axioms

MDPI

Article

Machine Learning in Quasi-Newton Methods

Vladimir Krutikov [1,2], Elena Tovbis [3], Predrag Stanimirović [1,4], Lev Kazakovtsev [1,3,*] and Darjan Karabašević [5,6,*]

[1] Laboratory "Hybrid Methods of Modeling and Optimization in Complex Systems", Siberian Federal University, 79 Svobodny Prospekt, 660041 Krasnoyarsk, Russia; krutikovvn@rambler.ru (V.K.); pecko@pmf.ni.ac.rs (P.S.)
[2] Department of Applied Mathematics, Kemerovo State University, 6 Krasnaya Street, 650043 Kemerovo, Russia
[3] Institute of Informatics and Telecommunications, Reshetnev Siberian State University of Science and Technology, 31, Krasnoyarskii Rabochii Prospekt, 660037 Krasnoyarsk, Russia; sibstu2006@rambler.ru
[4] Faculty of Sciences and Mathematics, University of Niš, 18000 Niš, Serbia
[5] College of Global Business, Korea University, Sejong 30019, Republic of Korea
[6] Faculty of Applied Management, Economics and Finance, University Business Academy in Novi Sad, Jevrejska 24, 11000 Belgrade, Serbia
* Correspondence: levk@bk.ru (L.K.); darjan.karabasevic@mef.edu.rs (D.K.)

Abstract: In this article, we consider the correction of metric matrices in quasi-Newton methods (QNM) from the perspective of machine learning theory. Based on training information for estimating the matrix of the second derivatives of a function, we formulate a quality functional and minimize it by using gradient machine learning algorithms. We demonstrate that this approach leads us to the well-known ways of updating metric matrices used in QNM. The learning algorithm for finding metric matrices performs minimization along a system of directions, the orthogonality of which determines the convergence rate of the learning process. The degree of learning vectors' orthogonality can be increased both by choosing a QNM and by using additional orthogonalization methods. It has been shown theoretically that the orthogonality degree of learning vectors in the Broyden–Fletcher–Goldfarb–Shanno (BFGS) method is higher than in the Davidon–Fletcher–Powell (DFP) method, which determines the advantage of the BFGS method. In our paper, we discuss some orthogonalization techniques. One of them is to include iterations with orthogonalization or an exact one-dimensional descent. As a result, it is theoretically possible to detect the cumulative effect of reducing the optimization space on quadratic functions. Another way to increase the orthogonality degree of learning vectors at the initial stages of the QNM is a special choice of initial metric matrices. Our computational experiments on problems with a high degree of conditionality have confirmed the stated theoretical assumptions.

Keywords: minimization algorithm; quasi-Newton method; convergence rate; machine learning

MSC: 90C53

Citation: Krutikov, V.; Tovbis, E.; Stanimirović, P.; Kazakovtsev, L.; Karabasevic, D. Machine Learning in Quasi-Newton Methods. *Axioms* **2024**, *13*, 240. https://doi.org/10.3390/axioms13040240

Academic Editor: Gustavo Olague

Received: 14 February 2024
Revised: 22 March 2024
Accepted: 2 April 2024
Published: 5 April 2024

Copyright: © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The problem of unconstrained minimization of smooth functions in a finite-dimensional Euclidean space has received a lot of attention in the literature [1,2]. In unconstrained optimization, in contrast to constrained optimization [3], the process of optimizing the objective function is carried out in the absence of restrictions on variables. Unconstrained problems arise also as reformulations of constrained optimization problems, in which the constraints are replaced by penalization terms in the objective function that have the effect of discouraging constraint violations [2].

Well-known methods [1,2] that enable us to solve such a problem include the gradient method, which is based on the idea of function local linear approximation, or Newton's method, which uses its quadratic approximation. The Levenberg–Marquardt

method is a modification of Newton's method, where the direction of descent differs from that specified by Newton's method. The conjugate gradient method is a two-step method in which the parameters are found from the solution of a two-dimensional optimization problem.

Quasi-Newton minimization methods are effective tools of solving smooth minimization problems when the function level curves have a high degree of elongation [4–7]. QNMs are commonly applied in a wide range of areas, such as biology [8], image processing [9], technics [10–15], and deep learning [16–18].

The QNM is based on the idea of using a matrix of second derivatives reconstructed from the gradients of a function. The first QNM was proposed in [19] and improved in [20]. The generally accepted notation for the matrix updating formula in this method is DFP. Nowadays, there are a significant number of equations for updating matrices in the QNM [4–7,21–28], and it is generally accepted [4,5] that among a variety of QNMs, the best methods use the BFGS matrix updating equation [29–31]. However, it has been experimentally established, but not theoretically explained, why the BFGS generates the best results among the QNMs [5].

A sampled version of the BFGS method named limited-memory BFGS (L-BFGS) [32] was presented to handle high-dimensional problems. The algorithm stores only a few vectors that represent the approximation of the Hessian instead of the entire matrix. A version with bound constraints was proposed in [33].

The penalty method [2] was developed for solving constrained optimization problems. The unconstrained problems are formed by adding a term, called a penalty function, to the objective function. The penalty is zero for feasible points and non-zero for infeasible points.

The development of QNMs occurred spontaneously through the search for matrix updating equations that satisfy certain properties of data approximation obtained in the problem solving process. In this paper, we consider a method for deriving matrix updating equations in QNMs by forming a quality functional based on learning relations for matrices, followed by obtaining matrix updating equations in the form of a step of the gradient method for minimizing the quality functional. This approach has shown high efficiency in organizing subgradient minimization methods [34,35].

In machine learning theory, the system in which the average risk (mathematical expectation of the total loss function) is minimal is considered optimal [36,37]. The goal of learning represents the state that has to be reached by the learning system in the process of learning. The selection of such a desired state is actually achieved by a proper choice of a certain functional that has an extremum which corresponds to the desired state [38]. Thus, in the matrix learning process, it is necessary to formulate a quality functional.

In QNMs, for each of the matrix rows, there is a product of the vector which exists as a learning relation. Consequently, we have a linear model with the coefficients of the matrix row as its parameters. Thus, we may formulate a quadratic learning quality functional for a linear model and obtain a gradient machine learning (ML) algorithm. This paper shows how one can obtain known methods for updating matrices in QNMs based on a gradient learning algorithm. Based on the general properties of convergence of gradient learning algorithms, it seems relevant to study the origins of the effectiveness of metric updating equations in QNMs.

In a gradient learning algorithm, the sequence of steps is represented as a method of minimization along a system of directions. The degree of orthogonality of these directions determines the convergence rate of the algorithm. The use of gradient learning algorithms for deriving matrix updating equations in QNMs enables us to analyze the quality of matrix updating algorithms based on the convergence rate properties of the learning algorithms. This paper shows that the higher degree of orthogonality of learning vectors in the BFGS method determines its advantage compared to the DFP method.

Studies on quadratic functions identify conditions under which the space dimension is reduced during the QNM iterations. The dimension of the minimization space is

reduced when the QNM includes iterations with an exact one-dimensional descent or an iteration with additional orthogonalization. It is possible to increase the orthogonality of the learning vectors and thereby increase the convergence rate of the method through special normalization of the initial matrix.

The computational experiment was carried out on functions with a high degree of conditionality. Various ways of increasing the orthogonality of learning vectors were assessed. The theoretically predicted effects of increasing the efficiency of QNMs confirmed their effectiveness in practice. It turned out that with an approximate one-dimensional descent, additional orthogonalization in iterations of the algorithm significantly increased the efficiency of the method. In addition, the efficiency of the method also increased significantly with the correct normalization of the initial matrix.

The rest of this paper is organized as follows. In Section 2, we provide basic information about matrix learning algorithms in QNMs. Section 3 contains an analysis of matrix updating formulas in QNMs. A symmetric positive definite metric is considered in Section 4. Section 5 gives a qualitative analysis of the BFGS and DFP matrix updating equations. Methods for reducing the minimization space of QNMs on quadratic functions are presented in Section 6. Methods for increasing the orthogonality of learning vectors in QNMs are considered in Section 7. In Section 8, we present a numerical study, and the last section summarizes the work.

