Another approach to build Lyapunov functions for the first order methods in the quadratic case

Daniil Merkulov*†
d.merkulov@skoltech.ru

Ivan Oseledets‡*
i.oseledets@skoltech.ru

Abstract

Lyapunov functions play a fundamental role in analyzing the stability and convergence properties of optimization methods. In this paper, we propose a novel and straightforward approach for constructing Lyapunov functions for first-order methods applied to quadratic functions. Our approach involves bringing the iteration matrix to an upper triangular form using Schur decomposition, then examining the value of the last coordinate of the state vector. This value is multiplied by a magnitude smaller than one at each iteration. Consequently, this value should decrease at each iteration, provided that the method converges. We rigorously prove the suitability of this Lyapunov function for all first-order methods and derive the necessary conditions for the proposed function to decrease monotonically. Experiments conducted with general convex functions are also presented, alongside a study on the limitations of the proposed approach.

Remarkably, the newly discovered Lyapunov function is straightforward and does not explicitly depend on the exact method formulation or function characteristics like strong convexity or smoothness constants. In essence, a single expression serves as a Lyapunov function for several methods, including Heavy Ball, Nesterov Accelerated Gradient, and Triple Momentum, among others. To the best of our knowledge, this approach has not been previously reported in the literature.

1 Introduction

Consider the following quadratic optimization problem:

$$\min_{x \in \mathbb{R}^d} f(x) = \min_{x \in \mathbb{R}^d} \frac{1}{2} x^T W x - b^T x + c, \text{ where } W \in \mathbb{S}^d_{++} \tag{1}$$

The problem itself is equivalent to solving a linear system with a positive definite matrix. However, this problem is not the main object of study. In the example of solving this problem, we want to study a specific way of constructing the *Lyapunov function* for the optimization methods. It is quite common for the convergence of accelerated first-order methods to be non-monotonic, even for quadratic objectives [4, 3] (see the numerical example in Figure **??**). These methods include the Polyak Heavy Ball method [12], Nesterov's accelerated gradient method [9], the triple momentum method [14], and others. It is important to note that these methods are widely used in modern neural network training, including in the currently

*Skolkovo Institute of Science and Technology
†Moscow Institute of Physics and Technology
‡Artificial Intelligence Research Institute

popular large language models. The quadratic problem appears not only in usual linear least squares problems but also in consensus search problems in decentralized systems [5]. Therefore, studying their convergence is of great significance.

Lyapunov functions have proven to be invaluable tools for understanding the stability properties of optimization algorithms. They offer a systematic approach for assessing convergence, boundedness, and even optimality in the iterative optimization process. These functions capture the behavior of an algorithm by assigning a scalar quantity to each iteration, enabling a broader understanding of the underlying dynamics and convergence behavior. One may think of them as energy functions that should not increase during the optimization trajectory of a given method. The main objective of this paper is to present a simple approach for constructing Lyapunov functions for such methods. For example, the following function will monotonically decrease for the Heavy Ball and Nesterov Accelerated Gradient methods applied to (**??**) (assuming that x^* is the solution and optimal hyperparameters were chosen):

$$V(x_k, x_{k-1}, x_{k-2}) = \|x_{k-1} - x^*\|^2 - \langle x_k - x^*, x_{k-2} - x^* \rangle \tag{2}$$

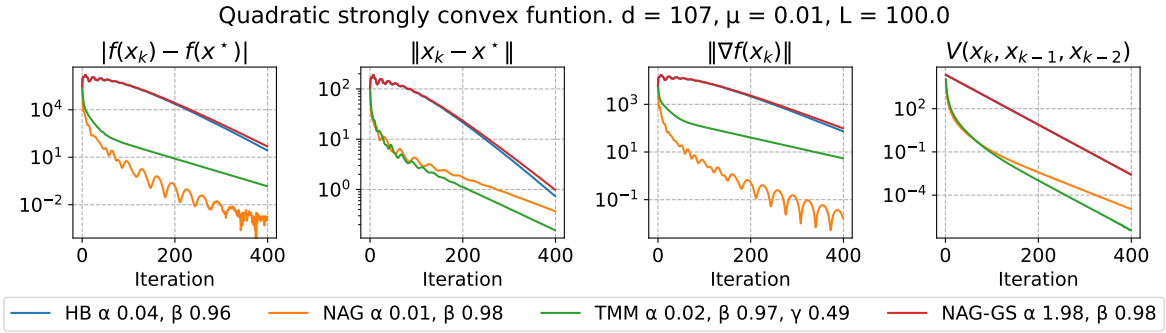


Figure 1: Dynamics of several accelerated methods with optimal hyperparameters α, β, γ applied to the general strongly convex quadratic problem (**??**) with dimension $d = 107$ are presented. Here and later, we use the following notation: **HB** - Heavy Ball [12], **NAG** - Nesterov Accelerated Gradient [8], **TMM** - Triple Momentum Method [14], **NAG-GS** - Nesterov Accelerated Gradient with Gauss-Seidel Splitting [7]. It is easy to see that the usual metrics on the first three subfigures on the left do not demonstrate a monotonic decrease, while the proposed Lyapunov function $V(x_k, x_{k-1}, x_{k-2})$ from (**??**) works for all these methods.

Recent advances in discovering Lyapunov functions can be found in the book [15], which covers a much more general setup of minimizing general (possibly non-quadratic) μ-strongly convex functions. Sometimes, in the context of accelerated methods, the Lyapunov function is also referred to as the *potential* function [1], especially when the problem is convex but not strongly convex. Important results have been obtained regarding the provably fastest Lyapunov functions [13] over a parameterized family of functions (called Lyapunov function candidates) for general strongly convex smooth functions. The parametric class of Lyapunov functions considered were quadratic functions. The class of minimized functions was μ-strongly convex with L-Lipschitz gradient. However, constructing such functions involves solving a small but rather complex SDP problem. A fair comparison of the functions from [13] and a function presented in the current work is a topic for further research.

2 Illustrative example: Heavy Ball method

If we have a first-order method applied to a quadratic function, which operates on the current point x_k (note that the gradient at the current point is also expressed in this term $\nabla f(x_k) = Wx_k - b$) and the previous point x_{k-1}, it can be formulated as follows:

$$z_{k+1} = Mz_k, \tag{3}$$

where z_k is a state vector (for example, $z_k = (x_k, x_{k-1})$). The system of equation (**??**) represents a linear dynamical system. The idea behind constructing a Lyapunov function involves creating a positive quantity that decreases along the trajectories of the dynamical system.

The presence of such a function ensures the convergence of the dynamical system. It should be noted that for the optimization problem of a function of general form (non-quadratic), it is not possible to write down the iteration of the method with the help of a dynamic linear system.

Consider the Heavy Ball method [12].

$$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1}). \tag{4}$$

It is well-known [12, 2] that α and β can be selected in such a way that the Heavy Ball method achieves global convergence. This can be demonstrated by finding a suitable *Lyapunov function* $V(x_k, x_{k-1})$ that decays along the iterations.

Our goal is to develop a systematic way to propose candidates for Lyapunov functions for the method (**??**). Table **??** illustrates how different methods can be expressed in this way and how the iteration matrix M appears in each particular case. The idea can be applied to any standard optimization method, but we will illustrate how it works for the Polyak method and later generalize the idea to any method presented in the form of (**??**) under some conditions on the iteration matrix M.

2.1 Reduction to a scalar case

For the sake of clarity, we will start with the case when $f(x)$ is a strongly convex quadratic function with the minimum at $x^\star = 0$, thus it could be written as

$$\begin{aligned} f(x) =& \frac{1}{2}\langle Wx, x\rangle \to \min_{x\in\mathbb{R}^d}, \\ & W \in \mathbb{S}^d_{++}. \end{aligned} \tag{5}$$

Where $W \in \mathbb{S}^d_{++}$ means, that matrix W is symmetric positive definite (SPD). Moreover, we will assume the following characteristics about it: $0 < \mu = \lambda_{min}(W)$; $\lambda_{max}(W) = L$. Note, that μ is a strong convexity constant, while L - is the Lipschitz constant for the gradient of the function.

