} \left( \frac{1}{m} \sum_{i=1}^{m} \frac{\beta^t_{i+1} \gamma_i}{1 - \beta_i} \right) \|\nabla f(x_t)\|^2 
\]
\[
(13), B_k \leq B_{k+1} \leq B_K F \sum_{t=0}^{k} \max_{i=1, m} \beta^t_{i+1} \|\nabla f(x_t)\|^2. 
\]
\[
\]
\[
^2 |a, b| \leq \frac{\|a\|^2}{2\lambda} + \frac{\|b\|^2}{2} \text{ for all } a, b \in \mathbb{R}^n \text{ and } \lambda > 0.
\]
For simplicity, we denote $B_K \equiv B$. Summing up these inequalities for $k = 0, 1, \ldots, K$ with weights $w_k = \left(1 - \frac{\mu F}{2}\right)^{-(k+1)}$, we get

$$3L F \sum_{k=0}^{K} w_k \|x_k - \bar{x}_k\|_2^2 \leq 3LBF^2 \cdot \sum_{k=0}^{K} \sum_{t=0}^{k-1} w_k \max_{i=1,m} \beta_i^{k-t} \|\nabla f(x_t)\|_2^2$$

$$\leq 3LBF^2 \cdot \sum_{k=0}^{K} \sum_{t=0}^{k} w_k \max_{i=1,m} \beta_i^{k-t} \|\nabla f(x_t)\|_2^2. \quad (31)$$

Next, we estimate $w_k$ using that $(1 - q/2)^{-1} \leq 1 + q$ for any $q \in (0, 1]$; for all $t = 0, 1, \ldots, k$

$$w_k = \left(1 - \frac{\mu F}{2}\right)^{-(k-t)} w_t \leq (1 + \mu F)^{k-t} w_t \leq \left(1 + \frac{1 - \max_{i=1,m} \beta_i}{2}\right)^{k-t} w_t.$$ 

Using an inequality above and $(1 + q/2) (1 - q) \leq 1 - q/2$ for $q = 1 - \max_{i=1,m} \beta_i$, we continue the previous derivation (31)

$$3L F \sum_{k=0}^{K} w_k \|x_k - \bar{x}_k\|_2^2 \leq 3LBF^2 \sum_{k=0}^{K} \sum_{t=0}^{k} w_t \|\nabla f(x_t)\|_2^2 \left(1 + \frac{1 - \max_{i=1,m} \beta_i}{2}\right)^{k-t} \max_{i=1,m} \beta_i^{k-t}$$

$$\leq 3LBF^2 \sum_{k=0}^{K} \sum_{t=0}^{k} w_t \|\nabla f(x_t)\|_2^2 \left(1 - \frac{1 - \max_{i=1,m} \beta_i}{2}\right)^{k-t}$$

$$\leq 3LBF^2 \left(\sum_{k=0}^{K} w_k \|\nabla f(x_k)\|_2^2\right) \left(\sum_{t=0}^{\infty} \left(1 - \frac{1 - \max_{i=1,m} \beta_i}{2}\right)^t\right)$$

$$= \frac{6LBF^2}{1 - \max_{i=1,m} \beta_i} \sum_{k=0}^{K} w_k \|\nabla f(x_k)\|_2^2$$

$$\leq \frac{12L^2BF^2}{1 - \max_{i=1,m} \beta_i} \sum_{k=0}^{K} w_k (f(x_k) - f(x_*)). \quad (32)$$

We take parameters $\gamma_i, \beta_i$ (12) implying (13). Combining this with the last result (32), we obtain (14).

### A.5 Proof of Theorem 2.5

Using Lemma 2.3 we get

$$\frac{F}{2} (f(x_k) - f(x_*)) \leq \left(1 - \frac{\mu F}{2}\right) \|\bar{x}_k - x_*\|_2^2 - \|\bar{x}_{k+1} - x_*\|_2^2 + 3LF \|x_k - \bar{x}_k\|_2^2.$$
Summing up these inequalities for \( k = 0, 1, \ldots, K \) with weights \( w_k = \left( 1 - \frac{\mu F}{2} \right)^{-(k+1)} \), we have