2. Matrix Learning Algorithms in Quasi-Newton Methods

Consider the minimization problem

$$f(x) \to min,\ x \in R^n.$$

The QNM for this problem is iterated as follows:

$$x^{k+1} = x^k + \beta_k s^k, \quad s^k = -H^k \nabla f(x^k), \tag{1}$$

$$\beta_k = \arg\min_{\beta \geq 0} f(x^k + \beta s^k), \tag{2}$$

$$\Delta x^k = x^{k+1} - x^k,\ y^k = \nabla f(x^{k+1}) - \nabla f(x^k), \tag{3}$$

$$H^{k+1} = H(H^k, \Delta x^k, y^k). \tag{4}$$

Here, $\nabla f(x)$ is the gradient of a function, s^k is the search direction, and β_k is chosen to satisfy the Wolfe conditions [2]. Further, $H^k \in R^{n \times n}$ is a symmetric matrix which is used as an approximation of the Hessian inverse. The operator

$$H(H, \Delta x, y) \in R^{n \times n},\ H \in R^{n \times n},\ \Delta x, y \in R^n \tag{5}$$

specifies a certain equation for updating the initial matrix H. At the input of the algorithm, the starting point x_0 and the symmetric strictly positive definite matrix H^0 must be specified. Such a matrix will be denoted as $H^0 > 0$.

Let us consider the relations for obtaining updating equations for H^k matrices on quadratic functions:

$$f(x) = \frac{1}{2}\langle x - x^*, A(x - x^*)\rangle + d, \quad A > 0, \tag{6}$$

where x^* is the minimum point. Here and below, the expression $\langle\cdot,\cdot\rangle$ means a scalar product of vectors. Without a loss of generality, we assume $d = 0$. The gradient of a quadratic function $f(x)$ is $\nabla f(x) = A(x - x^*)$. For $\Delta x \in R^n$, the gradient difference $y = \nabla f(x + \Delta x) - \nabla f(x)$ satisfies the relation:

$$A\Delta x = y \ \text{ or } A^{-1}y = \Delta x. \tag{7}$$

The equalities in (7) are used to obtain various equations for updating matrices H^k, which are approximations for A^{-1}, or matrices $B^k = (H^k)^{-1}$, which are approximations for A. An arbitrary equation for updating matrices H or B, the result of which is a matrix satisfying (7), will be denoted by $H(H, \Delta x, y)$ or $B(B, \Delta x, y)$, respectively.

Denoting as A_i and A_i^{-1} rows of the corresponding matrices A and A^{-1} with i-th index, then, according to (7), we obtain equations for the learning relations necessary to formulate algorithms for matrix rows' learning:

$$A_i \Delta x = y_i, \quad A_i^{-1} y = \Delta x_i, \quad i = 1,2,\ldots,n, \tag{8}$$

where y_i and Δx_i are the components of the vectors in (7). The relations in (8) make it possible to use machine learning algorithms of a linear model in the parameters to estimate the rows of the corresponding matrices.

Let us formulate the problem of estimating the parameters of a linear model from observational data.

ML problem: find unknown parameters $c^* \in R^n$ of the linear model

$$y = \langle z, c \rangle, \qquad z, c \in R^n, \; y \in R^1 \tag{9}$$

from observational data

$$y_k \in R^1, \; z^k \in R^n, k = 0,1,2,\ldots, \tag{10}$$

where $y_k = <c^*, z^k>$. We will use an indicator of training quality,

$$Q(z,c) = \frac{1}{2}(\langle z,c \rangle - y)^2, \tag{11}$$

which is an estimate of the quality functional required to find c^*.

Function (11) is a loss function. Due to the large dimension of the problem of estimating the elements of metric matrices, the use of the classical least squares method becomes difficult. We use the adaptive least squares method (recurrent least squares formulas).

The gradient learning algorithm based on (11) has the following form:

$$c^{k+1} = c^k - h_k \nabla Q(z^k, c^k) = c^k - h_k(\langle z^k, c^k \rangle - y_k) z^k. \tag{12}$$

Due to the orthogonality of the training vectors, the stochastic gradient method in the form "receiving of an observation-training-forgetting the observation information" in quasi-Newton methods enables us to obtain good approximations of the inverse matrices of second derivatives while maintaining their symmetry and positive definiteness.

In this paper, the value of such consideration is that we are able to identify the advantages of the BFGS method and obtain a method with orthogonalization of learning vectors and prove these provisions through testing.

The Kaczmarz algorithm [39] is a special case of (12) with the form

$$c^{k+1} = c^k - \frac{(\langle z^k, c^k \rangle - y_k)}{\langle z^k, c^k \rangle} z^k. \tag{13}$$

Let us list some of the properties of process (13), which we use to justify the properties of matrix updating in QNMs.

Property 1. *Process (13) ensures the equality*

$$y_k = \langle z^k, c^{k+1} \rangle, \tag{14}$$

and the solution is achieved under the condition of minimum changes in the parameters' values $||c^{k+1} - c^k||$.

Property 2. *If* $y_k = <c^*, z^k>$ *then the iteration of process (13) is equivalent to the step of minimizing the quadratic function*

$$\phi(c) = \langle c - c^*, c - c^* \rangle / 2 \tag{15}$$

from the point c^k along the direction z^k.

Proof. Property 2 is justified by the direct implementation of the function in (15) which is the minimizing step along the direction z^k, which is presented in Figure 1. Property 1 follows from the fact that movement to the point c^{k+1} is carried out along the normal to the hyperplane $\langle z^k, c \rangle = y_k$, that is, along the shortest path (Figure 1). Movement to other points on the hyperplane, for example to point A, satisfy only the condition in (14). □

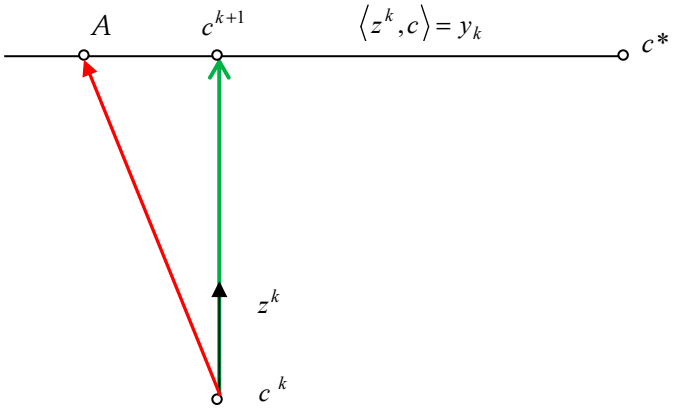

Figure 1. Step of process (13) on hyperplane $\langle z^k, c \rangle = y_k$ along the direction z^k.

Let us denote the residual as $r^k = c^k - c^*$. By subtracting c^* from both sides of (13) and making transformations, we obtain the following learning algorithm in the form of residuals:

$$r^{k+1} = W(z^k) r^k, \quad W(z) = I - \frac{z z^T}{z^T z}, \tag{16}$$

where I is the identity matrix. The sequence of minimization steps can be represented in the form of the residual transformation, where m is the number of iterations:

$$r^{k+1} = W_{k-m}^k(z)\, r^{k-m}, \quad W_{k-m}^k(z) = W(z^k)\, W(z^{k-1}) \cdots W(z^{k-m}). \tag{17}$$

The convergence rate of process (13) is significantly affected by the degree of orthogonality of the learning vectors z. The following property reflects the well-known fact of the minimization algorithm termination along orthogonal directions of the quadratic form of (15) with equal Hessian eigenvalues.

Property 3. *Let vectors z^k, $k = l, l+1, \ldots, l+n-1$ for a sequence of n iterations (13) be mutually orthogonal. Then, the solution c^* minimizing the function (15) is obtained in no more than n steps of the process (13) for an arbitrary initial c_l, wherein*

$$r^{l+n} = W_l^{l+n-1}(z) r^l = 0, \; W_l^{l+n-1}(z) = 0. \tag{18}$$

The following results are useful to estimate the convergence rate of the process in (13) as a method for minimizing the function in (15) without orthogonality of the descent vectors.

Consider a cycle of iterations for minimizing a function $\theta(x)$, $x \in R^n$, along the column vectors z^k, $\|z^k\| = 1$, $k = 1, \ldots, n$, of matrix $Z \in R^{n \times n}$:

$$x_{k+1} = x_k + \beta_k z_k, \; \beta_k = \arg\min_{\beta \geq 0} \theta(x_k + \beta z_k), \; k = 1, \ldots, n. \tag{19}$$

Here and below, we will use the Euclidean vector norm $\|x\| = \langle x, x\rangle^{1/2}$. Let us present the result of the iterations in (19) in the form of the operator $x_{n+1} = XP(x_1, Z)$. Consider the process

$$u^{q+1} = XP(u^q, Z^q), q = 0,1,\ldots, \tag{20}$$

where matrices Z^q and the initial approximation u^0 are given. To estimate the convergence rate of the QNM and the convergence rate of the metric matrix approximation, we need the following assumption about the properties of the function.

