S©MR

СИБИРСКИЕ ЭЛЕКТРОННЫЕ
МАТЕМАТИЧЕСКИЕ ИЗВЕСТИЯ
Siberian Electronic Mathematical Reports
http://semr.math.nsc.ru
ISSN 1813-3304

Vol. 20, No. 2, pp. 144–145 (2023) УДК 519.853.62
https://doi.org/10.33048/semi.2023.20.??? MSC 68T07

INCORPORATING PRECONDITIONING INTO ACCELERATED APPROACHES: THEORETICAL GUARANTEES AND PRACTICAL IMPROVEMENT

S.D. TRIFONOV, **L.I. LEVIN**, **S.A. CHEZHEGOV**,
AND **A.N. BEZNOSIKOV**

Communicated by P.P. PETROV

Abstract: Machine learning and deep learning are widely researched fields that provide solutions to many modern problems. Due to the complexity of new problems related to the size of datasets, efficient approaches are obligatory. In optimization theory, the Heavy Ball and Nesterov methods use *momentum* in their updates of model weights. On the other hand, the minimization problems considered may be poorly conditioned, which affects the applicability and effectiveness of the aforementioned techniques. One solution to this issue is *preconditioning*, which has already been investigated in approaches such as ADAGRAD, RMSPROP, ADAM and others. Despite this, momentum acceleration and preconditioning have not been fully explored together. Therefore, we propose the Preconditioned Heavy Ball (PHB) and Preconditioned Nesterov method (PN) with theoretical guarantees of convergence under *unified* assumption on the scaling matrix. Furthermore, we provide numerical experiments that demonstrate superior performance compared

TRIFONOV, S.D., LEVIN, L.I., CHEZHEGOV, S.A., BEZNOSIKOV, A.N., INCORPORATING PRECONDITIONING INTO ACCELERATED APPROACHES: THEORETICAL GUARANTEES AND PRACTICAL IMPROVEMENT.

The work is supported by the Ministry of Science and Higher Education of the Russian Federation, project No. FSMG-2024-0011.

Received June, 30, 2025, Published November, 30, 2025.

to the unscaled techniques in terms of iteration and oracle complexities.

Keywords: adaptive optimization, preconditioning, accelerated methods

1 Introduction

A classical optimization problem can be stated as

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

where the function f can have different meanings depending on the setting of the machine learning [25] and deep learning [12] problems, such as loss function, risk function, etc. The most intuitive, yet simple, approach for solving the problem (1) is Gradient Descent [2]. However, there are other first-order oracle approaches allow for both theoretical and practical improvements. These methods include the Heavy-Ball method [21] and Accelerated Gradient Descent [19], which serve as the foundation for a wide range of approaches aimed at addressing narrowly focused problems.

Although the momentum technique provides acceleration, it does not address ill-conditioned problems. Therefore, techniques related to the scaling of directions in the iterative process are relevant here. First, it is worth mentioning Newton method [20], which uses second-order information into the update rule. However, calculating the inverse matrix of second derivatives is regarded as an expensive oracle. Therefore, quasi-Newton methods [9, 11, 26, 15], based on the Hessian approximation, was emerged. The essence of this approach is quite simple: instead of using the exact inverse Hessian in Newton method, we use its approximation for efficiency. Despite the fact that quasi-Newton methods often perform well in deterministic settings, they tend to struggle in the presence of stochasticity.

Subsequently, approaches that are categorized as *adaptive* began to emerge due to their use of update rules that pull information from previous points. Such approaches include AdaGrad [7], RMSProp [27], Adam [14], NAdam [6], AMSGrad [23], OASIS [13], AdaHessian [28] and many others. Their distinctive feature from Newton and quasi-Newton methods is that the preconditioning matrix has a diagonal form (as a consequence, it becomes much cheaper to reverse it). Along with their empirical superiority over other existing techniques, adaptive methods are frequently chosen to train machine learning models.

At present, considerable research is devoted to the convergence analysis of adaptive methods, as well as to the integration of preconditioning into techniques such as variance reduction [24], extragradient [1] and distributed approaches [22, 4] in a generalized form. Despite this, scaling has not been considered under general assumptions on the preconditioning matrix for accelerated approaches.

Our contribution. Therefore, our contribution can be formulated as follows:

- We present two algorithms that are scaled versions of the Heavy-Ball and Nesterov methods.
- We propose theoretical guarantees of presented methods under unified assumption on the precondioning matrix and validate numerical experiments to show the outperformance compared to unscaled techniques.

2 Related Works

Currently, methods such as the Heavy-Ball method and Accelerated Gradient Descent have been investigated in detail under various assumptions concerning the target function and, in some cases, on the first-order stochastic oracle [21, 19, 10, 8]. In this paper, we rely on the analysis of the Heavy-Ball method from [5] and the classical analysis of the Nesterov three-point method [18].

If we discuss preconditioning techniques, we should mention methods that use only gradient, such as ADAGRAD [7], which originally motivated to solve the problem of nonsmooth optimization. Subsequent developments extended this approach through exponential moving averages (e.g., RMSPROP [27]) and momentum techniques (e.g., ADAM [14]), now widely adopted as default optimizers in deep learning. Moreover, several variants of ADAM have since been proposed, including Nesterov momentum – NADAM [6], modified second moment estimation – AMSGRAD [23], and decoupled weight decay – ADAMW [17].

There are also approaches where the analogue of second-order information rather than the gradient itself is used to adapt the scaling matrix. For example, such methods include OASIS [13], SOPHIA [16] and ADAHESSIAN [28] – methods with Hutchinson approximation which aim to leverage curvature information while retaining scalability. The main idea is based on avoiding explicit Hessian computation and instead evaluating the so-called Hessian-vector product. This approach is more efficient and yields a scaling matrix that approximately captures second-order information.

As already mentioned, preconditioning has already been considered under the general assumption as a modification of existing approaches. For example, variance reduced methods such as SCALED SARAH and SCALED L-SVRG [24] was derived. Moreover, for the saddle point problem, a scaled version of the Extragradient method [1] was considered. In the distributed optimization framework, the scaled LOCAL SGD method [4] was already analyzed.

3 Preliminaries, requirements and notations

Throughout this paper, we adopt the following notations. The scalar product in L_2-space is denoted as $\langle x, y\rangle$. We denote $\|x\| = \|x\|_2 = \sqrt{\langle x, y\rangle}$ as the norm in L_2-space. The induced scalar product is denoted as $\langle x, y\rangle_A =$

$\langle x, Ay\rangle$. The norm induced by a positive definite matrix A we denote as $\|x\|_A = \sqrt{\langle x, Ax\rangle}$. The Hadamard product of two same-dimensional elements A and B is denoted as $A \odot B$.

3.1. Objective function.

As for the assumptions on the target function, we suppose that the minimized function f from (1) is μ-strongly convex and L-smooth.

Assumption 1 (μ-strong convexity)**.** *The function f is μ-strongly convex, i.e $\forall x, y \in \mathbb{R}^d$*

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

Assumption 2 (L-smoothness)**.** *The function f is L-smooth, i.e $\forall x, y \in \mathbb{R}^d$*

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

Also L-smoothness can be written in an equivalent formulation which is more useful for theoretical proofs.