Since W is an SPD matrix, it has the eigendecomposition $W = Q\Lambda Q^*$, where Q is orthogonal and Λ is diagonal. If we assign new variables

$$\hat{x}_k = Q^* x_k. \tag{6}$$

The function will look like:

$$f(\hat{x}) = \frac{1}{2}\langle Wx, x\rangle = \frac{1}{2}\langle Q\Lambda Q^* x, x\rangle = \frac{1}{2}\langle \Lambda Q^* x, Q^* x\rangle = \frac{1}{2}\langle \Lambda \hat{x}, \hat{x}\rangle \tag{7}$$

Taking into account, that the gradient is $\nabla f(\hat{x}) = \Lambda \hat{x}$, the method (**??**) will be written as follows

$$\hat{x}_{k+1} = \hat{x}_k - \alpha\Lambda\hat{x}_k + \beta(\hat{x}_k - \hat{x}_{k-1}) \qquad \hat{x}_{k+1} = (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta\hat{x}_{k-1}. \tag{8}$$

We can now use the common reformulation:

$$\begin{cases} \hat{x}_{k+1} &= (I - \alpha\Lambda + \beta I)\hat{x}_k - \beta\hat{x}_{k-1} \\ \hat{x}_k &= \hat{x}_k, \end{cases} \tag{9}$$

Let's use the following notation $\hat{z}_k = \begin{bmatrix} \hat{x}_{k+1} \\ \hat{x}_k \end{bmatrix}$. Therefore $\hat{z}_{k+1} = M\hat{z}_k$, where the iteration matrix M is:

$$M = \begin{bmatrix} I - \alpha\Lambda + \beta I & -\beta I \\ I & 0_d \end{bmatrix} \tag{10}$$

Note, that M is $2d \times 2d$ matrix with 4 block-diagonal matrices of size $d \times d$ inside. It means, that we can rearrange the order of coordinates to make M block-diagonal in the following form (see Figure **??**). Note that in the equation below, the matrix M denotes the same as in the notation above, except for the described permutation of rows and columns. We use this slight abuse of notation for the sake of clarity. where $\hat{x}_k^{(i)}$ is i-th coordinate of vector $\hat{x}_k \in \mathbb{R}^d$ and

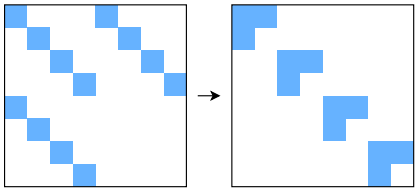


Figure 2: Illustration of matrix M rearrangement

$$\begin{bmatrix} \hat{x}_k^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \to \begin{bmatrix} \hat{x}_k^{(1)} \\ \hat{x}_{k-1}^{(1)} \\ \vdots \\ \hat{x}_k^{(d)} \\ \hat{x}_{k-1}^{(d)} \end{bmatrix} \qquad M = \begin{bmatrix} M_1 & & & \\ & M_2 & & \\ & & \dots & \\ & & & M_d \end{bmatrix} \tag{11}$$

M_i stands for 2×2 matrix. This rearrangement allows us to study the dynamics of the (**??**) independently for each dimension. One may observe, that the asymptotic convergence rate of the $2d$-dimensional vector sequence of $\hat{z}_k$ is defined by the worst convergence rate among its block of coordinates. Thus, it is enough to study the optimization in a one-dimensional case. For i-th coordinate with λ_i as an i-th eigenvalue of matrix W we have:

$$M_i = \begin{bmatrix} 1 - \alpha\lambda_i + \beta & -\beta \\ 1 & 0 \end{bmatrix} \tag{12}$$

2.2 Idea

From this moment we will consider the scalar case. Here we have the problem

$$f(x) = \frac{\lambda}{2}x^2 \to \min_{x \in \mathbb{R}^1}, \quad \lambda > 0 \tag{13}$$

It is worth mentioning, that we still keep in mind the original problem (**??**), when λ is some eigenvalue of initial matrix W, which means, that $0 < \mu \leq \lambda \leq L$. The iteration takes the form

$$x_{k+1} = x_k - \alpha\lambda x_k + \beta(x_k - x_{k-1}) \qquad \begin{cases} x_{k+1} &= (1 - \alpha\lambda)x_k - \beta x_{k-1} \\ x_k &= x_k, \end{cases} \tag{14}$$

which can be rewritten in the matrix form as

$$z_{k+1} = Mz_k, \quad M = \begin{bmatrix} 1 - \alpha\lambda + \beta & -\beta \\ 1 & 0 \end{bmatrix}, \quad z_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \tag{15}$$

The method will be convergent if $\rho(M) < 1$, and the optimal parameters can be computed by optimizing the spectral radius [12]

$$\alpha^*, \beta^* = \arg\min_{\alpha,\beta} \max_{\lambda \in [\mu, L]} \rho(M) \quad \alpha^* = \frac{4}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \beta^* = \left(\frac{\sqrt{L} - \sqrt{\mu}}{\sqrt{L} + \sqrt{\mu}} \right)^2 \tag{16}$$

It can be shown (**??**) that for such parameters the matrix M has complex eigenvalues, which forms a conjugate pair, so the distance to the optimum (in this case, $\|z_k\|$), generally, will not go to zero monotonically. One can use the machinery of the matrix Lyapunov equation [6] to derive the candidates for the Lyapunov function, which will be quadratic. However, there is a simpler way.

2.2.1 Lyapunov function formulation

Consider Schur decomposition of the matrix M :

$$M = UTU^*; \quad U^*U = I; \quad T = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} \tag{17}$$

where U is a unitary matrix, and T is a complex upper triangular matrix. If we manage to find T, we will have an easy iteration analysis. From the (**??**)

$$\begin{aligned} z_{k+1} &= Mz_k = UTU^*z_k \\ U^*z_{k+1} &= TU^*z_k \\ w_{k+1} &= Tw_k, \end{aligned} \tag{18}$$

where the substitution $w_k = U^*z_k$ was introduced. Since T is upper triangular, the last element of the vector w_k is just multiplied by the same number t_{22} at each iteration:

$$\begin{bmatrix} (w_{k+1})_1 \\ (w_{k+1})_2 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix} \begin{bmatrix} (w_k)_1 \\ (w_k)_2 \end{bmatrix} = \begin{bmatrix} t_{11} & t_{12} \\ 0 & t_{22} \end{bmatrix}^k \begin{bmatrix} (w_0)_1 \\ (w_0)_2 \end{bmatrix} \rightarrow (w_{k+1})_2 = (t_{22})^k (w_0)_2 \tag{19}$$

In equations (**??**), (**??**) the term $(m)_i$ denotes the i-th coordinate of the vector m. In this form, it's obvious, that if the method converges, the absolute value of t_{22} is lower than 1. Therefore, we can pick the absolute value of the w_k as the Lyapunov function, because it is the converging geometric progression with the convergence rate $|t_{22}|$. Since the diagonal elements t_{11}, t_{22} are the eigenvalues of the matrix M, the rate of convergence is the spectral radius of the iteration matrix, which makes the proposed Lyapunov function asymptotically optimal for the particular method.

$$V(x_k, x_{k-1}) = |(w_k)_2|^2 = |(U^*z_k)_2|^2 = \left| \left(U^* \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \right)_2 \right|^2 \tag{20}$$

2.2.2 Explicit Schur decomposition of the iteration matrix M

The diagonal elements of the matrix T are the eigenvalues of M. At the same time, the first column of the matrix U is the eigenvector of the matrix M. Let us obtain the expression for it. It is easy to verify, that the eigenvalues of the iteration matrix are

$$t_{11}, t_{22} = \lambda_1^M, \lambda_2^M = \lambda\left(\begin{bmatrix} 1-\alpha\lambda+\beta & -\beta \\ 1 & 0 \end{bmatrix}\right) = \frac{1+\beta-\alpha\lambda \pm \sqrt{(1+\beta-\alpha\lambda)^2-4\beta}}{2} \tag{21}$$

If we'll use the optimal values α^*, β^* from (**??**), assuming $\mu > 0$ and $L > 0$, we have:

$$t_{11}, t_{22} = \frac{\mu + L - 2\lambda \pm 2\sqrt{(L-\lambda)(\mu-\lambda)}}{(\sqrt{L}+\sqrt{\mu})^2} \tag{22}$$

Let us also verify, that the vector $(\lambda^M \ 1)^T$ will be the (unnormalized) eigenvector for iteration matrix M. Here we denote λ^M as any eigenvalue of matrix M (either λ_1^M or λ_2^M). We have

$$\begin{bmatrix} 1-\alpha\lambda+\beta & -\beta \\ 1 & 0 \end{bmatrix}\begin{bmatrix} \lambda^M \\ 1 \end{bmatrix} = \lambda^M \begin{bmatrix} \lambda^M \\ 1 \end{bmatrix} \to \lambda^M = \frac{1+\beta-\alpha\lambda \pm \sqrt{(1+\beta-\alpha\lambda)^2-4\beta}}{2}$$

To build matrix U from (**??**) we can take vector u_1 as an eigenvector of M:

$$U = \begin{bmatrix} u_1 \ u_2 \end{bmatrix} \quad u_1 \ = \ \frac{1}{\sqrt{1+(\lambda^M)^*\lambda^M}}\begin{bmatrix} \lambda^M \\ 1 \end{bmatrix}, \tag{23}$$

while the second vector u_2 can be taken as an orthogonal vector as $u_2 \sim \begin{bmatrix} 1 \\ -(\lambda^M)^* \end{bmatrix}$. Thus, we can write down the matrix U:

$$U = \frac{1}{\sqrt{1+(\lambda^M)^*\lambda^M}}\begin{bmatrix} \lambda^M & 1 \\ 1 & -(\lambda^M)^* \end{bmatrix} \tag{24}$$

Note, that if the eigenvalues λ_1^M, λ_2^M is not a conjugate pair, we couldn't write down the expression (**??**). The term *conjugate pair* refers to either complex eigenvalues, which satisfy $(\lambda_1^M)^* = \lambda_2^M$ or real equal eigenvalues $\lambda_1^M = \lambda_2^M$. Taking into account, that eigenvalues of M is a conjugate pair, the iteration matrix will take the form

$$M = \underbrace{\frac{1}{\sqrt{1+\lambda_1^M\lambda_2^M}}\begin{bmatrix} \lambda_1^M & 1 \\ 1 & -\lambda_2^M \end{bmatrix}}_{U}\underbrace{\begin{bmatrix} \lambda_1^M & * \\ 0 & \lambda_2^M \end{bmatrix}}_{T}\underbrace{\frac{1}{\sqrt{1+\lambda_1^M\lambda_2^M}}\begin{bmatrix} \lambda_2^M & 1 \\ 1 & -\lambda_1^M \end{bmatrix}}_{U^*} \tag{25}$$