Assumption 1. *Let the function be strongly convex, with a constant $\rho > 0$, and differentiable, and its gradient satisfy the Lipschitz condition with a constant $L > 0$.*

We assume that the function $f(x)$, $x \in R^n$, is differentiable and strongly convex in R^n, i.e., there exists $\rho > 0$ such that for all $x,y \in R^n$ and $\alpha \in [0, 1]$, the inequality holds,

$$f(\alpha x + (1-\alpha)y) \le \alpha f(x) + (1-\alpha) f(y) - \alpha(1-\alpha)\rho\|x-y\|^2/2,$$

and its gradient $\nabla f(x)$ satisfies the Lipschitz condition:

$$\|\nabla f(x) - \nabla f(y)\| \le L\|x-y\| \qquad \forall x, y \in R^n,\ L > 0. \tag{21}$$

Let us denote the minimum point of the function $\theta(x)$ by x^*. The following theorem [40] establishes the convergence rate of the iteration cycle (20).

Theorem 1. *Let the function $\theta(x)$, $x \in R^n$, satisfy Assumption 1; let matrices Z^q of the process in (20) be such that minimum eigenvalues μ^q of matrices $(Z^q)^T Z^q$ satisfy the constraint $\mu^q \ge \mu_0 > 0$. Then, the following inequality estimates the convergence rate of the process in (20):*

$$\theta(u^m) - \theta(x^*) \le [\theta(u^0) - \theta(x^*)]\exp\left(-\frac{m\rho^2\mu_0^2}{2L^2n^3}\right). \tag{22}$$

Estimate (22) enables us to formulate the following property of the process in (13).

Property 4. *Let vectors z_k, $k = 1,\ldots,n-1$, be given in (13), the columns of the matrices Z be composed of vectors $z_k/\|z_k\|$, and the minimum eigenvalue μ of the matrix $Z^T Z$ satisfy the constraint $\mu \ge \mu_0 > 0$. Then, the following inequality estimates the convergence rate:*

$$||c^n - c^*||^2 \le ||c^0 - c^*||^2 \, exp\left(-\frac{\mu_0^2}{2n^3}\right). \tag{23}$$

Proof. Let us apply the results of Theorem 1 to the process in (13). The strong convexity and Lipschitz constants for the gradient of the quadratic function in (15) are the same: $\rho = L = 1$. Using Property 2 and the estimate in (22) for $m = 1$, we obtain (23). □

The property of operators W_l^{l+n-1}, when the conditions of Property 4 are met, is determined by the estimate in (23), which can be represented in the following form:

$$r^n = \| W_0^{n-1}(z) r^0 \|^2 \le \| r^0 \|^2 \exp\left(-\frac{\mu_0^2}{2n^3}\right) \tag{24}$$

Thus, the Kaczmarz algorithm provides a solution to the equality in (14) for the last observation, while it implements a local learning strategy, i.e., a strategy for iteratively improving the approximation quality from a functional (15) point of view. If the learning vectors are orthogonal, the solution is found in no more than n iterations. When n learn-

ing vectors are linearly independent, the convergence rate (23) is determined by the degree of the learning vectors' orthogonality. The degree of the vectors' orthogonality will indicate the boundedness of the minimum eigenvalue $\mu \geq \mu_0 > 0$ of the matrix Z^TZ defined in Property 4.

Using the learning relations in (8), we obtain machine learning algorithms for estimating the rows of the corresponding matrices in the form of the process in (13). Consequently, the question of analyzing the quality of algorithms for updating matrices in QNMs will consist of analyzing learning relations like (8) and the degree of orthogonality of the vectors involved in training.

3. Gradient Learning Algorithms for Deriving and Analyzing Matrix Updating Equations in Quasi-Newton Methods

Well-known equations for matrix updating in QNMs were found as equations that eliminate mismatch on a new portion of training information. In machine learning theory, a quality measure is formulated. A gradient minimization algorithm is used to minimize this measure. Our goal is to give an account of QNMs from the standpoint of machine learning theory, i.e., to formulate quality measures of training and construct their minimization algorithms. This approach enables us to obtain a unified method for deriving matrix updating equations and extend the known facts and algorithms of learning theory to solve analysis of and achieve improvement in QNMs.

Let us obtain formulas for updating matrices in QNMs using the quadratic model of the minimized function in (6) and learning relations in (7). For one of the learning relations in (7), we present a complete study of Properties 1–4.

Let the current approximation H of the matrix $H^* = A^{-1}$ be known. It is required to construct a new approximation using the learning relations in (7) for the rows of the matrix in (8):

$$H^* y = \Delta x \quad \text{or} \quad H_i^* y = \Delta x_i \,, \quad i = 1,2,...,n \,. \tag{25}$$

To evaluate each row of the matrix H^* based on (25), we apply Algorithm (13). As a result, we obtain the following matrix updating equation:

$$H^+ = H_{B2}(H, \Delta x, y) = H + \frac{(\Delta x - Hy)y^T}{y^T y}, \tag{26}$$

which is known as the 2nd Broyden method for estimating matrices when solving systems of non-linear equations [5,6].

Equation (26) determines the step of minimizing a type of functional of (15) for each of the rows H_i of matrix H along the direction y:

$$\phi(H_i) = ||H_i - H_i^*||^2/2, \quad i = 1,2,...,n \,. \tag{27}$$

The matrix residual is $R = H - H^*$. Because of the iteration of (26), the residual is transformed according to the rule

$$R^+ = RW(y). \tag{28}$$

Let us denote the scalar product for matrices $A,B \in R^{n \times n}$ as

$$\langle A, B \rangle = \sum_{i=1}^{n} A_i^T B_i = \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} B_{ij} \,.$$

We use the Frobenius norm of matrices:

$$\| H \| = \left(\sum_{i=1}^{n} \| H_i \|^2 \right)^{1/2} \,.$$

Let us define the function,

$$\Phi(H) = \sum_{i=1}^{n} ||H_i - H_i^*||^2 /2 = ||H - H^*||^2/2, \quad (29)$$

and reformulate Properties 1–4 for the matrix updating process in (26).

Theorem 2. *Iteration (26) is equivalent to the minimization step $\Phi(H)$ from a point H along the direction ΔH:*

$$\Delta H = (\Delta x - Hy)y^T/y^T y, \quad (30)$$

where

$$H^+ y = \Delta x \quad (31)$$

$$||H^+ - H|| \le ||H^{\Delta x} - H|| \quad (32)$$

for arbitrary matrices $H^{\Delta x} \in R^{n\times n}$ satisfying the condition in (31).

Proof of Theorem 2. Let us show that the condition for the minimum of the function in (27) along the direction ΔH (30) is satisfied at the point H^+:

$$\begin{aligned}\langle \Delta H, \nabla\Phi(H^+)\rangle &= \sum_{j=1}^{n}\sum_{i=1}^{n}(\Delta x - Hy)_i y_j (H_{ij}^+ - H_{ij}^*) \\ &= \sum_{i=1}^{n}(\Delta x - Hy)_i (H_i^+ - H_i^*)y = \sum_{i=1}^{n}(\Delta x - Hy)_i(\Delta x_i - \Delta x_i) = 0.\end{aligned} \quad (33)$$

□

Next, we prove (32) by showing that ΔH is the normal of the hyperplane of matrices satisfying the condition in (31). To do this, we prove orthogonality of the vector in (30) to an arbitrary vector of the hyperplane, formed as the difference of matrices belonging to the hyperplane $V = H^1 - H^2$:

$$\begin{aligned}\langle \Delta H, H^1 - H^2\rangle &= \sum_{j=1}^{n}\sum_{i=1}^{n}(\Delta x - Hy)_i y_j (H_{ij}^1 - H_{ij}^2) \\ &= \sum_{i=1}^{n}(\Delta x - Hy)_i (H_i^1 - H_i^2)y = \sum_{i=1}^{n}(\Delta x - Hy)_i(\Delta x_i - \Delta x_i) = 0.\end{aligned}$$

Let us prove an analogue of Property 3 for (26).