\[
\frac{F}{2} \sum_{k=0}^{K} w_k \left( f(x_k) - f(x_*) \right) \leq \sum_{k=0}^{K} \left( w_k \left( 1 - \frac{\mu F}{2} \right) \|\bar{x}_k - x_*\|^2 - w_k \|\bar{x}_{k+1} - x_*\|^2 \right) + 3LF \sum_{k=0}^{K} w_k \|x_k - \bar{x}_k\|^2
\]

\[
\leq \sum_{k=0}^{K} \left( w_{k-1} \|\bar{x}_k - x_*\|^2 - w_k \|\bar{x}_{k+1} - x_*\|^2 \right) + \frac{F}{4} \sum_{k=0}^{K} w_k (f(x_k) - f(x_*))
\]

\[
\leq \|x_0 - x_*\|^2 + \frac{F}{4} \sum_{k=0}^{K} w_k (f(x_k) - f(x_*)).
\]

Rearranging and multiplying by \( \frac{1}{W_K} = \frac{1}{\sum_{k=0}^{K} w_k} \) this inequality, we have

\[
\frac{1}{W_K} \sum_{k=0}^{K} w_k (f(x_k) - f(x_*)) \leq \frac{4\|x_0 - x_*\|^2}{FW_K}.
\]

Next, we obtain (15) by using Jensen’s inequality:

\[
f(\bar{x}_K) \leq \frac{1}{W_K} \sum_{k=0}^{K} w_k f(x_k).
\]

In strongly convex case \((\mu > 0)\), we have \( W_K \geq w_{K-1} = \left( 1 - \frac{\mu F}{2} \right)^{-K} \), hence (16) holds. In convex case \((\mu = 0)\), \( W_K = K + 1 > K \) that implies (17).

**A.6 Proof of Corollary 2.6**

When \( \mu > 0 \), Theorem 2.5 implies (16)

\[
f(\bar{x}_K) - f(x_*) \leq \left( 1 - \frac{\mu F}{2} \right)^K \frac{4\|x_0 - x_*\|^2}{F} \leq \frac{4\|x_0 - x_*\|^2}{F} \exp \left( -\frac{\mu F}{2} K \right).
\]

Therefore, to ensure that the right-hand side is smaller than \( \varepsilon \), number of iterations \( K \) should satisfy

\[
K = \mathcal{O} \left( \frac{1}{\mu F} \ln \left( \frac{R_0^2}{\varepsilon F} \right) \right),
\]

where \( R_0 = \|x_0 - x_*\|^2 \). Assuming \( \gamma_i = \gamma \) for \( i = 1, \ldots, m \) and using constants \( \widetilde{\beta} \) and \( \hat{\beta} \) defined as

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1 - \beta_i)^2} = \frac{\widetilde{\beta}}{(1 - \beta)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta},
\]

Thus, the number of iterations to ensure that the right-hand side is smaller than \( \varepsilon \) is

\[
K = \mathcal{O} \left( \frac{1}{\mu F} \ln \left( \frac{R_0^2}{\varepsilon F} \right) \right),
\]

where \( R_0 = \|x_0 - x_*\|^2 \). Assuming \( \gamma_i = \gamma \) for \( i = 1, \ldots, m \) and using constants \( \widetilde{\beta} \) and \( \hat{\beta} \) defined as

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1 - \beta_i)^2} = \frac{\widetilde{\beta}}{(1 - \beta)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta},
\]

Thus, the number of iterations to ensure that the right-hand side is smaller than \( \varepsilon \) is

\[
K = \mathcal{O} \left( \frac{1}{\mu F} \ln \left( \frac{R_0^2}{\varepsilon F} \right) \right),
\]

where \( R_0 = \|x_0 - x_*\|^2 \). Assuming \( \gamma_i = \gamma \) for \( i = 1, \ldots, m \) and using constants \( \widetilde{\beta} \) and \( \hat{\beta} \) defined as

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1 - \beta_i)^2} = \frac{\widetilde{\beta}}{(1 - \beta)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta},
\]