The diagonal elements of the matrix T are the eigenvalues of M, so $t_{22} = \lambda^M$. Returning

to (**??**), we have:

$$
\begin{aligned}
V(x_k, x_{k-1}) &= \left| \frac{1}{\sqrt{1+\lambda_1^M \lambda_2^M}} \begin{bmatrix} 1 & -\lambda_1^M \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \right|^2 \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left| x_k - \lambda_1^M x_{k-1} \right|^2 \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(\mathrm{Re}^2(x_k - \lambda_1^M x_{k-1}) + \mathrm{Im}^2(x_k - \lambda_1^M x_{k-1}) \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(\left(x_k - \mathrm{Re}(\lambda_1^M) x_{k-1} \right)^2 + \left(\mathrm{Im}(\lambda_1^M) x_{k-1} \right)^2 \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(x_k^2 - 2\mathrm{Re}(\lambda_1^M) x_k x_{k-1} + \mathrm{Re}^2(\lambda_1^M) x_{k-1}^2 + \mathrm{Im}^2(\lambda_1^M) x_{k-1}^2 \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(x_k^2 - 2\mathrm{Re}(\lambda_1^M) x_k x_{k-1} + |\lambda_1^M|^2 x_{k-1}^2 \right)
\end{aligned}
\tag{26}
$$

2.2.3 Optimal hyperparameters for the method and the spectrum of the iteration matrix

Now we'll consider the eigenvalues of the iteration matrix M. We'll start with the optimal hyperparameters α^*, β^* from (**??**):

$$
\mathrm{Re}(\lambda_1^M) = \frac{L + \mu - 2\lambda}{(\sqrt{L} + \sqrt{\mu})^2}; \quad \mathrm{Im}(\lambda_1^M) = \frac{\pm 2\sqrt{(L-\lambda)(\lambda-\mu)}}{(\sqrt{L} + \sqrt{\mu})^2}; \tag{27}
$$

$$
|\lambda| = \frac{L - \mu}{(\sqrt{L} + \sqrt{\mu})^2} \quad |\lambda|^2 = \frac{(L-\mu)^2}{(\sqrt{L} + \sqrt{\mu})^4} \tag{28}
$$

$$
\begin{aligned}
V(x_k, x_{k-1}) &= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(x_k^2 - 2\frac{L + \mu - 2\lambda}{(\sqrt{L} + \sqrt{\mu})^2} x_k x_{k-1} + \frac{(L-\mu)^2}{(\sqrt{L} + \sqrt{\mu})^4} x_{k-1}^2 \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(\left(x_k - 2\frac{L + \mu - 2\lambda}{(\sqrt{L} + \sqrt{\mu})^2} x_{k-1} \right) x_k + \frac{(L-\mu)^2}{(\sqrt{L} + \sqrt{\mu})^4} x_{k-1}^2 \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \left(-\frac{(\sqrt{L} - \sqrt{\mu})^2}{(\sqrt{L} + \sqrt{\mu})^2} x_{k-2} x_k + \frac{(\sqrt{L} + \sqrt{\mu})^2 (\sqrt{L} - \sqrt{\mu})^2}{(\sqrt{L} + \sqrt{\mu})^4} x_{k-1}^2 \right) \\
&= \frac{1}{1+\lambda_1^M \lambda_2^M} \frac{(\sqrt{L} - \sqrt{\mu})^2}{(\sqrt{L} + \sqrt{\mu})^2} \left(x_{k-1}^2 - x_k x_{k-2} \right)
\end{aligned}
\tag{29}
$$

Overall, we have the simpler formula for the Lyapunov function in this setting:

$$
V(x_k, x_{k-1}) = x_{k-1}^2 - x_k x_{k-2} \tag{30}
$$

Let us highlight what was shown at the moment. We considered the problem (**??**) and the heavy ball method (**??**) applied to it. It was shown, that the dynamics happen independently on each coordinate with its iteration matrix M_i (which was denoted in this section as M for simplicity) (**??**) for each dimension. Then, we brought the iteration matrix M_i to the upper

triangular form using Schur decomposition. We demonstrated, that for each coordinate the proposed expression (**??**) is monotonically decreasing during the iteration procedure under the condition, that λ_1^M and λ_2^M are conjugate pair.

It can be shown (see Figure **??**), that for **HB**, **NAG** and **NAG-GS** with optimal parameters we have a spectrum, where for each dimension eigenvalues are conjugate pairs, while for the **TMM** this is not true, which means, that generally there is no guarantee, that proposed function (**??**) will serve as a Lyapunov function for this method. However, occasionally, sometimes it works even in this case, but we will show experiments, where $V(x_k, x_{k-1}, x_{k-2})$ will not monotonically decrease for **TMM** even in quadratic case.

However, optimal hyperparameter setting requires the knowledge of μ and L, which is not always possible in practice. Therefore, it is even more interesting, that we can derive convergence properties straightforward from the spectrum of the iteration matrix and verify, that some set of hyperparameters ensures proposed $V(x_k, x_{k-1}, x_{k-2})$ to be the Lyapunov function (see Figure **??**).

Method	a	b	$\rho(M)$	λ_1^M and λ_2^M are the conjugate pairs if
HB [12] with $\alpha^\star, \beta^\star$	$\frac{2(L+\mu-2\lambda)}{(\sqrt{L}+\sqrt{\mu})^2}$	$-\frac{(\sqrt{L}-\sqrt{\mu})^2}{(\sqrt{L}+\sqrt{\mu})^2}$	$\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	always
NAG[10] with $\alpha^\star, \beta^\star$	$\frac{2\sqrt{L}(L-\lambda)}{(\sqrt{L}+\sqrt{\mu})L}$	$-\frac{(\sqrt{L}-\sqrt{\mu})(L-\lambda)}{(\sqrt{L}+\sqrt{\mu})L}$	$1-\sqrt{\frac{\mu}{L}}$	always
TMM[14] with $\alpha^\star, \beta^\star, \gamma^\star$	$\frac{2L+\mu-3\lambda-(L-\lambda)\sqrt{\frac{\mu}{L}}}{L(1+\sqrt{\frac{\mu}{L}})}$	$-\frac{(\sqrt{L}-\sqrt{\mu})^2(L-\lambda)}{(\sqrt{L}+\sqrt{\mu})L}$	$\left(1-\sqrt{\frac{\mu}{L}}\right)^{\frac{3}{2}}$	not $\forall\lambda$
NAG-GS[7] with $\alpha^\star, \beta^\star$	$\frac{4\left(\mu+\sqrt{\mu L}\right)\left(-\lambda\left(\sqrt{\frac{L}{\mu}}+1\right)+\sqrt{\mu L}+L\right)}{\left(\mu+2\sqrt{\mu L}+L\right)^2}$	$-\frac{(L-\mu)^2}{(L+\mu+2\sqrt{\mu L})^2}$	$\frac{L-\mu}{L+\mu+2\sqrt{\mu L}}$	always

Table 1: Reformulation of first order methods with optimal parameters in the format given in theorem **??**

3 Lyapunov function for first-order methods for quadratic function

3.1 Scalar case

As soon as we formulate the result for the Heavy Ball method with optimal hyperparameters, we can generalize the idea to an arbitrary two-step method, which is convergent and has a conjugate pair of eigenvalues of the iteration matrix. This result is formulated in the following theorem.

Theorem 3.1. *For the quadratic optimization problem in the form of*

$$f(x) = \frac{\lambda}{2}x^2 \to \min_{x\in\mathbb{R}^1}, \quad \lambda > 0$$

Given any convergent optimization method, which could be written in the following form

$$x_{k+1} = ax_k + bx_{k-1} \tag{31}$$

, where $a^2 + 4b \leq 0$ it has the following Lyapunov function:

$$V(x_k, x_{k-1}, x_{k-2}) = x_{k-1}^2 - x_k x_{k-2}$$

Method	Iteration
HB [12]	$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1})$ $\mathbf{x_{k+1}} = (1+\beta-\alpha\lambda)\mathbf{x_k} - \beta\mathbf{x_{k-1}}$
NAG[8]	$\begin{cases} y_{k+1} = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha\nabla f(y_{k+1}) \end{cases}$ $\mathbf{x_{k+1}} = (1+\beta)(1-\alpha\lambda)\mathbf{x_k} - \beta(1-\alpha\lambda)\mathbf{x_{k-1}}$
TMM[14]	$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha\nabla f((1+\gamma)x_k - \gamma x_{k-1})$ $\mathbf{x_{k+1}} = (1+\beta-\alpha(1+\gamma)\lambda)\mathbf{x_k} + (\alpha\gamma\lambda - \beta)\mathbf{x_{k-1}}$
NAG-GS[7]	$\begin{cases} y_k = \beta y_{k-1} + (1-\beta)x_k - \alpha\nabla f(x_k) \\ x_{k+1} = \beta x_k + (1-\beta)y_k \end{cases}$ $\mathbf{x_{k+1}} = \left((2\beta + (1-\beta)^2) - \alpha(1-\beta)\lambda\right)\mathbf{x_k} - \beta^2\mathbf{x_{k-1}}$

Table 2: Correspondence between accelerated first-order methods and two-step notation given in (**??**). Notation $\mathbf{x_k}$ made only for the sake of clarity.