Theorem 3. *Let the vectors y_k, $k = l, l + 1, \ldots, l + n - 1$, for the sequence of n iterations in (26) be mutually orthogonal, then the solution H^* to the minimization problem in (29) will be obtained in no more than n steps of the process in (26),*

$$H^{k+1} = H_{B2}(H^k, \Delta x_k, y_k),\ k = l, l+1, \ldots, l+n-1, \quad (34)$$

for an arbitrary matrix H^l,

$$R^{l+n} = R^{l+n}[W_l^{l+n-1}(y)]^T = 0. \quad (35)$$

Proof of Theorem 3. From (28), the orthogonality of vectors y_k and (18) follows (35). □

Theorem 4. *Let vectors y_k, $k = 0, 1, \ldots, n - 1$, in (13) be given, vectors $y_k/\|y_k\|$ be columns of matrix P, and the minimum eigenvalue μ of a matrix P^TP satisfy the constraint $\mu \ge \mu_0 > 0$. Then, to estimate the convergence rate of the process in (34), the following inequality holds:*

$$||H^n - H^*||^2 \le ||H^0 - H^*||^2 \ exp\left(-\frac{\mu_0^2}{2n^3}\right). \quad (36)$$

Proof of Theorem 4. According to Property 4 and conditions of the theorem, the rows of matrices will have the following estimates (23):

$$||H_i^n - H_i^*||^2 \le ||H_i^0 - H_i^*||^2 \ exp\left(-\frac{\mu_0^2}{2n^3}\right), i = 0,1,\ldots,n-1.$$

A similar inequality will be true for the sums of the left and right sides. Considering the connection between the norms $||H^n - H^*||^2 = \sum_{i=1}^n ||H_i^n - H_i^*||^2$, we obtain the estimate in (36). □

In the case when the matrix H is symmetric, two products of the matrix H^* and the vector y are known:

$$H^* y = \Delta x, \ y^T H^* = \Delta x^T. \tag{37}$$

Applying the process in (28) twice for (37), we obtain a new process for updating the matrix residual:

$$R^+ = W(y)RW(y). \tag{38}$$

Expanding (38), we obtain the updating formula $H^+ = H_G(H, \Delta x, y)$ of J. Grinstadt [5,6], where

$$H_G(H, \Delta x, y) = H + \frac{\langle Hy - \Delta x, y\rangle yy^T}{\langle y, y\rangle^2} - \frac{y(Hy - \Delta x)^T + (Hy - \Delta x)y^T}{\langle y, y\rangle}. \tag{39}$$

Let us reformulate Properties 1–4 of the matrix updating process (26) for (39).

Theorem 5. *The iteration of* (*39*) *is equivalent to the minimization step* $\Phi(H)$ *from a point* H *along the* ΔH *direction:*

$$\Delta H = \frac{\langle Hy - \Delta x, y\rangle yy^T}{(y^T y)^2} - \frac{y(Hy - \Delta x)^T}{y^T y} - \frac{(Hy - \Delta x)y^T}{y^T y}. \tag{40}$$

At the same time,

$$H^+ y = \Delta x, \ \ y^T H^+ = \Delta x^T, \tag{41}$$

$$||H^+ - H|| \le ||H^{\Delta x} - H|| \tag{42}$$

for arbitrary matrices $H^{\Delta x} \in R^{n \times n}$ *satisfying the condition in (41).*

Proof of Theorem 5. Let us show that at the point H^+, the condition for the minimum of the function in (27) along the direction ΔH is satisfied:

$$\langle \Delta H, \nabla\Phi(H^+)\rangle = \langle \Delta H, H^+ - H^*\rangle = 0. \tag{43}$$

In (43), let us consider the scalar product for each term of (40) separately. The third term of Expression (40) coincides with (30). The equality to zero of the scalar product for it was obtained in (33). For the first term, the calculations are similar to (33):

$$\langle \Delta H^1, \nabla\Phi(H^+)\rangle = \frac{\langle Hy - \Delta x, y\rangle}{\langle y, y\rangle^2} \sum_{j=1}^n \sum_{i=1}^n y_i y_j (H_{ij}^+ - H_{ij}^*)$$
$$= \frac{\langle Hy - \Delta x, y\rangle}{\langle y, y\rangle^2} \sum_{i=1}^n y_i (H_i^+ - H_i^*) y = \frac{\langle Hy - \Delta x, y\rangle}{\langle y, y\rangle^2} \sum_{i=1}^n y_i (\Delta x_i - \Delta x_i) = 0.$$

Let us carry out calculations for the second term using the symmetry of matrices:

$$\langle y, y\rangle \langle \Delta H^2, \nabla\Phi(H^+)\rangle = \sum_{j=1}^n \sum_{i=1}^n y_i (Hy - \Delta x)_j (H_{ij}^+ - H_{ij}^*)$$
$$= \sum_{j=1}^n (Hy - \Delta x)_j \sum_{i=1}^n y_i (H_{ij}^+ - H_{ij}^*) = \sum_{j=1}^n (Hy - \Delta x)_j (H_j^+ - H_j^*) y$$
$$= \sum_{j=1}^n (Hy - \Delta x)_j (\Delta x_j - \Delta x_j) = 0.$$

Proof (43) is complete. Next, we prove that ΔH is the normal of the hyperplane of matrices satisfying the condition in (42). To do this, we prove that the vector ΔH is orthogonal to an arbitrary vector of the hyperplane, formed as the difference of matrices belonging to the hyperplane $V = H^1 - H^2$, that is, $<\Delta H, H^1 - H^2> = 0$. Since the matrices H^1 and H^2 satisfy the condition in (42), the proof is identical to the justification of the equality in (43). □

The following theorem establishes the convergence rate for a series of successive updates (39).

Theorem 6. *Let vectors y_k, $k = l, l + 1, \ldots, l + n - 1$, for the sequence of n iterations of (39) be mutually orthogonal. Then, the solution to the minimization problem in (29) can be obtained in no more than n steps of the process in (39),*

$$H^{k+1} = H_G(H^k, \Delta x_k, y_k), \; k = l, l+1, \ldots, l+n-1 \; , \tag{44}$$

for an arbitrary symmetric matrix H^l:

$$R^{l+n} = W_l^{l+n-1}(y) R^{l+n} [W_l^{l+n-1}(y)]^T = 0 \, . \tag{45}$$

Proof of Theorem 6. The update in (45) can be represented as two successive multiplications by $W_l^{l+n-1}(y)$, first from the left and then from the right. For each of the updates, the estimate in (35) is valid. □

Theorem 7. *Let vectors y_k, $k = 0, 1, \ldots, n - 1$, be given, vectors $y_k/\|y_k\|$ be columns of matrix P, and the minimum eigenvalue μ of a matrix P^TP satisfy the constraint $\mu \geq \mu_0 > 0$. Then, to estimate the convergence rate of the process in (44), the following inequality holds:*

$$\| H^n - H^* \|^2 \leq \| H^0 - H^* \|^2 \exp\left(- \frac{\mu_0^2}{n^3} \right). \tag{46}$$

Proof of Theorem 7. The matrix residual is updated according to the rule

$$R^{l+n} = W_l^{l+n-1}(y) R^{l+n} [W_l^{l+n-1}(y)]^T \, ,$$

which can be represented as two successive multiplications by $W_l^{l+n-1}(y)$, first from the left and then from the right. The estimate in (36) is valid for each of the updates, which proves (46). □

4. Symmetric Positive Definite Metric and Its Analysis

Let Function (6) be quadratic. We use the coordinate transformation

$$\hat{x} = Vx. \tag{47}$$

Let the matrix V satisfy the relation

$$V^{\mathrm{T}} V = \nabla^2 f(x) = A. \tag{48}$$

In the new coordinate system, the minimized function takes the following form:

$$\hat{f}(\hat{x}) = f(V^{-1}\hat{x}) = f(x). \tag{49}$$

Quadratic Function (6), considering (49), (47), and (48), takes the following form:

$$\hat{f}(\hat{x}) = \tfrac{1}{2}(\hat{x} - \hat{x}^*)^T V^{-T} A V^{-1} (\hat{x} - \hat{x}^*) = \tfrac{1}{2} \langle \hat{x} - \hat{x}^*, \hat{x} - \hat{x}^* \rangle. \tag{50}$$

Here, $\hat{x}^*$ is the minimum point of the function. According to (38) and (50), the matrix of second derivatives is the identity matrix $\nabla^2 \hat{f}(\hat{x}) = I$. Let us denote $\hat{r} = \hat{x} - \hat{x}^*$. The gradient is

$$\nabla \hat{f}(\hat{x}) = r(\hat{x}) = \hat{r} = \hat{x} - \hat{x}^*. \tag{51}$$

For the characteristics of functions $\hat{f}(\hat{x})$ and $f(x)$, the following relationships are valid:

$$\nabla \hat{f}(\hat{x}) = V^{-T} \nabla f(x), \quad \nabla^2 \hat{f}(\hat{x}) = V^{-T} \nabla^2 f(x) V^{-1}, \tag{52}$$

$$\Delta \hat{x} = \hat{x}^+ - \hat{x} = Vx^+ - Vx = V\Delta x, \tag{53}$$

$$\hat{y} = \nabla \hat{f}(x^+) - \nabla \hat{f}(x) = V^{-T} f(x^+) - V^{-T} f(x) = V^{-T} y \tag{54}$$

where notation $V^{-\mathrm{T}} = (V^{\mathrm{T}})^{-1}$ is used.