Proposition 1. *Assume that $f(x)$ satisfies Assumption 2. Therefore, it holds that $\forall x \in \mathbb{R}^d$*

$$\|\nabla f(x)\|^2 \leq 2L(f(x) - f(x^*)).$$

3.2. Preconditioning.

The update of the scaling matrix may differ from one approach to another. Nevertheless, this procedure can be described as sequence of diagonal matrices $\{D_k\}_{k=-1}^{\infty}$, where each D_k can be evaluated by the following rule:

$$D_k = \text{Update}(D_{k-1}, H_k), \tag{2}$$

where H_k is some information received at iteration k in the diagonal form. For instance, AdaGrad [7] update can be written as

$$D_k = \sqrt{D_{k-1}^2 + \text{diag}(\nabla f(x_k) \odot \nabla f(x_k))}.$$

Moreover, some methods use so-called *exponential smoothing*. Prominent examples of such approaches are RMSProp [27] and Adam [14], where the update (2) can be written as

$$D_k = \sqrt{\beta_2 D_{k-1}^2 + (1 - \beta_2)\text{diag}(\nabla f(x_k) \odot \nabla f(x_k))}, \tag{3}$$

where the parameter β_2 lies in $(0, 1)$. In theoretical studies [24, 4], the standard choice for the smoothing parameter β_2 is $\beta_2(k) = 1 - \frac{1}{k}$ or $\beta_2 = 1 - \frac{1}{K}$, while in practice a commonly used value is 0.999.
Sometimes, instead of exponential squared smoothing, exponential smoothing for linear summands can be used, such as in the OASIS [13] method:

$$D_k = \beta_2 D_{k-1} + (1 - \beta_2)|\text{diag}(v_k \odot \nabla^2 f(x_k) v_k)|, \tag{4}$$

where v_k is the vector of random variables sampled from the Rademacher distribution.

In general form, (3) and (4) can be represented as

$$D_k = \sqrt{\beta_2 D_{k-1}^2 + (1-\beta_2)H_k^2}, \tag{5}$$

or

$$D_k = \beta_2 D_{k-1} + (1-\beta_2)|H_k|. \tag{6}$$

We devote our analysis exactly to these rules, (5)-(6), for updating the matrix D_k. A natural assumption that is made on the sequence $\{D_k\}_{k=-1}^{\infty}$ is the boundedness of the scaling matrices.

Assumption 3. *The sequence $\{D_k\}_{k=-1}^{\infty}$ is bounded, i.e. there exist constants $\Gamma \geq e > 0$ such that for all k*

$$eI \preceq D_k \preceq \Gamma I.$$

This assumption is indeed valid in terms of Γ – usually this value is strongly correlated with the dimensionality of the problem and the smoothness constant (see [4], Table 1). However, the assumption of boundedness from below leaves questions. In this case, the following trick can be applied for the sequence $\{D_k\}$ for both rules (5)-(6):

$$[\hat{D}_k]_{ii} := \max\{e, [D_k]_{ii}\}. \tag{7}$$

Let us also formulate two auxiliary facts about preconditioning matrices.

Proposition 2. *Assume that the matrix A satisfies*

$$eI \preceq A \preceq \Gamma I$$

for some constants $\Gamma \geq e > 0$. Therefore,

$$e\|x\|^2 \leq \|x\|_A^2 \leq \Gamma\|x\|^2;$$
$$\frac{1}{\Gamma}\|x\|^2 \leq \|x\|_{A^{-1}}^2 \leq \frac{1}{e}\|x\|^2.$$

To prove Proposition 2, it is enough to apply the definition of the induced norm.

Proposition 3 ([1] – Lemma 1, [4] – Corollary 1)**.** *Assume that the sequence of matrices $\{D_k\}_{k=-1}^{\infty}$ satisfies Assumption 3. If the update rule (2) has the form (5)/(6) with (7), then the following inequality holds:*

$$\|x\|_{D_{k+1}}^2 \leq (1+(1-\beta_2)C)\|x\|_{D_k}^2,$$

where the constant C depends on the preconditioner update rule. To be more precise,

$$C = \begin{cases} \frac{\Gamma^2}{2e^2} & \textit{for } (5); \\ \frac{2\Gamma}{e} & \textit{for } (6). \end{cases}$$

4 Algorithms and Theoretical Results

In this section, we present the design of PRECONDITIONED HEAVY-BALL METHOD and PRECONDITIONED NESTEROV METHOD with its theoretical guarantees of convergence.

Algorithm 1 PRECONDITIONED HEAVY-BALL METHOD

```
Input: initial point $x_0$, parameter of momentum $\beta_1$, initial momentum
    $V_{-1} = 0$, initial scaling matrix $\hat{D}_{-1} \succeq eI$
for $k = 0, 1, 2, \dots, T-1$ do
    $V_k = \beta_1 V_{k-1} + \hat{D}_k^{-1} \nabla f(x_k)$
    $x_{k+1} = x_k - \gamma V_k$
    $[\hat{D}_{k+1}]_{ii} = \max\{e, [\mathrm{Update}(D_k, H_{k+1})]_{ii}\}$
end for
```

Algorithm 2 PRECONDITIONED NESTEROV METHOD

```
Input: initial point $x^0 = x_f^0 = x_g^0$, learning rate $\gamma_k$, initial scaling matrix
    $\hat{D}_{-1} \succeq eI$, momentums $\{\xi_k\}_{k=0} \geq 0, \{\theta_k\}_{k=0} \geq 0$
for $k = 0, 1, 2, \dots, T-1$ do
    $x_f^{k+1} = x_g^k - \gamma_k \hat{D}_k^{-1} \nabla f(x_g^k)$
    $x^{k+1} = \xi_k (x_f^{k+1} - x_f) + x_f^k$
    $x_g^{k+1} = \theta_{k+1} x_f^{k+1} + (1 - \theta_{k+1}) x^{k+1}$
    $[\hat{D}_{k+1}]_{ii} = \max\{e, [\mathrm{Update}(D_k, H_{k+1})]_{ii}\}$
end for
```

The algorithmic design is fairly standard. Unlike classical Gradient Descent, we employ either the heavy-ball momentum or Nesterov momentum as an acceleration mechanism. At the same time, the core modification of existing accelerated methods lies in the scaling matrix $\hat{D}_k$, which is updated according to rules (5)/(6) and (7). As previously mentioned, many adaptive methods naturally fit into such preconditioning schemes, making our proposed algorithms, in a sense, unified frameworks.

We now formulate the theorems that provide convergence guarantees for Algorithms 1 and 2.

Theorem 1. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations of Algorithm 1 with $\gamma = \frac{(1-\beta_1)^2 e}{12L}$, we have*

$$f\left(\frac{1}{W_{K-1}} \sum_{k=0}^{K-1} w_k x_k\right) - f(x^*) \leq 4 \exp\left(-\frac{(1-\beta_1)\mu e K}{48 L \Gamma}\right) L \|x_0 - x^*\|_{\hat{D}_{-1}}^2,$$

where $W_{K-1} = \sum_{k=0}^{K-1} w_k$, $w_k = \left(1 - \frac{\mu F}{4\Gamma}\right)^{-(k+1)}$.

Remark 1. *It is worth noting that the technique of point reweighting is standard in the analysis of scaling-based methods. For further details, see Appendix.*

As a consequence, the upper bound on the required number of iterations K for reaching ε-accuracy can be formulated as follows.