Proof. 1. Clearly, the method could be written in the form:

$$\begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} = M \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}, \qquad M = \begin{bmatrix} a & b \\ 1 & 0 \end{bmatrix}$$

While the eigenpairs are $\lambda^M = \lambda_1^M, \lambda_2^M$ or $\lambda^M = \dfrac{a \pm \sqrt{a^2+4b}}{2}$ and $v^M = \begin{bmatrix} \lambda^M \\ 1 \end{bmatrix}$

2. Thus, we can explicitly construct Schur decomposition:

$$M = \underbrace{\frac{1}{\sqrt{1+\left(\lambda_1^M\right)^*\lambda_1^M}} \begin{bmatrix} \lambda_1^M & 1 \\ 1 & -\left(\lambda_1^M\right)^* \end{bmatrix}}_{U} \underbrace{\begin{bmatrix} \lambda_1^M & * \\ 0 & \lambda_2^M \end{bmatrix}}_{T} \underbrace{\frac{1}{\sqrt{1+\left(\lambda_1^M\right)^*\lambda_1^M}} \begin{bmatrix} \left(\lambda_1^M\right)^* & 1 \\ 1 & -\lambda_1^M \end{bmatrix}}_{U^*} \tag{32}$$

Taking into account convergence condition: $\rho(M) < 1$, i.e. $\max\left(|\lambda^M|\right) < 1$ and the imaginary spectrum of the iteration matrix $a^2 + 4b \leq 0$, we will write down explicitly:

- $\left(\lambda_1^M\right)^* = \lambda_2^M$
- $|\lambda^M| = \sqrt{-b}$
- $|\lambda^M|^2 = -b$
- $\rho(M) = \sqrt{-b}$
- $\mathrm{Re}(\lambda^M) = \frac{a}{2}$
- $\mathrm{Im}(\lambda^M) = \frac{\sqrt{-a^2-4b}}{2}$
- Similarly to (**??**) we can say, that in the new variables (w_k), the absolute value of the last coordinate will monotonically decrease. Using the reformulation $z_{k+1} = Mz_k$ $w_k = U^*z_k$; $w_{k+1} = Tw_k$ and taking into account, that $\left(\lambda_1^M\right)^* = \lambda_2^M$ the

Method	a	b	$\rho(M)$ if λ_1^M and λ_2^M are conjugate pairs
HB [12] with α, β	$1 - \alpha\lambda + \beta$	$-\beta$	$\sqrt{\beta}$
NAG[10] with α, β	$(1 - \alpha\lambda)(1 + \beta)$	$-(1 - \alpha\lambda)\beta$	$\sqrt{(1 - \alpha\mu)\beta}$
TMM[14] with α, β, γ	$(1 + \beta - \alpha(1 + \gamma)\lambda)$	$(\alpha\gamma\lambda - \beta)$	$\sqrt{\beta - \alpha\gamma\mu}$
NAG-GS[7] with α, β	$2\beta + (1 - \beta)^2 - \alpha(1 - \beta)\lambda$	$-\beta^2$	β

Table 3: Reformulation of first-order methods with general parameters in the format given in theorem **??**

Lyapunov function will take the following form:

$$\begin{aligned} V(\cdot) = |(w_k)_2|^2 = |(U^* z_k)_2|^2 &= \left| \left(U^* \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} \right)_2 \right|^2 \\ &= \frac{1}{1 + \lambda_1^M \lambda_2^M} \left| x_k - \lambda_1^M x_{k-1} \right|^2 \\ &\simeq \mathrm{Re}^2(x_k - \lambda_1^M x_{k-1}) + \mathrm{Im}^2(x_k - \lambda_1^M x_{k-1}) \\ &= \left(x_k - \mathrm{Re}(\lambda_1^M) x_{k-1} \right)^2 + \left(\mathrm{Im}(\lambda_1^M) x_{k-1} \right)^2 \\ &= x_k^2 - 2\mathrm{Re}(\lambda_1^M) x_k x_{k-1} + \mathrm{Re}^2(\lambda_1^M) x_{k-1}^2 + \mathrm{Im}^2(\lambda_1^M) x_{k-1}^2 \\ &= x_k^2 - 2\mathrm{Re}(\lambda_1^M) x_k x_{k-1} + |\lambda_1^M|^2 x_{k-1}^2 \end{aligned} \tag{33}$$

- As soon as $x_k = a x_{k-1} + b x_{k-2}$, we can write the Lyapunov function

$$\begin{aligned} V(\cdot) &= x_k^2 - 2\mathrm{Re}(\lambda_1^M) x_k x_{k-1} + |\lambda_1^M|^2 x_{k-1}^2 \\ &= x_k \left(x_k - a x_{k-1} \right) + |\lambda_1^M|^2 x_{k-1}^2 \\ &= x_k b x_{k-2} - b x_{k-1}^2 \\ &\simeq x_{k-1}^2 - x_k x_{k-2} \end{aligned} \tag{34}$$

We used the $\simeq$ symbol above to denote equivalence from the Lyapunov function point of view between function $V(\cdot)$ and $aV(\cdot)$, $a > 0$. So, $V(\cdot) \simeq aV(\cdot)$

□

It is interesting, that the expression above serves as the Lyapunov function for the very wide class of methods. Several methods, which allow two-step reformulation like in (**??**) are presented in Table **??**

Therefore, we can study the hyperparameters of methods, presented in Table **??** for the meeting requirements of Theorem (**??**). Studying the specific requirements on the hyperparameters α, β, γ is of great interest and is the question of further research. Block matrix formulation for the vector version of the methods from Table **??** is presented in Appendix **??**.

3.2 General d-dimensional case

Theorem 3.2. *For the quadratic optimization problem in the form of (??):*

$$\min_{x\in\mathbb{R}^d} f(x) = \min_{x\in\mathbb{R}^d} \frac{1}{2}x^T W x - b^T x + c, \text{ where } W \in \mathbb{S}^d_{++} \tag{35}$$

with a unique solution $x^\star = W^{-1}b$, given any optimization method, which converges to $x^\star$ and could be written in the following form

$$x_{k+1} = Ax_k + Bx_{k-1},$$

where $A, B \in \mathbb{R}^{d\times d}$ are diagonal matrices, or, equivalently:

$$z_{k+1} = Mz_k, \quad M = \begin{bmatrix} A & B \\ I & 0_d \end{bmatrix} \quad z_k = \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix}$$

where the eigenvalues of the iteration matrix for each dimension (see the corresponding rearrangement on Figure ??) M_i forms the conjugate pairs, i.e. $(\lambda_1^{M_i})^* = \lambda_2^{M_i} = \lambda^{M_i}\ \forall i \in 1,\dots,d$ has the following Lyapunov function:

$$V(x_k, x_{k-1}, x_{k-2}) = \|x_{k-1} - x^*\|^2 - \langle x_k - x^*, x_{k-2} - x^*\rangle \tag{36}$$

Proof. 1. It is enough to observe, that we can easily change variables with the help of eigendecomposition $W = Q\Lambda Q^\star$ similarly as it was done in (??)

$$\hat{x}_k = Q^*(x_k - x^\star) \text{ or } x_k = Q\hat{x}_k + x^\star$$

Thus, our function became quadratic form with a diagonal matrix with the minimum at $\hat{x} = 0$:

$$f(\hat{x}) = \frac{1}{2}\langle \hat{x}, \Lambda\hat{x}\rangle - \frac{1}{2}\langle b, A^{-1}b\rangle + c$$

2. Due to the possible rearrangement of matrix (Figure ?? and (??)) and, therefore, independent coordinate-wise dynamics with matrix M_i.

$$M_i = \begin{bmatrix} a_i & b_i \\ 1 & 0 \end{bmatrix},$$

where a_i, b_i are the i-th diagonal elements of matrices A and B we can write down the Lyapunov function for each dimension of the vector $\hat{x}$. It follows from the Theorem ?? that for each dimension of the vector $\hat{x} \in \mathbb{R}^d$ one can write the Lyapunov function, which will decrease monotonically if $\rho(M_i)$ and $(\lambda_1^{M_i})^* = \lambda_2^{M_i}$:

$$V^i(\hat{x}^i_k, \hat{x}^i_{k-1}, \hat{x}^i_{k-2}) = \left(\hat{x}^i_{k-1}\right)^2 - \hat{x}^i_k\hat{x}^i_{k-2}$$

3. Now we can sum all the Lyapunov functions over dimensions $\forall i \in 1,\dots,d$:

$$
\begin{aligned}
V(\hat{x}_k, \hat{x}_{k-1}, \hat{x}_{k-2}) &= \sum_{i=1}^{d} V^i(\hat{x}_k^i, \hat{x}_{k-1}^i, \hat{x}_{k-2}^i) \\
&= \sum_{i=1}^{d} \left(\left(\hat{x}_{k-1}^i\right)^2 - \hat{x}_k^i \hat{x}_{k-2}^i \right) \\
&= \|\hat{x}_{k-1}\|^2 - (\hat{x}_k)^T (\hat{x}_{k-2})
\end{aligned}
$$