From (53), (54), and the properties of matrices V (48), the following equality holds:

$$\hat{y} = \Delta \hat{x} \equiv z \tag{55}$$

For the symmetric matrix $\widehat{H}$, two products of the matrix $\widehat{H}^*$ and the vector y are known:

$$\hat{H}^* \hat{y} = \Delta \hat{x}, \quad \hat{y}^T \hat{H}^* = \Delta \hat{x}^T. \tag{56}$$

Applying the process in (28) twice to (56), we obtain a new process for updating the matrix residual $\hat{R} = \hat{H} - I$:

$$\hat{R}^+ = W(\hat{y})\, \hat{R}\, W(\hat{y}) = W(z)\, \hat{R}\, W(z). \tag{57}$$

Taking into account (55), the update in (39) takes the form

$$\widehat{H}^+_{BFGS} = \hat{H}_G\left(\widehat{H}, \Delta \hat{x}, \hat{y}\right) = \widehat{H} + \frac{\langle \widehat{H}z - z, z\rangle zz^T}{\langle z, z\rangle^2} - \frac{z\left(\widehat{H}z - z\right)^T + \left(\widehat{H}z - z\right)z^T}{\langle z, z\rangle}. \tag{58}$$

Let us consider the methods in (1)-(4) in relation to the function $\hat{f}(\hat{x})$ in the new coordinate system.

$$\hat{x}^{k+1} = \hat{x}^k + \hat{\beta}_k \hat{s}^k, \quad \hat{s}^k = -\widehat{H}^k \nabla \hat{f}\left(\hat{x}^k\right), \tag{59}$$

$$\hat{\beta}_k = \arg\min_{\hat{\beta} \geq 0} \hat{f}\left(\hat{x}^k + \hat{\beta}\hat{s}^k\right), \tag{60}$$

$$\Delta \hat{x}^k = \hat{x}^{k+1} - \hat{x}^k = z^k, \ \hat{y}^k = \nabla \hat{f}(\hat{x}^{k+1}) - \nabla \hat{f}(\hat{x}^k) = z^k, \tag{61}$$

$$\hat{H}^{k+1} = H\left(\hat{H}^k, \Delta \hat{x}^k, \hat{y}^k\right). \tag{62}$$

Parameter $\hat{\beta}_k$ in (59) characterizes the accuracy of a one-dimensional descent. If the matrices are correlated by

$$\widehat{H}^k = VH^k V^T, \quad H^k = V^{-1} \widehat{H}^k V^{-T}, \tag{63}$$

and the initial conditions are

$$\hat{x}^0 = Vx^0, \quad \widehat{H}^0 = VH^0 V^T, \tag{64}$$

then these processes generate identical sequences $\hat{f}(\hat{x}^k) = f(x^k)$ and characteristics connected by the relations in (47) and (52)–(54). In this case, the equality $\hat{\beta}_k = \beta_k$ holds.

Considering the equality $\hat{y} = \Delta\hat{x}$ from (55), Equation (58) can be transformed. As a result, we obtain the BFGS equation:

$$H_{BFGS}(\widehat{H}, \Delta\hat{x}, \hat{y}) = \widehat{H} - \frac{(\Delta\hat{x} - \widehat{H}\hat{y}, \hat{y})\Delta\hat{x}\,\Delta\hat{x}^T}{\langle\hat{y}, \Delta\hat{x}\rangle^2} + \frac{(\Delta\hat{x} - \widehat{H}\hat{y})\Delta\hat{x}^T + \Delta\hat{x}(\Delta\hat{x} - \widehat{H}\hat{y})^T}{\langle\hat{y}, \Delta\hat{x}\rangle}. \quad (65)$$

Equation (65) satisfies the requirement of (63) and has the same form in various coordinate systems. Similar properties have the matrix transformation equation H_{DFP}, which can be represented as a transformed formula H_{BFGS} [29–31]:

$$H_{DFP}(\widehat{H}, \Delta\hat{x}, \hat{y}) = H_{BFGS}(\widehat{H}, \Delta\hat{x}, \hat{y}) - vv^T,$$
$$v = \langle\hat{y}, \hat{H}\hat{y}\rangle^{\frac{1}{2}}\left[\frac{\Delta\hat{x}}{\langle\Delta\hat{x}, \hat{y}\rangle} - \frac{\hat{H}\hat{y}}{\langle\hat{y}, \hat{H}\hat{y}\rangle}\right]. \quad (66)$$

Taking into account (55) and (58), we obtain the following expression in the new coordinate system:

$$\hat{H}_{DFP} = \hat{H}_{BFGS} - \hat{v}\hat{v}^T, \ \hat{v} = \langle z, \widehat{H}z\rangle^{\frac{1}{2}}\left[\frac{z}{\langle z, z\rangle} - \frac{\hat{H}z}{\langle z, \hat{H}z\rangle}\right]. \quad (67)$$

The form of the matrices in (65) and (66) does not change depending on the coordinate system. Consequently, the form of the processes in (1)–(4) and (59)–(62) is completely identical in different coordinate systems when using Formulas (65) and (67). Thus, for further studies of the properties of QNMs on quadratic functions, we can use Equations (58) and (67) in the coordinate system specified by the transformation in (47).

Within the iteration of the processes in (59)–(62) for a quadratic function with an identity matrix of second derivatives, the residual can be represented in the form of components

$$\hat{r}^k \equiv r(\hat{x}^k) = \hat{r}_z^k + \hat{r}_{\perp z}^k, \quad (68)$$

where $\hat{r}_z^k$ is a component along the vector z^k (or, which is the same, along $\hat{s}^k$), and $\hat{r}_{\perp z}^k$ is a component orthogonal to z^k. With an inexact one-dimensional descent in (59), the component $\hat{r}_z^k$ decreases but does not disappear completely. For the convenience of theoretical studies, the residual transformation in Equation (68) in this case can be represented by introducing parameter $\gamma_k \in (0, 2)$ instead of $\hat{\beta}_k$, characterizing the degree of descent accuracy:

$$\hat{r}^{k+1} = W(z^k, \gamma_k)\hat{r}^k = (1 - \gamma_k)\hat{r}_z^k + \hat{r}_{\perp z}^k, \ \ W(z, \gamma) = I - \gamma\frac{zz^T}{z^T z}, \ \ \gamma_k \in (0, 2). \quad (69)$$

Here, at arbitrary $\gamma_k \in (0, 2)$, the objective function decreases. With an inexact one-dimensional descent, a certain value $\gamma_k \in (0, 2)$ will be attained, at which the new value of the function becomes smaller.

The restriction on the one-dimensional search in (59), imposed on γ_k in (69), ensures a reduction in the objective function

$$\hat{f}(\hat{x}^{k+1}) = ||\hat{r}^{k+1}||^2/2 = ||W(z^k, \gamma_k)\hat{r}^k||^2/2$$
$$= (1 - \gamma_k)^2||\hat{r}_z^k||^2 + ||\hat{r}_{\perp z}^k||^2 < ||\hat{r}_z^k||^2 + ||\hat{r}_{\perp z}^k||^2 = \hat{f}(\hat{x}^k).$$

As a result of the iterations in (59)–(62) with (65) and according to (57), the matrix residual $\hat{R}^k = \hat{H}^k - \hat{H}^* = \hat{H}^k - I$ is transformed according to the rule

$$\hat{R}^{k+1} = W(z^k)\hat{R}^k W(z^k). \quad (70)$$

Therefore, one system of vectors z^k is used in the new coordinate system of the QNM iteration with the aim of minimizing the function and residual functional for matrices (29). With the orthogonality of vectors z^k and an exact one-dimensional search, the solution $\hat{r}^k = 0$ will be obtained in no more than n iterations. By virtue of the equality

$\langle z^i, z^j \rangle = \langle A\Delta x^i, \Delta x^j \rangle$, the orthogonality of vectors z^k in the chosen coordinate system is equivalent to the conjugacy of vectors Δx^k.

Due to the type of identity which defines the QNM iteration in different coordinate systems, we further denote the iteration of processes (59)–(62) and (1)–(4), considering the accuracy of one-dimensional descent (introduced in (69) by the parameter $\gamma_k \in (0,2)$) by the operator

$$QN(x^k, H^k, x^{k+1}, H^{k+1}, \gamma_k). \tag{71}$$

To simplify the notation in further studies of quasi-Newton methods on quadratic functions, without a loss of generality, we use an iteration of the method in (71) adjusted to minimize the function

$$f(x) = \frac{1}{2}\langle x - x^*, x - x^* \rangle, \tag{71a}$$

which allows us, without transforming the coordinate system (47), to use all associated relations for the processes in (59)–(62) with the function in (50) for studying the process in (71), omitting the hats above the variables in the notation.

Let us note some of the properties of the QNM.