Corollary 1. *Under the conditions of Theorem 1, the required number of iterations K of Algorithm 1 for reaching ε-accuracy, i.e. $f(x_{out}) - f(x^*) \leq \varepsilon$, can be upper bounded as*

$$K = \mathcal{O}\left(\frac{\Gamma}{e}\frac{L}{\mu(1-\beta_1)}\log\left(\frac{L\|x_0 - x^*\|^2}{\varepsilon}\right)\right).$$

As for the Algorithm 2, the theorem of convergence can be formulated as follows.

Theorem 2. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations of Algorithm 2 with $\gamma_k \equiv \frac{e}{L}, \xi_k \equiv \sqrt{\frac{L\Gamma}{\mu e}}$ and $\theta_k \equiv \frac{\sqrt{L\Gamma}}{\sqrt{\mu e}+\sqrt{L\Gamma}}$, we have*

$$\|x^{k+1} - x^*\|^2_{\hat{D}_k} \leq \exp\left(-K\sqrt{\frac{\mu e}{4L\Gamma}}\right)\left[\|x^0 - x^*\|^2_{\hat{D}_{-1}} + \frac{2\Gamma}{\mu}(f(x^0) - f(x^*))\right].$$

As a result, the next corollary holds.

Corollary 2. *Under the conditions of Theorem 2, the required number of iterations K of Algorithm 2 for reaching ε-accuracy, i.e. $\left\|x^k - x^*\right\|^2 \leq \varepsilon$, can be upper bounded as*

$$K = \mathcal{O}\left(\sqrt{\frac{\Gamma}{e}}\sqrt{\frac{L}{\mu}}\log\left(\frac{\|x_0 - x^*\|^2 + \frac{\Gamma}{\mu}(f(x^0) - f(x^*))}{\varepsilon}\right)\right).$$

5 Discussion of the results

Let us highlight the differences between the convergence guarantees of the proposed Algorithms 1 and 2 compared to their unpreconditioned versions. Two key aspects are worth noting:

(1) **The induced norm**. As a convergence criterion in Theorems 1 and 2, we employ the norm induced by the preconditioning matrix. Essentially, one can transition to the L_2-space by applying Proposition 2, which introduces a multiplicative factor Γ related to the initial distance to the optimum. However, this factor is not

critical, as it appears inside a *logarithmic* term in the iteration complexity bound on K.

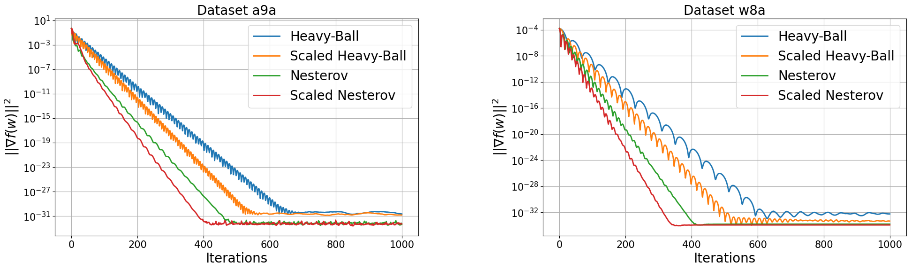
(2) **Exponential term.** In turn, the exponential factor in Theorems 1 and 2 no longer yields a polylogarithmic dependence on $\frac{\Gamma}{e}$ in the final bound on K. Specifically, Theorem 1 contributes a term $\frac{\Gamma}{e}$, while Theorem 2 contributes a term $\sqrt{\frac{\Gamma}{e}}$.

As a result, for the heavy-ball method with scaling, no improvement in the dependence on the multiplicative factor is observed – similar dependence had already been established in prior works [1, 24, 4] considering various techniques. This effect can be explained rather straightforwardly: the heavy-ball method does not provide *theoretical* acceleration for the class of functions satisfying Assumptions 1 and 2. In contrast, the scaled version of Nesterov method demonstrates, to the best of our knowledge, the first improvement of its kind with respect to the factor $\frac{\Gamma}{e}$, as this method enables theoretical acceleration under the given assumptions.
This phenomenon can be attributed to the fact that the scaling matrix allows one to operate over a modified landscape of the objective function, where, due to Proposition 2, the effective smoothness and strong convexity constants are altered accordingly; that is, μ becomes $\frac{\mu}{\Gamma}$, and L becomes $\frac{L}{e}$.

6 Experiments

In this section, we describe the experimental setups and present the numerical results.

6.1. Setup. As the target task, we consider a classical yet illustrative machine learning problem – binary classification. In the following, we describe the experimental setup in more detail.
•*Datasets.* We utilize the `a9a` and `w8a` LibSVM [3] datasets in our experiments. These datasets are chosen due to their diverse characteristics and their applicability to classification tasks. We divide the data into training and test parts in a percentage ratio of 80% for training and 20% for testing.
•*Metric.* Since we solve the classification problem, we use standard metrics such as cross-entropy loss. To estimate the rate of convergence we use the square norm of the gradient.
•*Model.* For our experiments, we chose linear model, which in combination with cross-entropy loss handles the binary classification problem effectively.
•*Optimization methods.* For our experiments, we implemented two optimization methods such as Heavy Ball, Nesterov and their scaled versions. Preconditioning matrix is chosen as for the Adam approach. For Algorithm 1, we chose a momentum parameter $\beta_1 = 0.9$, for Algorithm 2 we chose the hyperparameters according to the theory. Learning rate γ is chosen as the best option after tuning.



(A) Dataset a9a

(B) Dataset w8a

FIG. 1. Comparison of Scaled Heavy-Ball and Nesterov methods with unscaled versions.

6.2. Experiment Results. The results of numerical experiments are presented above. Although the theoretical estimates get worse (due to the $\frac{\Gamma}{e}$ times increase in the condition number of the problem), in practice we see a rather expected effect – scaled versions of the algorithms allow us to converge faster to the optimum compared to the unpreconditioned techniques. This indicates the applicability of the approach we presented.

7 Conclusion

In this work, we proposed the design of two accelerated algorithms incorporating a preconditioning matrix. For the proposed methods PRECONDITIONED HEAVY BALL and PRECONDITIONED NESTEROV, we also provided theoretical convergence guarantees. These guarantees involve an additional multiplicative factor $\frac{\Gamma}{e}$ that slightly worsens the upper bound on the number of iterations. However, such a factor is standard for methods of this class. Moreover, our empirical results demonstrate that the scaled versions of the algorithms significantly outperform their unscaled counterparts in terms of convergence speed, confirming the practical effectiveness of the proposed approach. The presented study also opens avenues for future research, particularly in the directions of stochastic extensions and analysis under generalized smoothness assumptions.

References

[1] A. Beznosikov, A. Alanov, D. Kovalev, M. Takáč, and A. Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.

[2] A. Cauchy et al. Méthode générale pour la résolution des systemes d'équations simultanées. *Comp. Rend. Sci. Paris*, 25(1847):536–538, 1847.

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

[4] S. Chezhegov, S. Skorik, N. Khachaturov, D. Shalagin, A. Avetisyan, M. Takáč, Y. Kholodov, and A. Beznosikov. Local methods with adaptivity via scaling. *arXiv preprint arXiv:2406.00846*, 2024.