4. Switching back to the original variables with $\hat{x}_k = Q^*(x_k - x^\star)$

$$
\begin{aligned}
V(x_k, x_{k-1}, x_{k-2}) &= \|Q^*(x_{k-1} - x^\star)\|^2 - (Q^*(x_k - x^\star))^T (Q(x_{k-2} - x^\star)) \\
&= (x_{k-1} - x^\star)^T Q Q^* (x_{k-1} - x^\star) - ((x_k - x^\star))^T Q Q^* ((x_{k-2} - x^\star)) \\
&= \|x_{k-1} - x^\star\|^2 - \langle x_k - x^\star, x_{k-2} - x^\star \rangle
\end{aligned}
$$

□

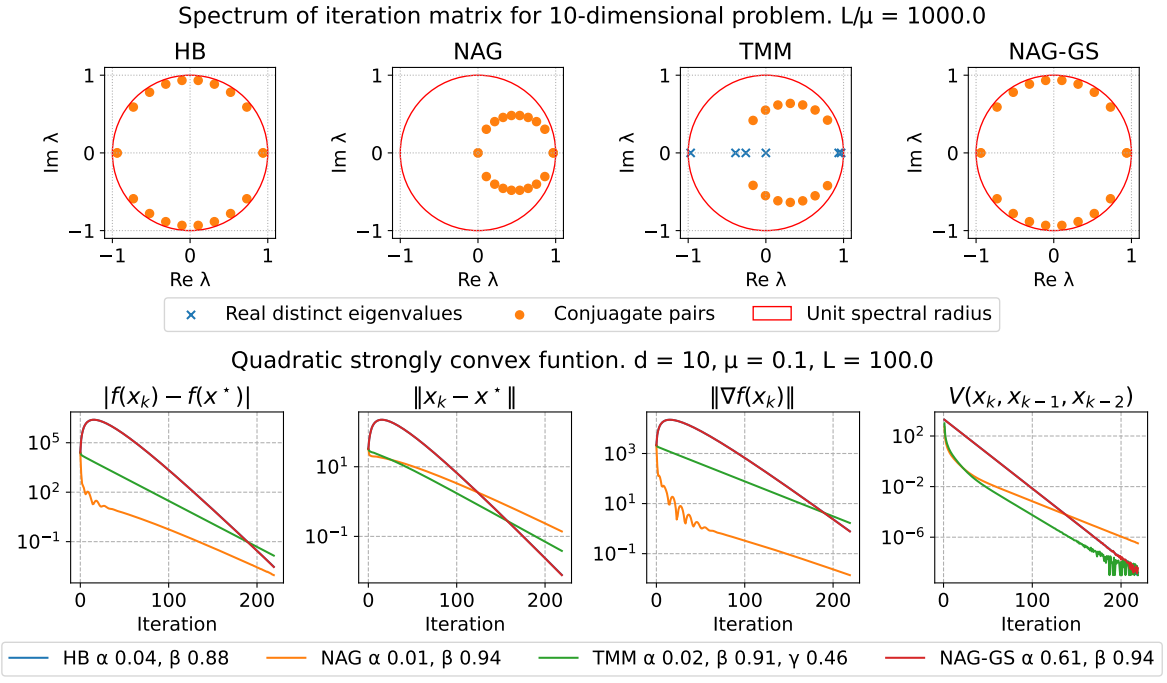


Figure 3: The correspondance between the Spectrum of iteration matrix for **HB**, **NAG**, **TMM**, **NAG-GS** methods with optimal hyperparameters applied to strongly convex 10-dimensional quadratics and convergence characteristics.

It is important to mention, that for d-dimensional case the proposed Lyapunov function is a sum of geometric progressions for each coordinate with rates $|\lambda^{M_1}|, |\lambda^{M_2}|, \ldots, |\lambda^{M_d}|$ and thus the asymptotic convergence rate is determined by the worst among them, which means, that starting from some iteration number the convergence rate will be the spectral radius

of the iteration matrix $\rho(M) = \max\limits_{i=1,\dots,d} |\lambda^{M_i}|$. It is also interesting that in the **NAG** case, the eigenvalues of the iteration matrix in the multidimensional case differ significantly in magnitude, which leads to the fact that at the beginning the convergence of the Lyapunov function is determined by the convergence along those coordinates with the smallest magnitude eigenvalues (orange line in Figure **??** from the bottom), and then it comes to the asymptotic convergence rate determined by the largest magnitude eigenvalues.

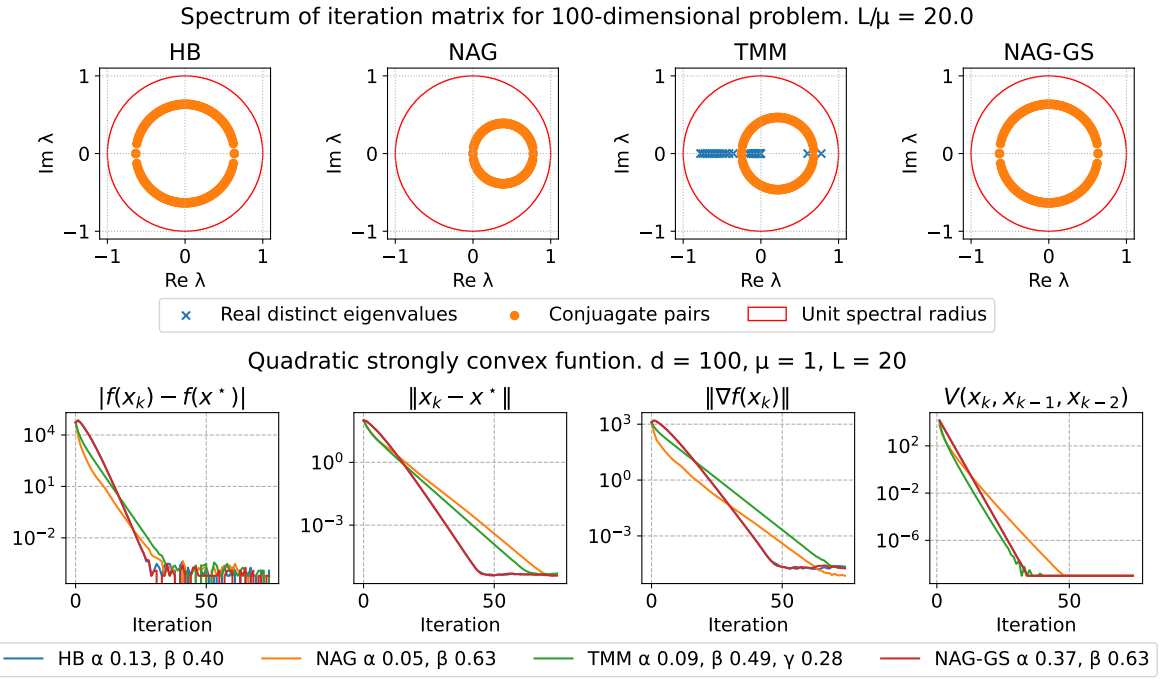


Figure 4: Dynamics of methods from Table **??** with optimal hyperparameters $\alpha^\star, \beta^\star, \gamma^\star$ applied to the strongly convex quadratic problem (**??**)

The spectrum of the considered method may say a lot about the convergence. For example on Figure **??** it is easy to verify, that the **TMM** method does not satisfy the Theorem**??** requirements, therefore we can see, that the expression (**??**) is not Lyapunov function for the method. Moreover, we can see, that eigenvalue distribution for **HB** and **NAG-GS** significantly differs from the **NAG** and **TMM**. For the first group, the absolute values of the eigenvalues are the same (they form a circle on the complex plane), while for the latter the magnitudes of the eigenvalues vary. This is the reason why the corresponding $V(x_k, x_{k-1}, x_{k-2})$ dynamics is faster at the beginning of the optimization process and slows down at the end - convergence rate at the end depends only on the spectral radius (largest magnitude).

Note that the considered class of methods with the diagonal matrices A and B contains many popular methods (see Table **??**). However, the whole idea of constructing the described Lyapunov function relies essentially on the fact that we can consider the dynamics of each component of the vector x independently (see Section **??**). Arbitrary methods with an arbitrary iteration matrix, in general, cannot fail to be suitable for such a procedure of Lyapunov function construction.

The proposed Lyapunov function works well for a variety of scenarios. However, it is not a Lyapunov function for a general (strongly) convex optimization case. The counter-examples

are provided in the corresponding sections below.

4 Numerical experiments

All the code for experiments is available on the GitHub Repository:
https://github.com/MerkulovDaniil/SimpleLyapunov

4.1 Quadratic problem

To validate the theoretical claims, we conducted experiments on quadratic problems defined in equation (**??**). We applied various first-order methods, including Heavy Ball and Nesterov Accelerated Gradient, to minimize the quadratic function.

We started by randomly generating matrices $W \in \mathbb{S}^d_{++}$ of dimensions $d = 100, 200, 500$. The generated matrices were ensured to be positive definite. Moreover, the spectrum of the matrices is uniformly spread from μ to L. Then, we generated random vector $x^\star \in \mathbb{R}^d$ and a vector $b \in \mathbb{R}^d$ was also calculated as $b = Ax^\star$ for each test case. For each setup, we performed iterations using various algorithms and monitored the value of the proposed Lyapunov function $V(x_k, x_{k-1}, x_{k-2}) = \|x_{k-1} - x^*\|^2 - \langle x_k - x^*, x_{k-2} - x^*\rangle$. In our experiments, the tolerance of V measuring is 10^{-9}, which is why we can see a plateau at this level at the end.