Thus, the number of iterations to ensure that the right-hand side is smaller than \( \varepsilon \) is

\[
K = \mathcal{O} \left( \frac{1}{\mu F} \ln \left( \frac{R_0^2}{\varepsilon F} \right) \right),
\]

where \( R_0 = \|x_0 - x_*\|^2 \). Assuming \( \gamma_i = \gamma \) for \( i = 1, \ldots, m \) and using constants \( \widetilde{\beta} \) and \( \hat{\beta} \) defined as

\[
\frac{1}{m} \sum_{i=1}^{m} \frac{\beta_i}{(1 - \beta_i)^2} = \frac{\widetilde{\beta}}{(1 - \beta)^2}, \quad \frac{1}{m} \sum_{i=1}^{m} \frac{1}{1 - \beta_i} = \frac{1}{1 - \beta},
\]
we get that
\[
B = \frac{1}{m} \sum_{i=1}^{m} \beta_i \gamma_i \left(1 - \beta_i^{K+1}\right) \leq \frac{\gamma \tilde{\beta}}{(1 - \beta)^2}, \quad F = \frac{1}{m} \sum_{i=1}^{m} \gamma_i (1 - \beta_i) = \frac{\gamma}{1 - \beta}.
\]

Therefore, the conditions from (13) hold when
\[
\frac{\gamma}{1 - \beta} \leq \frac{1}{4L}, \quad \frac{\gamma^2 \tilde{\beta}}{(1 - \beta)^2(1 - \beta)} \leq \frac{1 - \max_{i=1,m} \beta_i}{48L^2}.
\]

These inequalities together with (12) are satisfied for
\[
\gamma = \min \left\{ \left(1 - \max_{i=1,m} \beta_i\right)^2, \frac{1 - \tilde{\beta}}{4L}, \frac{(1 - \tilde{\beta}) \sqrt{(1 - \tilde{\beta}) (1 - \max_{i=1,m} \beta_i)}}{4\sqrt{3L} \sqrt{\tilde{\beta}}} \right\}.
\] (35)

Plugging (35) in (34), we derive the following complexity result:
\[
K = \mathcal{O} \left(\frac{1}{\mu F \ln \frac{R_0^2}{F \varepsilon}}\right) = \mathcal{O} \left(\frac{1 - \tilde{\beta}}{\gamma \mu} \ln \left(\frac{(1 - \tilde{\beta}) R_0^2}{\gamma \varepsilon}\right)\right)
\]
\[
= \mathcal{O} \left(\max \left(\frac{L}{\mu}, \frac{1 - \tilde{\beta}}{(1 - \max_{i=1,m} \beta_i)^2}, \frac{L \sqrt{\tilde{\beta} (1 - \tilde{\beta})}}{\mu (1 - \tilde{\beta}) \sqrt{(1 - \max_{i=1,m} \beta_i)}}\right)\right)
\]
\[
\cdot \ln \left(\frac{R_0^2}{\varepsilon} \cdot \max \left(L, \frac{1 - \tilde{\beta}}{(1 - \max_{i=1,m} \beta_i)^2}, \frac{L \sqrt{\tilde{\beta} (1 - \tilde{\beta})}}{(1 - \tilde{\beta}) \sqrt{(1 - \max_{i=1,m} \beta_i)}}\right)\right)
\]
\[
= \mathcal{O} \left(\left(\frac{L}{\mu} + \frac{1 - \tilde{\beta}}{(1 - \max_{i=1,m} \beta_i)^2} + \frac{L \sqrt{\tilde{\beta} (1 - \tilde{\beta})}}{\mu (1 - \tilde{\beta}) \sqrt{(1 - \max_{i=1,m} \beta_i)}}\right)\right)
\]
\[
\cdot \ln \left(\frac{R_0^2}{\varepsilon} \cdot \left(L + \frac{1 - \tilde{\beta}}{(1 - \max_{i=1,m} \beta_i)^2} + \frac{L \sqrt{\tilde{\beta} (1 - \tilde{\beta})}}{(1 - \tilde{\beta}) \sqrt{(1 - \max_{i=1,m} \beta_i)}}\right)\right).
\]

When \(\mu = 0\), Theorem 2.5 implies (16)
\[
f(\bar{x}_K) - f(x_\ast) \leq \frac{4\|x_0 - x_\ast\|^2_2}{FK}.
\] (36)

Therefore, to ensure that the right-hand side is smaller than \(\varepsilon\), number of iterations \(K\) should satisfy
\[
K = \mathcal{O} \left(\frac{R_0^2}{\varepsilon^2 F}\right).
\] (37)

Plugging (35) in (37), we get (19).