[5] M. Danilova. On the convergence analysis of aggregated heavy-ball method. In *International Conference on Mathematical Optimization Theory and Operations Research*, pages 3–17. Springer, 2022.

[6] T. Dozat. Incorporating nesterov momentum into adam. 2016.

[7] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[8] N. Flammarion and F. Bach. From averaging to acceleration, there is only a step-size. In *Conference on learning theory*, pages 658–695. PMLR, 2015.

[9] R. Fletcher and M. J. Powell. A rapidly convergent descent method for minimization. *The computer journal*, 6(2):163–168, 1963.

[10] S. Ghadimi and G. Lan. Optimal stochastic approximation algorithms for strongly convex stochastic composite optimization i: A generic algorithmic framework. *SIAM Journal on Optimization*, 22(4):1469–1492, 2012.

[11] D. Goldfarb. A family of variable-metric methods derived by variational means. *Mathematics of computation*, 24(109):23–26, 1970.

[12] I. Goodfellow, Y. Bengio, and A. Courville. Deep feedforward networks. *Deep learning*, 1:161–217, 2016.

[13] M. Jahani, S. Rusakov, Z. Shi, P. Richtárik, M. W. Mahoney, and M. Takáč. Doubly adaptive scaled algorithm for machine learning using second-order information. *arXiv preprint arXiv:2109.05198*, 2021.

[14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] D. C. Liu and J. Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[16] H. Liu, Z. Li, D. Hall, P. Liang, and T. Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. *arXiv preprint arXiv:2305.14342*, 2023.

[17] I. Loshchilov. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[18] Y. Nesterov. *Introductory lectures on convex optimization: A basic course*, volume 87. Springer Science & Business Media, 2013.

[19] Y. E. Nesterov. A method for unconstrained convex minimization problem with the rate of convergence $o(1/k^2)$. *Doklady AN SSSR*, 269:543–547, 1983.

[20] I. Newton. *De analysi per aequationes numero terminorum infinitas*. 1711.

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

[22] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan. Adaptive federated optimization. *arXiv preprint arXiv:2003.00295*, 2020.

[23] S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. *arXiv preprint arXiv:1904.09237*, 2019.

[24] A. Sadiev, A. Beznosikov, A. J. Almansoori, D. Kamzolov, R. Tappenden, and M. Takáč. Stochastic gradient methods with preconditioned updates. *Journal of Optimization Theory and Applications*, 201(2):471–489, 2024.

[25] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[26] D. F. Shanno. Conditioning of quasi-newton methods for function minimization. *Mathematics of computation*, 24(111):647–656, 1970.

[27] T. Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

[28] Z. Yao, A. Gholami, S. Shen, M. Mustafa, K. Keutzer, and M. Mahoney. Adahessian: An adaptive second order optimizer for machine learning. In *proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 10665–10673, 2021.

A Supplementary materials

Before we start, let us formulate auxiliary propositions.

Proposition 4 (Young's inequality)**.** *For all $x, y \in \mathbb{R}^d$ and for any $\lambda > 0$ the next inequality holds:*

$$\langle x, y\rangle \leq \frac{\|x\|^2}{2\lambda} + \frac{\lambda \|y\|^2}{2}.$$

Proposition 5 (Jensen's inequality)**.** *For any convex $f : \mathbb{R}^d \to \mathbb{R}$ and $\{a_i\}_{i=1}^n : a_i \geq 0$ and $\sum\limits_{i=1}^n a_i = 1$, the next inequality holds:*

$$f\left(\sum_{i=1}^n a_i x_i\right) \leq \sum_{i=1}^n a_i f\left(x_i\right).$$

A.1. Proof of Preconditioned Heavy-Ball method.

Lemma 1 (Descent lemma). *Suppose that Assumptions 1, 2 and 3 hold. Then, if the sequence $\{x_k\}_{k=0}$ is produced by Algorithm 1 and γ satisfies*

$$F := \frac{\gamma}{1-\beta_1} \leq \frac{e}{4L},$$

the next inequality holds:

$$\frac{F}{2}\left(f(x_k) - f(x^*)\right) \leq \left(1 - \frac{\mu F}{4\Gamma}\right) \|\tilde{x}_k - x^*\|^2_{\hat{D}_{k-1}} - \|\tilde{x}_{k+1} - x^*\|^2_{\hat{D}_k} + \frac{3LF}{e}\|x_k - \tilde{x}_k\|^2_{\hat{D}_k}$$

for all $k \geq 0$, where $\tilde{x}_k = x_k - \frac{\beta_1\gamma}{1-\beta_1}V_{k-1}$.

Proof. Starting with the notation of virtual sequence $\{\tilde{x}_k\}$ and the update rule:

$$\begin{aligned}\tilde{x}_{k+1} &= x_{k+1} - \frac{\beta_1\gamma}{1-\beta_1}\cdot V_k = x_k - \gamma V_k - \frac{\beta_1\gamma}{1-\beta_1}V_k = x_k - \frac{(1-\beta_1)\gamma + \beta_1\gamma}{1-\beta_1}V_k \\ &= x_k - \frac{\gamma}{1-\beta_1}V_k = x_k - \frac{\gamma}{1-\beta_1}\left(\beta_1 V_{k-1} + \hat{D}_k^{-1}\nabla f(x_k)\right) \\ &= x_k - \frac{\gamma\beta_1}{1-\beta_1}V_{k-1} - \frac{\gamma}{1-\beta_1}\hat{D}_k^{-1}\nabla f(x_k) = \tilde{x}_k - \frac{\gamma}{1-\beta_1}\hat{D}_k^{-1}\nabla f(x_k).\end{aligned} \tag{8}$$

Therefore, using (8), we get

$$\begin{aligned}\|\tilde{x}_{k+1} - x^*\|^2_{\hat{D}_k} &= \|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + F^2\left\|\hat{D}_k^{-1}\nabla f(x_k)\right\|^2_{\hat{D}_k} \\ &\quad - 2F\left\langle \tilde{x}_k - x^*, \hat{D}_k\hat{D}_k^{-1}\nabla f(x_k)\right\rangle \\ &= \|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + F^2\|\nabla f(x_k)\|^2_{\hat{D}_k^{-1}} - 2F\langle x_k - x^*, \nabla f(x_k)\rangle \\ &\quad - 2F\langle \tilde{x}_k - x_k, \nabla f(x_k)\rangle.\end{aligned} \tag{9}$$

To upper bound the second term in (9), we apply Proposition 1 and Proposition 2:

$$\|\nabla f(x_k)\|^2_{\hat{D}_k^{-1}} \leq \frac{1}{e}\|\nabla f(x_k)\|_2^2 \leq \frac{2L}{e}\left(f(x_k) - f(x^*)\right). \tag{10}$$

For the third term in (9), according to Assumption 1, one can obtain

$$\begin{aligned}\langle x_k - x^*, \nabla f(x_k)\rangle &\geq f(x_k) - f(x^*) + \frac{\mu}{2}\|x_k - x^*\|_2^2 \\ &\geq f(x_k) - f(x^*) + \frac{\mu}{2\Gamma}\|x_k - x^*\|^2_{\hat{D}_k}. \end{aligned} \tag{11}$$