Theorem 8. *Let $H^k > 0$ and the iteration of (71) be carried out with matrix transformation equations H_{BFGS} and H_{DFP} (67). Then, the vector z^k is an eigenvector of the matrices H_{BFGS}^{k+1}, H_{DFP}^{k+1}, R_{BFGS}^{k+1}, and R_{DFP}^{k+1}:*

$$R_{BFGS}^{k+1} z^k = 0, \quad H_{BFGS}^{k+1} z^k = z^k. \tag{72}$$

$$R_{DFP}^{k+1} z^k = 0, \quad H_{DFP}^{k+1} z^k = z^k. \tag{73}$$

Proof of Theorem 8. The first of the equalities in (72) follows from (70). The second of the equalities in (72) follows from this fact and the definition of the matrix residual.

By direct verification, based on (67), we establish that the vectors z^k and v^k are orthogonal. Therefore, the additional term $v\,v^{\mathrm{T}}$ in Equation (67) does not affect the multiplication of vector z^k by a matrix, which together with (72) proves (73). □

As consequence of Theorem 8, the dimension of the space being minimized is reduced by one in the case of an exact one-dimensional descent, which will be shown below. Section 5 justifies the advantages of the BFGS equation (65) over the DFP equation (66) for matrix transformation.

5. Qualitative Analysis of the Advantages of the BFGS Equation over the DFP Equation

The effectiveness of the learning algorithm is determined by the degree of orthogonality of the learning vectors in the operator factors $W_{k-m}^k(y)$. In the new coordinate system, the transformation in (70) is determined by the factors $W_{k-m}^k(z)$ in the residual expressions. Therefore, to analyze the orthogonality degree of the system of vectors z, it is necessary to involve the method of their formation. Let us show that the vectors z^k in (69) and (70) generated by the BFGS equation have a higher degree of orthogonality compared to those generated by DFP. To get rid of a large number of indices, consider the iteration of the QNM (71) in the form

$$QN(\hat{f}, \hat{x}, \hat{H}, \hat{x}^+, \hat{H}^+, \gamma). \tag{74}$$

Theorem 9. *Let $\hat{H} > 0$ and the iteration of (74) be carried out with the matrix updating equations $\hat{H}_{BFGS}$ (58) and $\hat{H}_{DFP}$ (67), and*

$$||\hat{v}|| \neq 0. \tag{75}$$

Then, the following statements are valid.
1. The descent directions for the next iteration are of the form

$$\hat{s}^{+}_{BFGS} = \hat{H}^{+}_{BFGS}\hat{r}^{+} = (1-\gamma_k)\hat{r}_z + \langle z, \hat{H}z\rangle^{\frac{1}{2}}\hat{v}, \tag{76}$$

$$\hat{s}^{+}_{DFP} = \hat{H}^{+}_{DFP}\hat{r}^{+} = (1-\gamma_k)\hat{r}_z + q\langle z, \hat{H}z\rangle^{\frac{1}{2}}\hat{v}, \tag{77}$$

where

$$0 < q = \frac{\langle \hat{H}\hat{r}, \hat{H}\hat{r}\rangle^2}{\langle \hat{r}, \hat{H}\hat{r}\rangle\langle \hat{H}\hat{H}\hat{r}, \hat{H}\hat{r}\rangle} < 1. \tag{78}$$

2. With respect to the cosine of the angle between adjacent directions of the descent, we have the following estimate:

$$\frac{\langle \hat{s}^{+}_{BFGS}, z\rangle^2}{\langle z,z\rangle\langle \hat{s}^{+}_{BFGS}, \hat{s}^{+}_{BFGS}\rangle} \leq \frac{\langle \hat{s}^{+}_{DFP}, z\rangle^2}{\langle z,z\rangle\langle \hat{s}^{+}_{DFP}, \hat{s}^{+}_{DFP}\rangle}. \tag{79}$$

3. In the subspace of vectors orthogonal to z, the trace of the matrix $\hat{H}^{+}_{BFGS}$ *does not change,*

$$sp_{\perp z}(\hat{H}^{+}_{BFGS}) = sp_{\perp z}(\hat{H}), \tag{80}$$

and the trace of the matrix $\hat{H}^{+}_{DFP}$ *decreases,*

$$sp_{\perp z}(\hat{H}^{+}_{DFP}) = sp_{\perp z}(\hat{H}) - \frac{\langle \hat{v}, \hat{H}z\rangle^2}{\langle \hat{v}, \hat{v}\rangle\langle z, \hat{H}z\rangle} < sp_{\perp z}(\hat{H}). \tag{81}$$

Proof of Theorem 9. We represent the residual, similarly to (69), in the following form:

$$\hat{r} = \hat{r}_z + \hat{r}_{\perp z}, \quad ||\hat{r}_{\perp z}|| \neq 0. \tag{82}$$

After performing the iteration of (74), the residual takes the form

$$\hat{r}^{+} = W(z,\gamma)\hat{r} = (1-\gamma)\hat{r}_z + \hat{r}_{\perp z}. \tag{83}$$

According to (83), in $\hat{r}^{+}$, the component $\hat{r}_{\perp z}$ does not depend on the accuracy of the one-dimensional search. Therefore, initially, we find new descent directions in (76) and (77) under the condition of an exact one-dimensional search, that is, with $\hat{r}^{+} = \hat{r}_{\perp z}$.

Considering the gradient expression in (51), the direction of minimization in the iteration of (74) is $\hat{s} = -\hat{H}\hat{r}$. Based on that result, considering (55) and the equality $\langle \hat{r}^{+}, z\rangle = 0$, following from the condition of exact one-dimensional minimization (60), we obtain

$$\hat{r}^{+} = W(z)\hat{r} = \hat{r} + z = \hat{r} - \hat{H}\hat{r}\frac{\langle \hat{r}, \hat{H}\hat{r}\rangle}{\langle \hat{H}\hat{r}, \hat{H}\hat{r}\rangle}. \tag{84}$$

This implies

$$z = -\hat{H}\hat{r}\frac{\langle \hat{r}, \hat{H}\hat{r}\rangle}{\langle \hat{H}\hat{r}, \hat{H}\hat{r}\rangle}, \tag{85}$$

$$\hat{H}\hat{r} = -z\frac{\langle \hat{H}\hat{r}, \hat{H}\hat{r}\rangle}{\langle \hat{r}, \hat{H}\hat{r}\rangle} = -z\frac{\langle \hat{H}\hat{r}, z\rangle}{\langle \hat{r}, z\rangle}. \tag{86}$$

From (84), taking into account the orthogonality of the vectors $\hat{r}^{+}$, z, we obtain the equality

$$\langle \hat{r}, z\rangle = -\langle z, z\rangle. \tag{87}$$

Let us find the expression $\hat{H}^{+}\hat{r}^{+}$ necessary to form the descent direction $\hat{s}^{+} = -\hat{H}^{+}\hat{r}^{+}$ in the next iteration. Considering the orthogonality of the vectors $\hat{r}^{+}$ and z, using the BFGS matrix transformation formula (58), we obtain

$$\begin{aligned}\widehat{H}^{+}\hat{r}^{+} &= \widehat{H}\,\hat{r}^{+} + z\frac{\langle z-\widehat{H}z,\hat{r}^{+}\rangle}{\langle z,z\rangle} = \widehat{H}\,\hat{r}^{+} - z\frac{\langle \widehat{H}z,\hat{r}^{+}\rangle}{\langle z,z\rangle}\\ &= \widehat{H}\,\hat{r} + \widehat{H}z - z\frac{\langle \widehat{H}z,\hat{r}^{+}\rangle}{\langle z,z\rangle}\\ &= \widehat{H}\,\hat{r} - z\frac{\langle \widehat{H}z,\hat{r}\rangle}{\langle z,z\rangle} + \widehat{H}z - z\frac{\langle \widehat{H}z,z\rangle}{\langle z,z\rangle}.\end{aligned} \tag{88}$$

Transformation of the equality in (86) based on (87) leads to

$$\widehat{H}\hat{r} = -z\frac{\langle \widehat{H}\hat{r},z\rangle}{\langle \hat{r},z\rangle} = z\frac{\langle \widehat{H}\hat{r},z\rangle}{\langle z,z\rangle} = z\frac{\langle \widehat{H}z,\hat{r}\rangle}{\langle z,z\rangle}. \tag{89}$$

Making the replacement (89) in the last expression from (88), we find

$$\widehat{H}^{+}\hat{r}^{+} = \widehat{H}z - z\frac{\langle \widehat{H}z,z\rangle}{\langle z,z\rangle}. \tag{90}$$

According to (90), the new descent vector can be represented using the expression for $\hat{v}$ from (67)

$$\hat{s}^{+} = -\widehat{H}^{+}\hat{r}^{+} = z\frac{\langle \widehat{H}z,z\rangle}{\langle z,z\rangle} - \widehat{H}z = \langle \widehat{H}z,z\rangle^{-\frac{1}{2}}\hat{v}. \tag{91}$$

Since the component $\hat{r}_{\perp z}$ in (83) does not depend on the accuracy of the one-dimensional search, Expression (91) determines its contribution to the direction of descent in (76). Finally, the property of (72) together with the residual $\hat{r}$ representation in (82) proves (76).