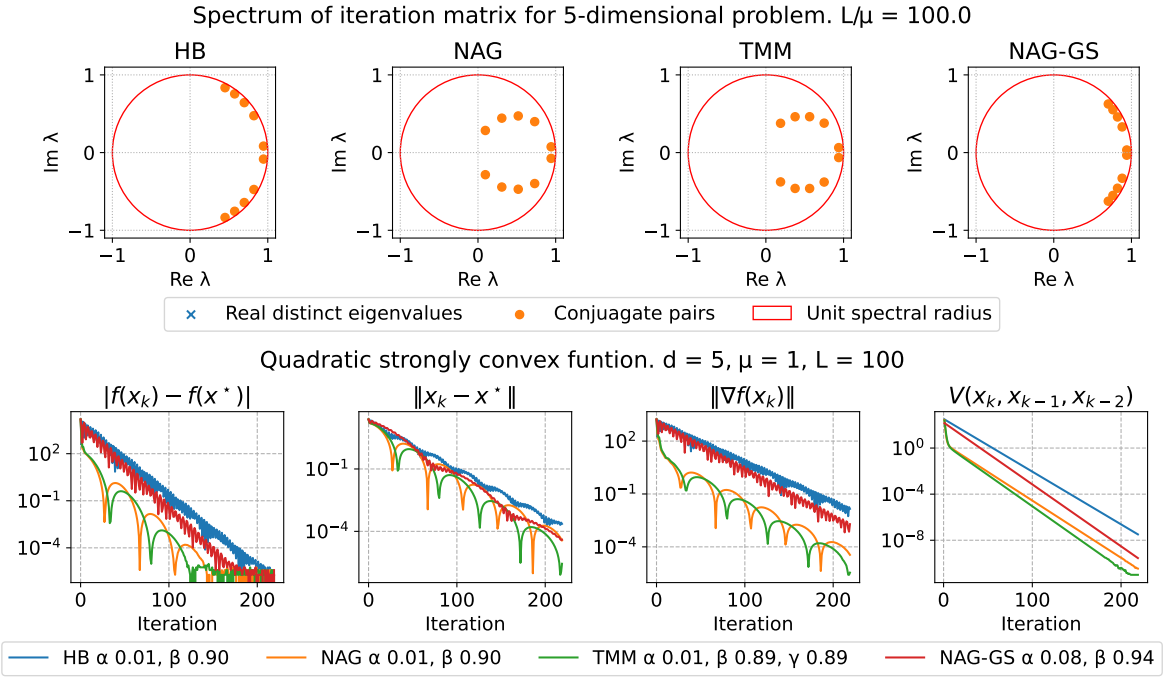


Figure 5: Dynamics of methods from Table **??** with non-optimal hyperparameters α, β, γ applied to the strongly convex quadratic problem (**??**)

4.1.1 Optimal hyperparameters for methods

The results are consistent with the theoretical predictions. The Lyapunov function monotonically decreased and approached zero as the methods converged. This indicates that

our Lyapunov function provides an accurate measure of algorithmic behavior for quadratic problems. Note, that for **TMM** method we don't have theoretical guarantees for V to be a Lyapunov function. Here are the results for the ill-conditioned quadratic problem:

4.1.2 Non-optimal, but suitable hyperparameters

It is especially interesting to look at Figure **??**, where all considered methods meet Theorem **??** requirements, despite having non-optimal hyperparameters. Nowadays, such formulation of methods, where hyperparameters are to be tuned, is widely spread in Applications - Neural Networks training.

4.1.3 Convex quadratic problem with $\mu = 0$.

It follows from the structure of the matrix M, that if the spectrum of the original matrix W has k zero eigenvalues, then we will have k real unit eigenvalues, which corresponding summands $V^i(x_k, x_{k-1}, x_{k-2})$ will not decrease during the iteration process. Practically speaking it means, that some terms of expression (**??**) will linearly decrease, which leads to an almost linear decrease of the $V(x_k, x_{k-1}, x_{k-2})$ until some level, after that we will have some oscillations. This is supported by Figure **??**.

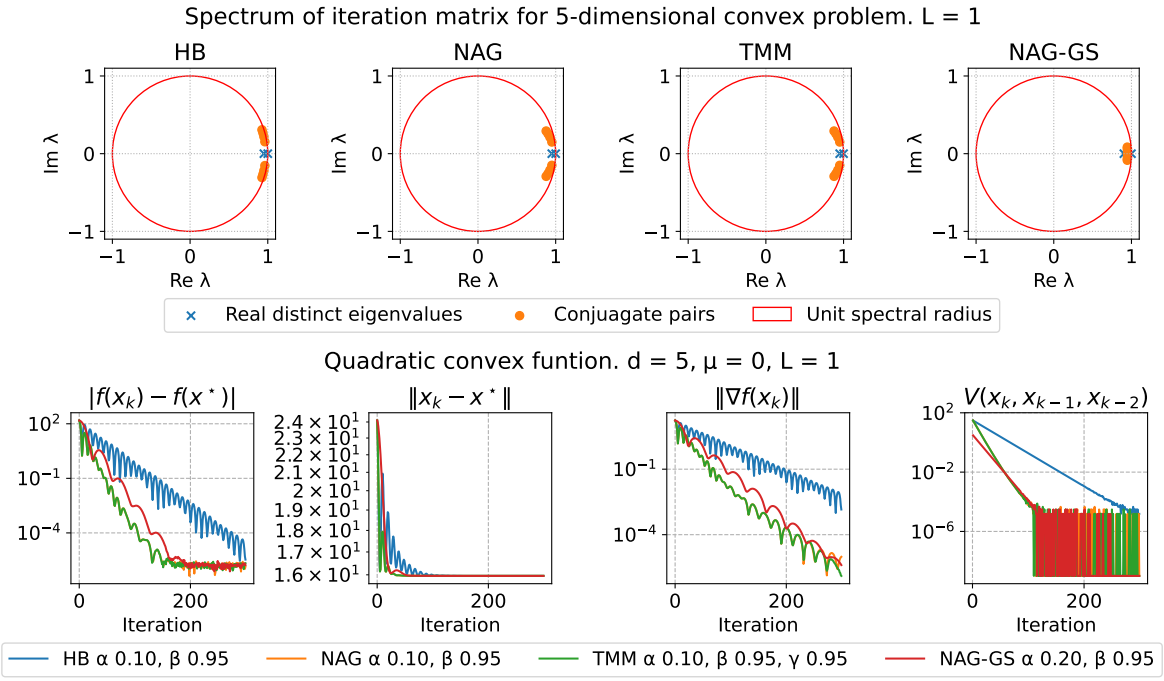


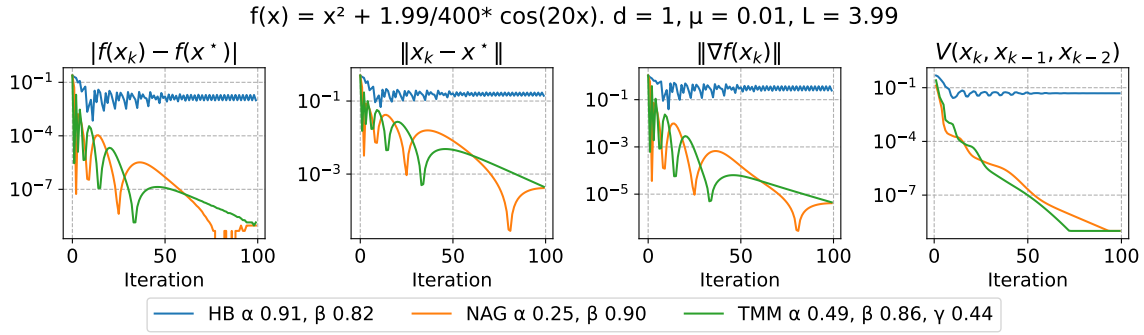
Figure 6: Dynamics of methods from Table **??** with some hyperparameters α, β, γ applied to the convex quadratic problem (**??**)

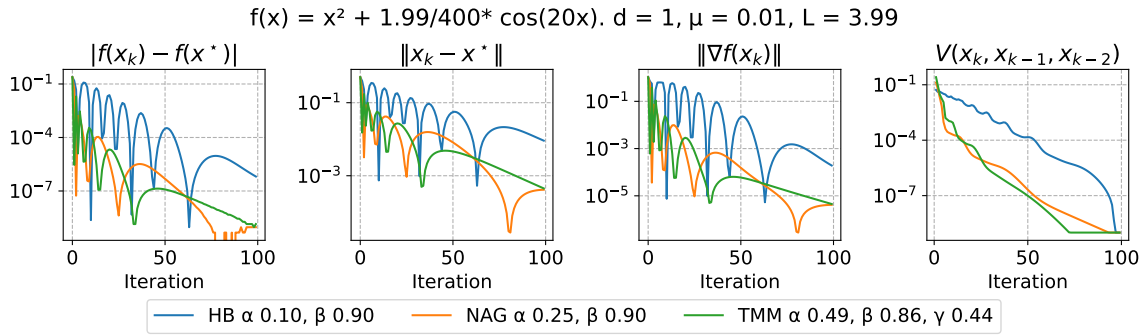
4.2 Strongly convex non-quadratic problem

We considered an example of the convex problem, where **HB** method failed to converge with optimal hyperparameters for the strongly convex function [11]

$$f(x) = x^2 + \frac{1.99}{400}\cos(20x)$$

It has $\mu = 0.01, L = 3.99$





One can conclude, that such a simple function won't serve as a general-purpose Lyapunov function.