For the last term in (9), we use the Young's inequality (Proposition 4) with $\lambda = \frac{1}{2L}$ and Proposition 1:

$$\begin{aligned} -2F\langle \tilde{x}_k - x_k, \nabla f(x_k)\rangle &\leq 2LF\|\tilde{x}_k - x_k\|_2^2 + \frac{F}{2L}\|\nabla f(x_k)\|_2^2 \\ &\leq \frac{2LF}{e}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} + F\left(f(x_k) - f(x^*)\right). \end{aligned} \tag{12}$$

Substituting (10), (11) and (12) into (9), we obtain

$$\begin{aligned} \|\tilde{x}_{k+1} - x^*\|^2_{\hat{D}_k} &= \|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + F^2\|\nabla f(x_k)\|^2_{\hat{D}_k^{-1}} - 2F\langle \tilde{x}_k - x^*, \nabla f(x_k)\rangle \\ &- 2F\langle \tilde{x}_k - x_k, \nabla f(x_k)\rangle \\ &\leq \|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + \frac{2LF^2}{e}\left(f(x_k) - f(x_*)\right) \\ &- 2F\left(f(x_k) - f(x^*) + \frac{\mu}{2\Gamma}\|x_k - x^*\|^2_{\hat{D}_k}\right) \\ &+ \frac{2LF}{e}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} + F\left(f(x_k) - f(x^*)\right) \\ &= \|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + \left(\frac{2LF^2}{e} - F\right)\left(f(x_k) - f(x^*)\right) \\ &+ \frac{2LF}{e}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} - \frac{\mu F}{\Gamma}\|x_k - x^*\|^2_{\hat{D}_k}. \end{aligned}$$

To bound the last term, we apply one-dimensional Young's inequality (Proposition 4) with $\lambda = 1$ to obtain

$$\begin{aligned} \frac{\mu F}{2\Gamma}\|\tilde{x}_k - x^*\|^2_{\hat{D}_k} &= \frac{\mu F}{2\Gamma}\sum_{i=1}^d [\hat{D}_k]_{ii}[\tilde{x}_k - x^*]_i^2 \\ &\leq \frac{\mu F}{2\Gamma}\sum_{i=1}^d [\hat{D}_k]_{ii}\left(2[x_k - x^*]_i^2 + 2[\tilde{x}_k - x_k]_i^2\right) \\ &= \frac{\mu F}{\Gamma}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} + \frac{\mu F}{\Gamma}\|x_k - x^*\|^2_{\hat{D}_k}. \end{aligned}$$

Therefore, one can get

$$\begin{aligned}
\|\tilde{x}_{k+1} - x^*\|^2_{\hat{D}_k} &\leq \left(1 - \frac{\mu F}{2\Gamma}\right)\|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + \left(\frac{2LF^2}{e} - F\right)(f(x_k) - f(x^*)) \\
&\quad + \left(\frac{2LF}{e} + \frac{\mu F}{\Gamma}\right)\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} \\
&\leq \left(1 - \frac{\mu F}{2\Gamma}\right)\|\tilde{x}_k - x^*\|^2_{\hat{D}_k} + \left(\frac{2LF^2}{e} - F\right)(f(x_k) - f(x^*)) \\
&\quad + \frac{3LF}{e}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k},
\end{aligned}$$

where in the last inequality we applied $\frac{\mu}{\Gamma} \leq \frac{L}{e}$. Due to the bound on F we have $\frac{2LF^2}{e} \leq \frac{F}{2}$. What is more, applying Proposition 3 with $\beta_2 \geq 1 - \frac{\mu F}{4\Gamma C}$ allows to get

$$\|\tilde{x}_k - x^*\|^2_{\hat{D}_k} \leq \left(1 + \frac{\mu F}{4\Gamma}\right)\|\tilde{x}_k - x^*\|^2_{\hat{D}_{k-1}}.$$

Combining this with $(1 - x)\left(1 + \frac{x}{2}\right) \leq 1 - \frac{x}{2}$, one can obtain

$$\begin{aligned}
\|\tilde{x}_{k+1} - x^*\|^2_{\hat{D}_k} &\leq \left(1 - \frac{\mu F}{4\Gamma}\right)\|\tilde{x}_k - x^*\|^2_{\hat{D}_{k-1}} - \frac{F}{2}(f(x_k) - f(x^*)) \\
&\quad + \frac{3LF}{e}\|\tilde{x}_k - x_k\|^2_{\hat{D}_k},
\end{aligned}$$

what concludes the proof. □

Lemma 2 (Auxiliary lemma)**.** *Suppose that Assumptions 2 and 3 hold. Thus, if the sequence $\{x_k\}_{k=0}$ is produced by Algorithm 1, the following inequality is satisfied for all k:*

$$\|\tilde{x}_k - x_k\|^2_{\hat{D}_k} \leq \frac{2L\beta_1^2\gamma^2}{(1-\beta_1)^3 e}\sum_{t=0}^{k-1}\beta_1^{k-1-t}\left(1 + \frac{\mu F}{4\Gamma}\right)^{k-t}(f(x_t) - f(x^*)).$$

Proof. Let us substitute the analytical form of $\tilde{x_k}$:

$$\begin{aligned}
\|\tilde{x_k} - x_k\|^2_{\hat{D}_k} &= \left\|\frac{\beta_1\gamma}{1-\beta_1}V_{k-1}\right\|^2_{\hat{D}_k} = \left\|\frac{\beta_1\gamma}{1-\beta_1}\sum_{t=0}^{k-1}\beta_1^{k-1-t}\hat{D}_t^{-1}\nabla f(x_t)\right\|^2_{\hat{D}_k} \\
&= \frac{\beta_1^2\gamma^2}{(1-\beta_1)^2}\left(\sum_{j=0}^{k-1}\beta_1^{k-1-j}\right)^2\left\|\sum_{t=0}^{k-1}\frac{\beta_1^{k-1-t}}{\sum\limits_{j=0}^{k-1}\beta_1^{k-1-j}}\hat{D}_t^{-1}\nabla f(x_t)\right\|^2_{\hat{D}_k} \\
&\leq \frac{\beta_1^2\gamma^2}{(1-\beta_1)^2}\left(\sum_{j=0}^{k-1}\beta_1^{k-1-j}\right)\sum_{t=0}^{k-1}\beta_1^{k-1-t}\left\|\hat{D}_t^{-1}\nabla f(x_t)\right\|^2_{\hat{D}_k}, \quad (13)
\end{aligned}$$

where we apply the Jensen's inequality (Proposition 5) for the convex function $\|\cdot\|^2$. Moreover, due to Proposition 3 with $\beta_2 \geq 1 - \frac{\mu F}{4\Gamma C}$, we have

$$\begin{aligned}
\left\|\hat{D}_t^{-1}\nabla f(x_t)\right\|_{\hat{D}_k}^2 &\leq \left(1+\frac{\mu F}{4\Gamma}\right)\left\|\hat{D}_t^{-1}\nabla f(x_t)\right\|_{\hat{D}_{k-1}}^2 \leq \dots \\
&\leq \left(1+\frac{\mu F}{4\Gamma}\right)^{k-t}\left\|\hat{D}_t^{-1}\nabla f(x_t)\right\|_{\hat{D}_t}^2 \\
&= \left(1+\frac{\mu F}{4\Gamma}\right)^{k-t}\|\nabla f(x_t)\|_{\hat{D}_t^{-1}}^2 \\
&\leq \left(1+\frac{\mu F}{4\Gamma}\right)^{k-t}\frac{1}{e}\|\nabla f(x_t)\|_2^2 \\
&\leq \left(1+\frac{\mu F}{4\Gamma}\right)^{k-t}\frac{2L}{e}\left(f(x_t)-f(x^*)\right), \qquad (14)
\end{aligned}$$

where in the last inequality we applied Proposition 1 and Proposition 2. Substituting (14) into (13), we finish the proof. □

Now we are ready to formulate the theorem of convergence of Preconditioned Heavy-Ball.