The condition in (75) according to (91) prevents the completion of the minimization process. If $\hat{v}=0$, then as a result of exact one-dimensional minimization, we obtain $\hat{s}^{+} = -\widehat{H}^{+}\hat{r}^{+} = \langle \widehat{H}z,z\rangle^{-0.5}\hat{v} = 0$, which, taking into account $\widehat{H} > 0$, means $\hat{r}^{+}=0$. As before, using (67), we find a new descent direction for the DFP method, assuming that the one-dimensional search is exact:

$$\hat{s}^{+}_{DFP} = -\widehat{H}^{+}_{DFP}\hat{r}^{+} = -\widehat{H}^{+}_{BFGS}\hat{r}^{+} + \hat{v}\langle \hat{v},\hat{r}^{+}\rangle = \hat{s}^{+}_{BFGS} + \hat{v}\langle \hat{v},\hat{r}^{+}\rangle\,. \tag{92}$$

The last term in (92), taking into account (91) and the orthogonality of the vectors $\hat{r}^{+}, z$, can be represented in the form

$$\begin{aligned}\hat{v}\langle \hat{v},\hat{r}^{+}\rangle &= \langle z,\widehat{H}z\rangle\langle \frac{z}{\langle z,z\rangle} - \frac{\widehat{H}z}{\langle z,\widehat{H}z\rangle},\hat{r}^{+}\rangle\left[\frac{z}{\langle z,z\rangle} - \frac{\widehat{H}z}{\langle z,\widehat{H}z\rangle}\right] = -\langle \frac{\widehat{H}z}{\langle z,\widehat{H}z\rangle},\hat{r}^{+}\rangle\,\hat{s}^{+}_{BFGS}\\ &= -\langle \frac{\widehat{H}z}{\langle z,\widehat{H}z\rangle},\hat{r}+z\rangle\,\hat{s}^{+}_{BFGS} = \left(-\frac{\langle \widehat{H}z,\hat{r}\rangle}{\langle z,\widehat{H}z\rangle} - 1\right)\hat{s}^{+}_{BFGS}.\end{aligned} \tag{93}$$

Let us transform the scalar value as follows:

$$q = -\frac{\langle \widehat{H}z,\hat{r}\rangle}{\langle \widehat{H}z,z\rangle} = -\frac{\langle \widehat{H}\,\widehat{H}\,\hat{r},\hat{r}\rangle}{\langle \widehat{H}\,\widehat{H}\,\hat{r},z\rangle} = \frac{\langle \widehat{H}\,\widehat{H}\,\hat{r},\hat{r}\rangle^{2}}{\langle \widehat{H}\,\widehat{H}\,\hat{r},\widehat{H}\hat{r}\rangle\langle \widehat{H}\,\hat{r},\hat{r}\rangle}. \tag{94}$$

Based on (92), together with (93) and (94), we obtain the expression

$$\hat{s}^{+}_{DFP} = -\widehat{H}^{+}_{DFP}\hat{r}^{+} = \hat{s}^{+}_{BFGS} + (q-1)\hat{s}^{+}_{BFGS} = q\hat{s}^{+}_{BFGS}.$$

And finally, the last expression, using the property of (73) together with the representation of the residual, considering the accuracy of the one-dimensional descent (82), proves (77).

Since $\widehat{H} > 0$, the left inequality in (78) will hold. We prove the right inequality by contradiction. Let us denote by $\widehat{H}^{L} > 0 (L > 0)$ a matrix with eigenvectors of the matrix H and eigenvalues in the form of powers of the corresponding eigenvalues of the matrix H, given by $\lambda_i^{\widehat{H}^L} = (\lambda_i^{H})^{L}, i = 1,2,\ldots,n$. Let $u = (\widehat{H})^{0.5}\hat{r}$. Then,

$$q = \langle \widehat{H}u,u\rangle^{2}/\langle \widehat{H}u,\widehat{H}u\rangle\langle u,u\rangle.$$

Consequently, the equality $\widehat{H}u = \rho u$ holds if q = 1. Therefore, u is an eigenvector of the matrix H, and therefore, all matrices $\widehat{H}^L$ also have such an eigenvector. Due to this fact and the equality $u = (\widehat{H})^{1/2}\hat{r}$, the vector $\hat{r}$ is also an eigenvector, and $u = (\widehat{H})^{1/2}\hat{r} = \rho^{1/2}\hat{r}$, where ρ is the eigenvalue of the matrix $\widehat{H}$. In this case, considering the representation in (85) of vector z, vector $\hat{v}$, according to its representation in (67), is zero, which cannot be true according to the condition in (75). Therefore, the right inequality in (78) also holds.

Due to the orthogonality of vectors $\hat{v}$ and z and according to (76) and (77), the numerators in (79) are the same, and for the denominators, taking into account (78), the inequality $\langle \hat{s}^+_{DFP}, \hat{s}^+_{DFP}\rangle < \langle \hat{s}^+_{BFGS<}, \hat{s}^+_{BFGS}\rangle$ holds, which proves (79). In an exact one-dimensional search, the equality is satisfied in (79) since the numerators in (79) are zero.

Let us justify point 3 of the theorem. In accordance with the notation of equations H_{BFGS} (58) and H_{DFP} (67), we introduce an orthogonal coordinate system in which the first two orthonormal vectors are determined by the following equations:

$$e_1 = z/\|z\| \ , e_2 = p/\|p\|, p = \widehat{H}z - z\frac{\langle z,\widehat{H}z\rangle}{\langle z,z\rangle}, \tag{95}$$

where vectors p and z are orthogonal and $\hat{v} = -\langle z, \widehat{H}z\rangle^{-1/2}p$. In such a coordinate system, these vectors are defined by

$$z^T = (||z||,0,\ldots,0) \ \ p^T = (0,||p||,0,\ldots,0). \tag{96}$$

Let us consider the form of matrix $\hat{H}$ in the selected coordinate system. Let us determine the type of vector p based on its representation in (95). Taking into account $\left\langle z, \hat{H}z\right\rangle / \left\langle z, z\right\rangle = \hat{H}_{1,1}$, components of vector p have the form

$$(\widehat{H}z)^T = ||z||\left(\widehat{H}_{1,1}, \widehat{H}_{2,1}, \widehat{H}_{3,1}, \ldots, \widehat{H}_{n,1}\right), \quad z^T\frac{\langle z,\widehat{H}z\rangle}{\langle z,z\rangle} = ||z||(\widehat{H}_{1,1}, 0,\ldots,0).$$

Hence, $p^T = ||z||\left(0, \widehat{H}_{2,1}, \widehat{H}_{3,1}, \ldots, \widehat{H}_{n,1}\right)$. Comparing the last expression with the expression in (96), we conclude that in the chosen coordinate system, the first column $\hat{H}_1$ of matrix $\hat{H}$ has the following form:

$$\widehat{H}_1 = \left(\widehat{H}_{11}, \widehat{H}_{21}, 0,\ldots,0\right)^T. \tag{97}$$

From (97) and (96), it follows that

$$p^T = ||z||\left(0, \widehat{H}_{2,1}, 0,\ldots,0\right), \ \hat{v} = -\langle z, \widehat{H}z\rangle^{-1/2}, \ p = \left(0, \widehat{H}_{2,1}/\widehat{H}^{1/2}_{1,1}, 0,\ldots,0\right), \tag{98}$$

and the original matrix will have the form

$$\widehat{H} = \begin{pmatrix} \widehat{H}_{11} & \widehat{H}_{12} & 0 & \ldots & 0 \\ \widehat{H}_{21} & \widehat{H}_{22} & \widehat{H}_{23} & \ldots & \widehat{H}_{2n} \\ 0 & \widehat{H}_{32} & \widehat{H}_{33} & \ldots & \widehat{H}_{3n} \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 0 & \widehat{H}_{n2} & \widehat{H}_{n3} & \ldots & \widehat{H}_{nn} \end{pmatrix}. \tag{99}$$

When correcting matrices with formulas BFGS (58) and DFP (67), changes will occur only in the space of the first two variables, determined by the unit vectors in (95). As a result of the BFGS transformation in (58), we obtain the following two-dimensional matrix:

$$\begin{aligned} \widehat{H}^+_{2\times 2_{BFGS}} = \begin{pmatrix} \widehat{H}_{11} & \widehat{H}_{12} \\ \widehat{H}_{12} & \widehat{H}_{22} \end{pmatrix} + \begin{pmatrix} \widehat{H}_{11}-1 & 0 \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} \widehat{H}_{11}-1 & \widehat{H}_{12} \\ 0 & 0 \end{pmatrix} - \begin{pmatrix} \widehat{H}_{11}-1 & 0 \\ \widehat{H}_{12} & 0 \end{pmatrix} \\ = \begin{pmatrix} 1 & 0 \\ 0 & \widehat{H}_{22} \end{pmatrix}. \end{aligned} \tag{100}$$