5 Conclusion

We presented a novel (to the best of our knowledge) approach to construct a Lyapunov function for a quadratic optimization problem and first-order algorithms, based on the bounding of the last diagonal element of the iteration matrix after Schur decomposition. It is interesting to mention, that such a simple expression serves as a Lyapunov function for the wide family of two-step methods, such as Heavy Ball, Nesterov Accelerated Gradient, Triple Momentum Method, and Nesterov Accelerated Gradient with Gauss-Seidel splitting method under some conditions, which is formulated as a main result of the paper. We have conducted experiments on quadratics, that support our claims and presented a counter-example of a general strongly convex function, where the constructed function is not a Lyapunov function.

References

[1] Alexandre d'Aspremont, Damien Scieur, Adrien Taylor, et al. "Acceleration methods". In: *Foundations and Trends in Optimization* 5.1-2 (2021), pp. 1–245.

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

[3] Pontus Giselsson and Stephen Boyd. "Monotonicity and restart in fast gradient methods". In: *53rd IEEE Conference on Decision and Control*. IEEE. 2014, pp. 5058–5063.

[4] Gabriel Goh. "Why Momentum Really Works". In: *Distill* (2017). DOI: 10.23915/distill.00006. URL: http://distill.pub/2017/momentum.

[5] Eduard Gorbunov, Alexander Rogozin, Aleksandr Beznosikov, Darina Dvinskikh, and Alexander Gasnikov. "Recent theoretical advances in decentralized distributed convex optimization". In: *High-Dimensional Optimization and Probability: With a View Towards Data Science*. Springer, 2022, pp. 253–325.

[6] Sven J Hammarling. "Numerical solution of the stable, non-negative definite lyapunov equation". In: *IMA Journal of Numerical Analysis* 2.3 (1982), pp. 303–323.

[7] Valentin Leplat, Daniil Merkulov, Aleksandr Katrutsa, Daniel Bershatsky, and Ivan Oseledets. "NAG-GS: semi-implicit, accelerated and robust stochastic optimizer." In: (2022).

[8] Yurii Nesterov. "A method of solving a convex programming problem with convergence rate $\mathcal{O}\left(1/k^2\right)$". In: *Doklady Akademii Nauk*. Vol. 269. 3. Russian Academy of Sciences. 1983, pp. 543–547.

[9] Yurii Nesterov. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science & Business Media, 2003.

[10] Yurii Nesterov et al. *Lectures on convex optimization*. Vol. 137. Springer.

[11] Boris T Polyak. "Introduction to optimization. Optimization software". In: *Inc., Publications Division, New York* 1 (1987), p. 32.

[12] Boris T Polyak. "Some methods of speeding up the convergence of iteration methods". In: *Ussr computational mathematics and mathematical physics* 4.5 (1964), pp. 1–17.

[13] Adrien Taylor, Bryan Van Scoy, and Laurent Lessard. "Lyapunov functions for first-order methods: Tight automated convergence guarantees". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 4897–4906.

[14] Bryan Van Scoy, Randy A Freeman, and Kevin M Lynch. "The fastest known globally convergent first-order method for minimizing strongly convex functions". In: *IEEE Control Systems Letters* 2.1 (2017), pp. 49–54.

[15] Евгения Воронцова , Роланд Хильдебранд , Александр Гасников , Фёдор Стонякин. *Выпуклая оптимизация*. Vol. 364. MIPT, 2021. ISBN: 978-5-7417-0776-0.

A Two-step notation of gradient methods for quadratic minimization

Method	Iteration
HB [12] Optimal α, β: $\alpha^\star = \frac{4}{\left(\sqrt{L}+\sqrt{\mu}\right)^2},\ \beta^\star = \frac{\left(\sqrt{L}-\sqrt{\mu}\right)^2}{\left(\sqrt{L}+\sqrt{\mu}\right)^2}$	$x_{k+1} = x_k - \alpha\nabla f(x_k) + \beta(x_k - x_{k-1})$ $x_{k+1} = ((1+\beta)I - \alpha\Lambda)x_k - \beta x_{k-1}$ $A = (1+\beta)I - \alpha\Lambda \qquad B = -\beta I$
NAG [9] Optimal α, β: $\alpha^\star = \frac{1}{L},\ \beta^\star = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$	$\begin{cases} y_{k+1} = x_k + \beta(x_k - x_{k-1}) \\ x_{k+1} = y_{k+1} - \alpha\nabla f(y_{k+1}) \end{cases}$ $x_{k+1} = (1+\beta)(I - \alpha\Lambda)x_k - \beta(I - \alpha\Lambda)x_{k-1}$ $A = (1+\beta)(I - \alpha\Lambda) \qquad B = -\beta(I - \alpha\Lambda)$
TMM [14] Optimal $\alpha, \beta, \gamma^\star$: $\rho = 1 - \sqrt{\frac{\mu}{L}}$ $\alpha^\star = \frac{1+\rho}{L},\ \beta^\star = \frac{\rho^2}{2-\rho},\ \gamma^\star = \frac{\rho^2}{(1+\rho)(2-\rho)}$	$x_{k+1} = (1+\beta)x_k - \beta x_{k-1} - \alpha\nabla f((1+\gamma)x_k - \gamma x_{k-1})$ $x_{k+1} = ((1+\beta)I - \alpha(1+\gamma)\Lambda)x_k + (\alpha\gamma\Lambda - \beta I)x_{k-1}$ $A = (1+\beta)I - \alpha(1+\gamma)\Lambda \qquad B = \alpha\gamma\Lambda - \beta I$
NAG-GS [7] Optimal α, β: $\alpha^\star = \frac{2+2\sqrt{\frac{L}{\mu}}}{L+\mu+2\sqrt{\mu L}},\ \beta^\star = \frac{L-\mu}{L+\mu+2\sqrt{\mu L}}$	$\begin{cases} y_k = \beta y_{k-1} + (1-\beta)x_k - \alpha\nabla f(x_k) \\ x_{k+1} = \beta x_k + (1-\beta)y_k \end{cases}$ $x_{k+1} = (2\beta + (1-\beta)^2)I - \alpha(1-\beta)\Lambda x_k - \beta^2 x_{k-1}$ $A = (2\beta + (1-\beta)^2)I - \alpha(1-\beta)\Lambda \qquad B = -\beta^2 I$

Table 4: Correspondence between several accelerated methods for strongly convex functions and its reformulations concerning two-step notation.

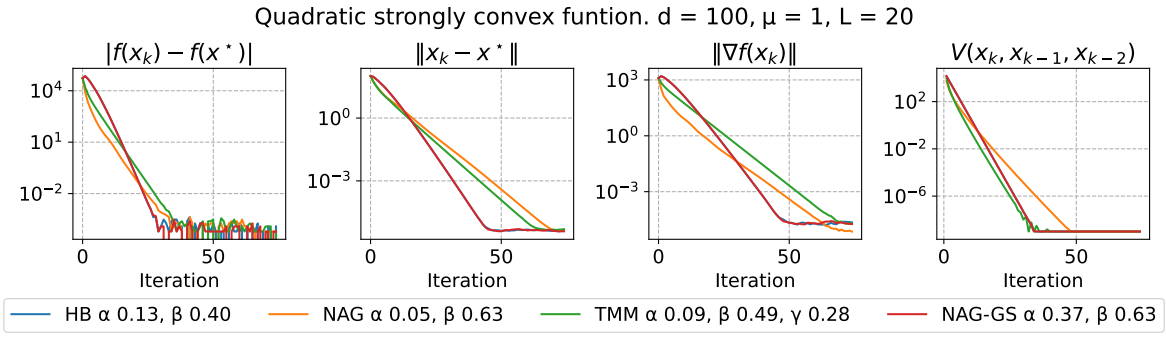
B Experiments

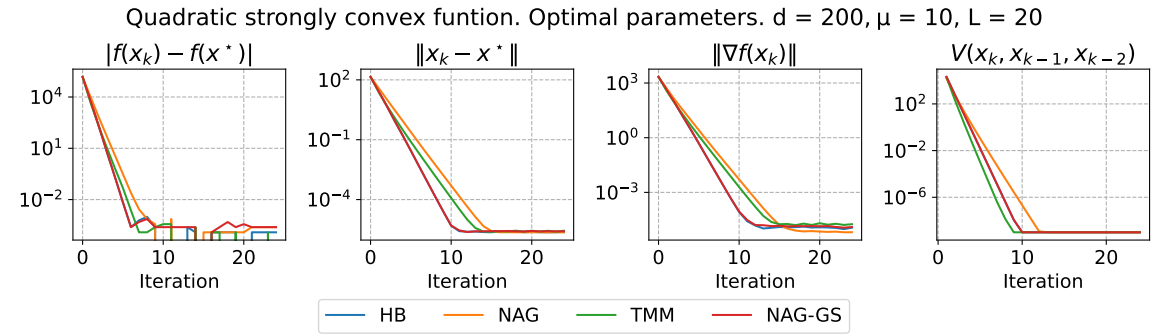
B.1 Quadratic Problem

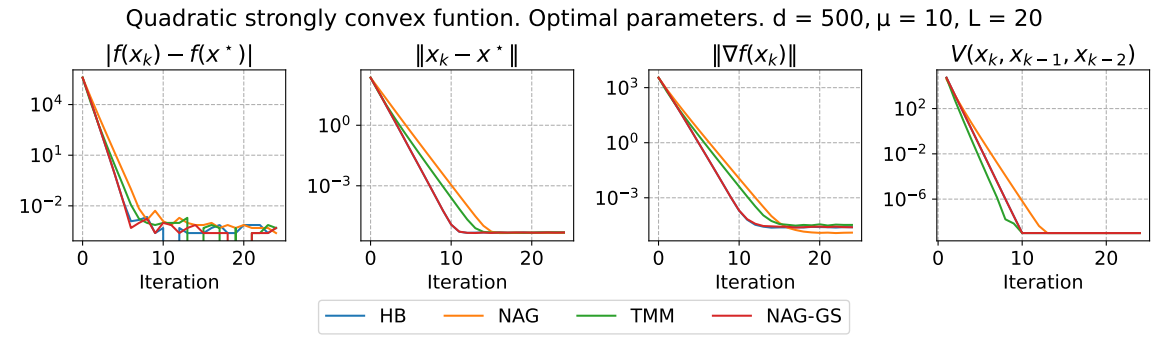
$$\min_{x\in\mathbb{R}^d} \frac{1}{2}x^T W x - b^T x + c, \text{ where } W \in \mathbb{S}^d_{++}$$