Theorem 3. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations of Algorithm 1 with $\gamma = \frac{(1-\beta_1)^2 e}{12L}$, we have*

$$f\left(\frac{1}{W_{K-1}}\sum_{k=0}^{K-1} w_k x_k\right) - f(x^*) \leq 4\exp\left(-\frac{(1-\beta_1)\mu e K}{48L\Gamma}\right) L\|x_0 - x^*\|_{\hat{D}_{-1}}^2,$$

where $W_{K-1} = \sum_{k=0}^{K-1} w_k$, $w_k = \left(1-\frac{\mu F}{4\Gamma}\right)^{-(k+1)}$.

Proof. Let us denote $w_k = \left(1-\frac{\mu F}{4\Gamma}\right)^{-(k+1)}$. Therefore, summing the results from Lemma 1 with weights w_k, we get

$$\begin{aligned}
&\sum_{k=0}^{K-1}\frac{w_k F}{2}(f(x_k)-f(x^*)) \\
&\quad\leq \sum_{k=0}^{K-1} w_k\Bigg[\left(\left(1-\frac{\mu F}{4\Gamma}\right)\|\tilde{x}_k - x^*\|_{\hat{D}_{k-1}}^2 - \|\tilde{x}_{k+1}-x^*\|_{\hat{D}_k}^2\right) \\
&\quad + \frac{3LF}{e}\|\tilde{x}_k - x_k\|_{\hat{D}_k}^2\Bigg] \\
&\quad= \sum_{k=0}^{K-1}\left[w_{k-1}\|\tilde{x}_k - x^*\|_{\hat{D}_{k-1}}^2 - w_k\|\tilde{x}_{k+1}-x^*\|_{\hat{D}_k}^2 + \frac{3LFw_k}{e}\|\tilde{x}_k - x_k\|_{\hat{D}_k}^2\right].
\end{aligned}$$

The main question arises for the last term. Applying Lemma 2, one can obtain

$$\sum_{k=0}^{K-1} \frac{3LFw_k}{e} \|\tilde{x}_k - x_k\|^2_{\hat{D}_k}$$
$$\leq \frac{6L^2F\beta_1^2\gamma^2}{(1-\beta_1)^3e^2} \sum_{k=0}^{K-1}\sum_{t=0}^{k-1} w_k \beta_1^{k-1-t} \left(1 + \frac{\mu F}{4\Gamma}\right)^{k-t} (f(x_t) - f(x^*)).$$

Now we decompose w_k as

$$w_k = w_t \left(1 - \frac{\mu F}{4\Gamma}\right)^{-(k-t)} \leq w_t \left(1 + \frac{\mu F}{2\Gamma}\right)^{k-t} \leq w_t \left(1 + \frac{1-\beta_1}{2}\right)^{k-t},$$

where the last inequality holds due to the choice of γ. Consequently, using that $\beta_1 = 1 - (1-\beta_1)$ and $(1-x)\left(1+\frac{x}{2}\right) \leq \left(1 - \frac{x}{2}\right)$, we have

$$\sum_{k=0}^{K-1} \frac{3LFw_k}{e} \|\tilde{x}_k - x_k\|^2_{\hat{D}_k}$$
$$\leq \frac{6L^2F\beta_1\gamma^2}{(1-\beta_1)^3e^2} \sum_{k=0}^{K-1}\sum_{t=0}^{k-1} w_t \left(1 - \frac{1-\beta_1}{2}\right)^{k-t} \left(1 + \frac{\mu F}{4\Gamma}\right)^{k-t} (f(x_t) - f(x^*)).$$

Moreover, since $1 - \beta_1 \geq \frac{\mu F}{\Gamma}$, we get

$$\sum_{k=0}^{K-1} \frac{3LFw_k}{e} \|\tilde{x}_k - x_k\|^2_{\hat{D}_k}$$
$$\leq \frac{6L^2F\beta_1\gamma^2}{(1-\beta_1)^3e^2} \sum_{k=0}^{K-1}\sum_{t=0}^{k-1} w_t \left(1 - \frac{1-\beta_1}{4}\right)^{k-t} (f(x_t) - f(x^*)).$$

Therefore, we obtain

$$\sum_{k=0}^{K-1} \frac{w_kF}{2}(f(x_k) - f(x^*))$$
$$\leq w_{-1} \|\tilde{x}_0 - x^*\|^2_{\hat{D}_{-1}} + \frac{6L^2F\beta_1\gamma^2}{(1-\beta_1)^3e^2} \sum_{k=0}^{K-1}\sum_{t=0}^{k-1} w_t \left(1 - \frac{1-\beta_1}{4}\right)^{k-t} (f(x_t) - f(x^*)).$$

It can be easily shown that the coefficient related to $(f(x_r) - f(x^*))$ in the right-hand side can be upper bounded as

$$\frac{6L^2F\beta_1\gamma^2}{(1-\beta_1)^3e^2} \sum_{k=0}^{\infty} \left(1 - \frac{1-\beta_1}{4}\right)^k \leq \frac{24L^2F\beta_1\gamma^2}{(1-\beta_1)^4e} w_r = \frac{24L^2F^3\beta_1}{(1-\beta_1)^2e^2} w_r.$$

Applying $F \leq \frac{(1-\beta_1)e}{12L}$, we obtain

$$\frac{24L^2F^3\beta_1}{(1-\beta_1)^2e^2} \leq \frac{F}{4}.$$

As a result, one can get

$$\sum_{k=0}^{K-1} \frac{w_k F}{4}(f(x_k) - f(x^*)) \leq w_{-1} \|\tilde{x}_0 - x^*\|^2_{\hat{D}_{-1}} .$$

Dividing both sides by $W_{K-1} = \sum_{k=0}^{K-1} w_k$ and using that $\sum_{k=0}^{K-1} w_k \geq w_{K-1} \geq \exp\left(\frac{\mu F K}{4\Gamma}\right)$, we have.

$$\frac{1}{W_{K-1}} \sum_{k=0}^{K-1} w_k(f(x_k) - f(x^*)) \leq 4\exp\left(-\frac{\mu F K}{4\Gamma}\right) \|\tilde{x}_0 - x^*\|^2_{\hat{D}_{-1}} .$$

Applying the Jensen's inequality (Proposition 5) to the left-hand side and substituting the choice of γ, we conclude the proof. □

A.2. Proof of Preconditioned Nesterov Method.

Lemma 3 (Auxiliary lemma)**.** *Suppose that Assumptions 1, 2 and 3 hold. Then, if the sequence $\{x_k\}_{k=0}$ is produced by Algorithm 2, for all k and any $u \in \mathbb{R}^d$ the following inequality is satisfied:*

$$\begin{aligned} f(x_f^{k+1}) \leq f(u) - \left\langle \nabla f(x_g^k), u - x_g^k \right\rangle - \frac{\mu}{2} \left\| u - x_g^k \right\|_2^2 \\ + \left(\frac{L\gamma_k}{2e} - 1\right) \gamma_k \left\| \nabla f(x_g^k) \right\|^2_{\hat{D}_k^{-1}} . \end{aligned}$$

Proof. Start with the L-smoothness of the function f:

$$f(x_f^{k+1}) \leq f(x_g^k) + \langle \nabla f(x_g^k), x_f^{k+1} - x_g^k \rangle + \frac{L}{2} \left\| x_f^{k+1} - x_g^k \right\|_2^2 .$$

Applying Proposition 2 to the last term, we obtain

$$f(x_f^{k+1}) \leq f(x_g^k) + \left\langle \nabla f(x_g^k), x_f^{k+1} - x_g^k \right\rangle + \frac{L}{2e} \left\| x_f^{k+1} - x_g^k \right\|^2_{\hat{D}_k} . \tag{15}$$

The last term can be decomposed due to the update rule of the algorithm:

$$\left\| x_f^{k+1} - x_g^k \right\|^2_{\hat{D}_k} = \gamma_k^2 \left\| \nabla f(x_g^k) \right\|^2_{\hat{D}_k^{-1}} . \tag{16}$$

Moreover, the same update rule can be applied to the second term:

$$\left\langle \nabla f(x_g^k), x_f^{k+1} - x_g^k \right\rangle = -\gamma_k \left\| \nabla f(x_g^k) \right\|^2_{\hat{D}_k^{-1}} . \tag{17}$$

The first term can be upper bounded with the Assumption 1:

$$f(x_g^k) \leq f(u) - \left\langle \nabla f(x_g^k), u - x_g^k \right\rangle - \frac{\mu}{2} \left\| u - x_g^k \right\|_2^2 . \tag{18}$$

Substituting (16), (17) and (18) into (13) gives the final result. □

Lemma 4 (**Descent lemma**). *Suppose that Assumptions 1, 2 and 3 hold. Then, if the sequence* $\{x_k\}_{k=0}$ *is produced by Algorithm 2, for all* k *with* $\xi_k \geq 1$, $\frac{\xi_k^2 \gamma_k \mu}{\Gamma} \geq 1$, $\gamma_k \leq \frac{e}{L}$ *and* $\theta_k = \frac{\xi_k}{1+\xi_k}$*, the next inequality is satisfied:*

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|_{\hat{D}_k}^2 &+ 2\gamma_k \xi_k^2 (f(x_f^{k+1}) - f(x^*)) \\ &\leq \left(1 - \frac{1}{\xi_k}\right)\left[\left\|x^k - x^*\right\|_{\hat{D}_k}^2 + 2\gamma_k \xi_k^2 (f(x_f^k) - f(x^*))\right].\end{aligned}$$

Proof. Starting with the update rule:

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|_{\hat{D}_k}^2 &= \left\|\xi_k x_f^{k+1} + (1-\xi_k) x_f^k - x^*\right\|_{\hat{D}_k}^2 \\ &= \left\|\xi_k \left(x_g^k - \gamma_k \hat{D}_k^{-1} \nabla f(x_g^k)\right) + (1-\xi_k) x_f^k - x^*\right\|_{\hat{D}_k}^2 \\ &= \left\|\xi_k x_g^k + (1-\xi_k) x_f^k - x^*\right\|_{\hat{D}_k}^2 + \gamma_k^2 \xi_k^2 \left\|\hat{D}_k^{-1} \nabla f(x_g^k)\right\|_{\hat{D}_k}^2 \\ &- 2\gamma_k \xi_k \left\langle \hat{D}_k^{-1} \nabla f(x_g^k), \xi_k x_g^k + (1-\xi_k) x_f^k - x^* \right\rangle_{\hat{D}_k} \\ &= \left\|\xi_k x_g^k + (1-\xi_k) x_f^k - x^*\right\|_{\hat{D}_k}^2 + \gamma_k^2 \xi_k^2 \left\|\nabla f(x_g^k)\right\|_{\hat{D}_k^{-1}}^2 \\ &- 2\gamma_k \xi_k \left\langle \nabla f(x_g^k), \xi_k x_g^k + (1-\xi_k) x_f^k - x^* \right\rangle. \qquad (19)\end{aligned}$$

Let us decompose the first term. According to the update rule, we get

$$\begin{aligned}\xi_k x_g^k + (1-\xi_k) x_f^k - x^* &= \xi_k x_g^k + \frac{(1-\xi_k)}{\theta_k} \theta_k x_f^k - x^* \\ &= \xi_k x_g^k + \frac{(1-\xi_k)}{\theta_k}(x_g^k - (1-\theta_k) x^k) - x^*.\end{aligned}$$

Choosing θ_k as $\frac{\xi_k}{1+\xi_k}$, after simple estimations, one can obtain

$$\xi_k x_g^k + \frac{(1-\xi_k)}{\theta_k}(x_g^k - (1-\theta_k) x^k) - x^* = x^k - \frac{1}{\xi_k}(x^k - x_g^k).$$

Hence, the first term in (19) can be decomposed as

$$\begin{aligned}\left\|\xi_k x_g^k + (1-\xi_k) x_f^k - x^*\right\|_{\hat{D}_k}^2 &= \left\|x^k - \frac{1}{\xi_k}(x^k - x_g^k) - x^*\right\|_{\hat{D}_k}^2 \\ &= \left\|x^k - x^*\right\|_{\hat{D}_k}^2 - \frac{2}{\xi_k}\left\langle x^k - x^*, x^k - x_g^k \right\rangle_{\hat{D}_k} \\ &+ \frac{1}{\xi_k^2}\left\|x^k - x_g^k\right\|_{\hat{D}_k}^2. \qquad (20)\end{aligned}$$

Let us note that

$$-2\left\langle x^k - x^*, x_g^k - x^k \right\rangle_{\hat{D}_k} = \left\|x_g^k - x^k\right\|_{\hat{D}_k}^2 + \left\|x^k - x^*\right\|_{\hat{D}_k}^2 - \left\|x_g^k - x^*\right\|_{\hat{D}_k}^2.$$

Thus, continue with (20):

$$\begin{aligned}\left\|\xi_k x_g^k + (1-\xi_k)x_f^k - x^*\right\|^2_{\hat{D}_k} &= \left(1-\frac{1}{\xi_k}\right)\left\|x^k - x^*\right\|^2_{\hat{D}_k} \\ &\quad + \left(\frac{1}{\xi^2}-\frac{1}{\xi}\right)\left\|x^k - x_g^k\right\|^2_{\hat{D}_k} + \frac{1}{\xi_k}\left\|x_g^k - x^*\right\|^2_{\hat{D}_k}.\end{aligned} \tag{21}$$

From Lemma 3 with $u = x_f^k$ and $u = x^*$, under Proposition 2 we get

$$\begin{aligned}f(x_f^{k+1}) \leq f(x_f^k) &- \left\langle \nabla f(x_g^k), x_f^k - x_g^k\right\rangle - \frac{\mu}{2\Gamma}\left\|x_f^k - x_g^k\right\|^2_{\hat{D}_k} \\ &+ \left(\frac{L\gamma_k}{2e}-1\right)\gamma_k\left\|\nabla f(x_g^k)\right\|^2_{\hat{D}_k^{-1}}.\end{aligned} \tag{22}$$

$$\begin{aligned}f(x_f^{k+1}) \leq f(x^*) &- \left\langle \nabla f(x_g^k), x^* - x_g^k\right\rangle - \frac{\mu}{2\Gamma}\left\|x^* - x_g^k\right\|^2_{\hat{D}_k} \\ &+ \left(\frac{L\gamma_k}{2e}-1\right)\gamma_k\left\|\nabla f(x_g^k)\right\|^2_{\hat{D}_k^{-1}}.\end{aligned} \tag{23}$$