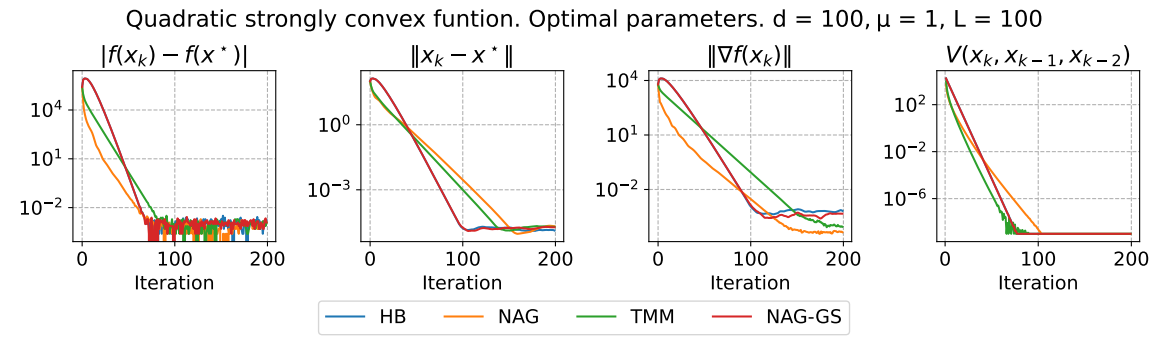
B.2 Convex problem

$$\min_{x\in\mathbb{R}^d} f(x) = e^{\|x\|_2^2}$$



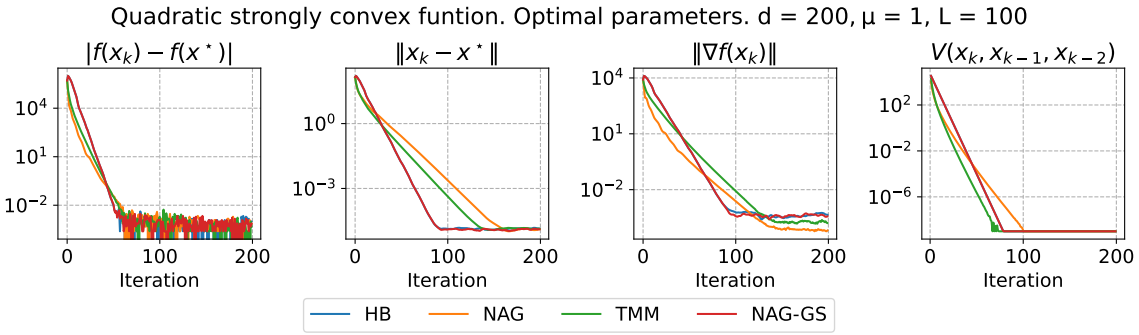


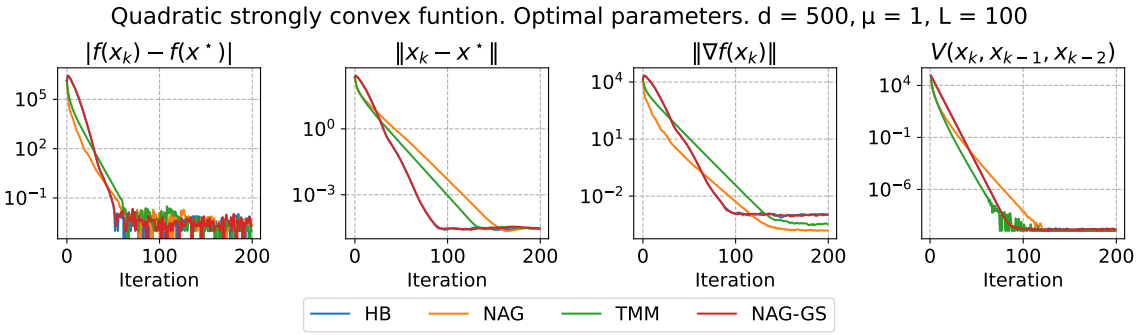


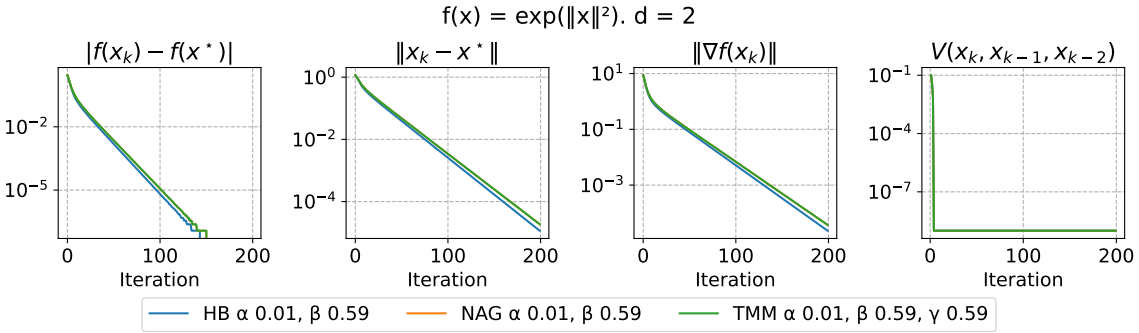


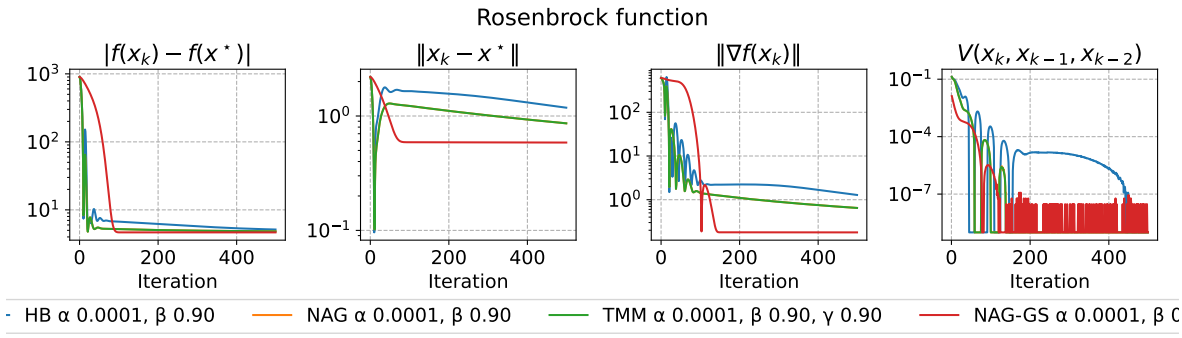
B.3 Non-convex problem

$$\min_{x,y\in\mathbb{R}^2} f(x,y) = (1-x)^2 + 100(y-x^2)^2,\ x^\star = (1,1)$$


Quadratic strongly convex funtion. Optimal parameters. d = 200, μ = 1, L = 100
$|f(x_k) - f(x^\star)|$
$\|x_k - x^\star\|$
$\|\nabla f(x_k)\|$
$V(x_k, x_{k-1}, x_{k-2})$
Iteration
HB
NAG
TMM
NAG-GS


Quadratic strongly convex funtion. Optimal parameters. d = 500, μ = 1, L = 100
$|f(x_k) - f(x^\star)|$
$\|x_k - x^\star\|$
$\|\nabla f(x_k)\|$
$V(x_k, x_{k-1}, x_{k-2})$
Iteration
HB
NAG
TMM
NAG-GS


f(x) = exp(‖x‖²). d = 2
$|f(x_k) - f(x^\star)|$
$\|x_k - x^\star\|$
$\|\nabla f(x_k)\|$
$V(x_k, x_{k-1}, x_{k-2})$
Iteration
HB α 0.01, β 0.59
NAG α 0.01, β 0.59
TMM α 0.01, β 0.59, γ 0.59


Rosenbrock function
$|f(x_k) - f(x^\star)|$
$\|x_k - x^\star\|$
$\|\nabla f(x_k)\|$
$V(x_k, x_{k-1}, x_{k-2})$
Iteration
HB α 0.0001, β 0.90
NAG α 0.0001, β 0.90
TMM α 0.0001, β 0.90, γ 0.90
NAG-GS α 0.0001, β 0