Summing (22) and (23) with multiplicative factors $2\gamma_k\xi_k(\xi_k-1)$ and $2\gamma_k\xi_k$ respectively, and substituting the result with (21) into (19), we have

$$\begin{aligned}&\left\|x^{k+1} - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k^2 f(x_f^{k+1}) \\ &\qquad\leq \left(1-\frac{1}{\xi_k}\right)\left\|x^k - x^*\right\|^2_{\hat{D}_k} + \left(\frac{1}{\xi^2}-\frac{1}{\xi}\right)\left\|x^k - x_g^k\right\|^2_{\hat{D}_k} \\ &\qquad+ \frac{1}{\xi_k}\left\|x_g^k - x^*\right\|^2_{\hat{D}_k} + \gamma_k^2\xi_k^2\left\|\nabla f(x_g^k)\right\|^2_{\hat{D}_k^{-1}} + 2\gamma_k\xi_k f(x^*) \\ &\qquad- \frac{\gamma_k\xi_k\mu}{\Gamma}\left\|x^* - x_g^k\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k(\xi_k-1)f(x_f^k) \\ &\qquad- \frac{\gamma_k\xi_k(\xi_k-1)\mu}{\Gamma}\left\|x_f^k - x_g^k\right\|^2_{\hat{D}_k} + 2\gamma_k^2\xi_k^2\left(\frac{L\gamma_k}{2e}-1\right)\left\|\nabla f(x_g^k)\right\|^2_{\hat{D}_k^{-1}}.\end{aligned}$$

Combining the terms, we have

$$\begin{aligned}&\left\|x^{k+1} - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k^2 f(x_f^{k+1}) \\ &\qquad\leq \left(1-\frac{1}{\xi_k}\right)\left\|x^k - x^*\right\|^2_{\hat{D}_k} + \left(\frac{1}{\xi^2}-\frac{1}{\xi}\right)\left\|x^k - x_g^k\right\|^2_{\hat{D}_k} \\ &\qquad+ \left(\frac{1}{\xi_k}-\frac{\gamma_k\xi_k\mu}{\Gamma}\right)\left\|x_g^k - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k f(x^*) \\ &\qquad+ 2\gamma_k\xi_k(\xi_k-1)f(x_f^k) - \frac{\gamma_k\xi_k(\xi_k-1)\mu}{\Gamma}\left\|x_f^k - x_g^k\right\|^2_{\hat{D}_k} \\ &\qquad+ \gamma_k^2\xi_k^2\left(\frac{L\gamma_k}{e}-1\right)\left\|\nabla f(x_g^k)\right\|^2_{\hat{D}_k^{-1}},\end{aligned}$$

where the scalar products are eliminated by multiplicative factors mentioned before. Subtracting $2\gamma_k\xi_k^2 f(x^*)$ from both sides, and using that $\xi_k \geq 1$, $\frac{\xi_k^2\gamma_k\mu}{\Gamma} \geq 1$ and $\gamma_k \leq \frac{e}{L}$, we obtain

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2_{\hat{D}_k} &+ 2\gamma_k\xi_k^2(f(x_f^{k+1}) - f(x^*)) \\ &\leq \left(1 - \frac{1}{\xi_k}\right)\left\|x^k - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k(\xi_k - 1)(f(x_f^k) - f(x^*)) \\ &= \left(1 - \frac{1}{\xi_k}\right)\left[\left\|x^k - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k^2(f(x_f^k) - f(x^*))\right],\end{aligned}$$

what finishes the proof. □

To obtain the convergence, let us formulate a final theorem.

Theorem 4. *Suppose that Assumptions 1, 2 and 3 hold. Then, after K iterations of Algorithm 2 with $\gamma_k \equiv \frac{e}{L}, \xi_k \equiv \sqrt{\frac{L\Gamma}{\mu e}}$ and $\theta_k \equiv \frac{\sqrt{L\Gamma}}{\sqrt{\mu e}+\sqrt{L\Gamma}}$, we have*

$$\|x^{k+1} - x^*\|^2_{\hat{D}_k} \leq \exp\left(-K\sqrt{\frac{\mu e}{4L\Gamma}}\right)\left[\|x^0 - x^*\|^2_{\hat{D}_{-1}} + \frac{2\Gamma}{\mu}(f(x^0) - f(x^*))\right].$$

Proof. Starting with Lemma 4:

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2_{\hat{D}_k} &+ 2\gamma_k\xi_k^2(f(x_f^{k+1}) - f(x^*)) \\ &\leq \left(1 - \frac{1}{\xi_k}\right)\left[\left\|x^k - x^*\right\|^2_{\hat{D}_k} + 2\gamma_k\xi_k^2(f(x_f^k) - f(x^*))\right].\end{aligned}$$

We use the Proposition 3 with $\beta_2 \geq 1 - \frac{1}{2C}\sqrt{\frac{\mu e}{L\Gamma}}$ such that

$$\left\|x^k - x^*\right\|^2_{\hat{D}_k} \leq \left(1 + \frac{1}{2\xi_k}\right)\left\|x^k - x^*\right\|^2_{\hat{D}_{k-1}}.$$

Consequently, with $(1 - x)\left(1 + \frac{x}{2}\right) \leq \left(1 - \frac{x}{2}\right)$ we get

$$\begin{aligned}\left\|x^{k+1} - x^*\right\|^2_{\hat{D}_k} &+ 2\gamma_k\xi_k^2(f(x_f^{k+1}) - f(x^*)) \\ &\leq \left(1 - \frac{1}{2\xi_k}\right)\left[\left\|x^k - x^*\right\|^2_{\hat{D}_{k-1}} + 2\gamma_k\xi_k^2(f(x_f^k) - f(x^*))\right].\end{aligned}$$

After substituting γ_k, ξ_k and θ_k, with $(1 - x) \leq \exp(-x)$, the recursion provides a final bound. □

Remark 2. *For Theorem 2 it is enough to apply Proposition 2 to obtain the convergence in L_2-space.*

Stepan Denisovich Trifonov
Moscow Institute of Physics and Technology,
9 Institutskiy Per.,
141701, Dolgoprudny, Russia
Email address: trifonov.sd@phystech.edu

Leonid Ilyich Levin
Moscow Institute of Physics and Technology,
9 Institutskiy Per.,
141701, Dolgoprudny, Russia
Email address: levin.li@phystech.edu

Savelii Andreevich Chezhegov
Moscow Institute of Physics and Technology, ISP RAS
9 Institutskiy Per.,
141701, Dolgoprudny, Russia
Email address: chezhegov.sa@phystech.edu

Aleksandr Nikolaevich Beznosikov
Moscow Institute of Physics and Technology, ISP RAS, Innopolis University
9 Institutskiy Per.,
141701, Dolgoprudny, Russia
Email address: beznosikov.an@phystech.edu