Based on the relationship of matrices expressed in (67), using (98), we obtain the result of the transformation according to the DFP equation in (67):

$$\widehat{H}^{+}_{2\times 2\_DFP} = \widehat{H}^{+}_{2\times 2\_BFGS} - \hat{v}\hat{v}^{T} = \begin{pmatrix} 1 & 0 \\ 0 & \widehat{H}_{22} - \frac{\widehat{H}^{2}_{12}}{\widehat{H}_{11}} \end{pmatrix}. \quad (101)$$

Thus, the resulting two-dimensional matrices have the following form:

$$\widehat{H}^{+}_{2\times 2\_BFGS} = \begin{pmatrix} 1 & 0 \\ 0 & \widehat{H}_{22} \end{pmatrix}. \ \widehat{H}^{+}_{2\times 2\_DFP} = \begin{pmatrix} 1 & 0 \\ 0 & \widehat{H}_{22} - \frac{\widehat{H}^{2}_{12}}{\widehat{H}_{11}} \end{pmatrix}. \quad (102)$$

The corresponding complete matrices are presented below:

$$\widehat{H}^{+}_{BFGS} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \widehat{H}_{22} & \widehat{H}_{23} & \dots & \widehat{H}_{2n} \\ 0 & \widehat{H}_{32} & \widehat{H}_{33} & \dots & \widehat{H}_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \widehat{H}_{n2} & \widehat{H}_{n3} & \dots & \widehat{H}_{nn} \end{pmatrix}, \quad (103)$$

$$\widehat{H}^{+}_{DFP} = \begin{pmatrix} 1 & 0 & 0 & \dots & 0 \\ 0 & \widehat{H}_{22} - \frac{\widehat{H}^{2}_{12}}{\widehat{H}_{11}} & \widehat{H}_{23} & \dots & \widehat{H}_{2n} \\ 0 & \widehat{H}_{32} & \widehat{H}_{33} & \dots & \widehat{H}_{3n} \\ \dots & \dots & \dots & \dots & \dots \\ 0 & \widehat{H}_{n2} & \widehat{H}_{n3} & \dots & \widehat{H}_{nn} \end{pmatrix}. \quad (104)$$

Due to the condition in (75) from Expression (98) for $\hat{v}$, it follows that $\widehat{H}_{2,1} \neq 0$. Consequently, the trace of matrix $\widehat{H}^{+}_{DFP}$, according to (102) and (104), will decrease by $\widehat{H}^{2}_{12}/\widehat{H}_{11}$. The last expression can be transformed considering the definition of the coordinate system in (96). As a result, we obtain (81). From (103), we obtain (80). □

Regarding the results of Theorem 9, we can draw the following conclusions.

1. With an inexact one-dimensional descent in the DFP method, the successive descent directions are less orthogonal than in the BFGS method (79).
2. The trace of matrix $\widehat{H}$ in the DFP method in the unexplored space decreases (81). This makes it difficult to enter a new subspace during subsequent minimization. Moreover, in the case of an exact one-dimensional descent, in the next step, this decrease is restored; however, a new one appears.
3. Theorem 9 also shows that in the case of an exact one-dimensional search, the minimization space on quadratic functions is reduced by one.

Due to the limited computational accuracy on ill-conditioned problems (i.e., problems with a high condition number), the noted effects can significantly worsen the convergence of the DFP method.

In conjugate gradient methods [39], if the accuracy of the one-dimensional descent is violated, the sequence of vectors ceases to be conjugated. In QNMs, due to the reduction in the minimization subspace by one during exact one-dimensional descent, the effect of reducing the minimization space accumulates. In Section 6, we look at methods for replenishing the space excluded from the minimization process.

6. Methods for Reducing the Minimization Space of Quasi-Newton Methods on Quadratic Functions

We will assume that the quadratic function has the form expressed in (71a):

$$f(x) = \frac{1}{2}\langle x - x^{*}, x - x^{*}\rangle.$$

For matrices H^{k+1} and R^{k+1} obtained using the iteration of (71), $QN(x^{k}, H^{k}, x^{k+1}, H^{k+1}, \gamma_{k})$, the relations in (72) and (73) hold:

$$R^{k+1}z^{k} = 0, \ H^{k+1}z^{k} = z^{k}. \quad (105)$$

Vector z^k is an eigenvector for matrices H^{k+1} and R^{k+1} with one and zero eigenvalues, respectively. Let us consider ways to increase the dimension of the quasi-Newton relations' execution subspace.

Let us denote by $H \in I_m$ a matrix $H > 0$ that has m eigenvectors with unit eigenvalues, and the corresponding matrix $R = H - I$ with the corresponding eigenvectors and zero eigenvalues we will denote by $R \in O_m$. Let us denote by Q_m a subspace of dimension m spanned by a system of eigenvectors with unit eigenvalues of the matrix $H \in I_m$, and its complement by $D_m = R^n \backslash Q_m$.

An arbitrary orthonormal system of m vectors $e_1, \ldots, e_m$, of subspace Q_m is a system of eigenvectors of matrices $H \in I_m$ and $R \in O_m$:

$$H e_i = e_i \ , Re_i = 0 \ , \ \ i = 1, \ldots, m. \tag{106}$$

It follows that an arbitrary vector, which is a linear combination of vectors e_i, will satisfy the quasi-Newton relations.

Lemma 1. *Consider the matrix $H \in I_m$ and the vectors*

$$r = r_Q + r_D, \, r_Q \in Q_m, \, r_D \in D_m. \tag{107}$$

Then,

$$Hr = Hr_Q + Hr_D, \, Hr_Q = r_Q \in Q_m, \, Hr_D \in D_m. \tag{108}$$

Proof of Lemma 1. The system of m eigenvectors of matrix $H \in I_m$ is contained in the set Q_m. Due to the orthogonality of the eigenvectors, the remaining part of the matrix $H \in I_m$ is contained in the set D_m. Therefore, the operation of multiplying the vectors in (107) by the matrix in (108) does not take them beyond their subspace. In this case, for the vector r_Q, the equality $Hr_Q = r_Q \in Q_m$ holds, which follows from the definition of the subspace Q_m. □

Lemma 2. *Let $H^k > 0$, $H^k \in I_m$, $m < n$, $r_Q^k = 0, r_D^k \neq 0$, and iteration $QN(x^k, H^k, x^{k+1}, H^{k+1}, \gamma_k)$ be completed. Then,*

$$\text{if } \gamma_k = 1, \text{ then } H^{k+1} \in I_{m+1} \text{ and } \ r_Q^{k+1} = 0; \tag{109}$$

$$\text{if } \gamma_k \neq 1, \text{ then } H^{k+1} \in I_{m+1} \text{ and } \ r_Q^{k+1} \neq 0. \tag{110}$$

Proof of Lemma 2. The descent direction, taking into account (51), has the form $s^k = -H^k \nabla f(x^k) = -H^k r^k = -H^k r_D^k$. Based on Lemma 1, it follows that $H^k r_D^k \in D_m$. As follows from Theorem 8, a new eigenvector expressed in (72) and (73) with a unit eigenvalue appears in the subspace D_m, regardless of the accuracy of the one-dimensional descent, which proves (109), taking into account the accuracy of the one-dimensional search. With an inexact descent, part of the residual remains along the vector z^k, which proves (110). □

Lemma 3. *Let $H^k > 0$, $H^k \in I_m$, $m \leq n$, $r_Q^k \neq 0$, $r_D^k \neq 0$, and iteration $QN(x^k, H^k, x^{k+1}, H^{k+1}, \gamma_k)$ be completed. Then, it follows that*

$$\text{if } \gamma_k = 1, \text{ then } H^{k+1} \in I_m \text{ and } \ r_Q^{k+1} = 0; \tag{111}$$

$$\text{if } \gamma_k \neq 1, \text{ then } H^{k+1} \in I_m \text{ and } \ r_Q^{k+1} \neq 0. \tag{112}$$

Proof of Lemma 3. Since $r_Q^k \neq 0$, we take a system, where one of the eigenvectors is the vector r_Q^k, as an orthogonal system of eigenvectors in Q_m. From the remaining eigenvec-

tors, we form a subspace Q_{m-1} in which there is no residual. Applying to Q_{m-1} the results of Lemma 2 under the condition $H^k \in I_{m-1}$, we obtain (111) and (112). □

By alternating operations with an exact and inexact one-dimensional descent, it is possible to obtain finite convergence on quadratic functions of QNMs.

Theorem 10. *Let $H^k > 0$, $H^k \in I_m$, $r_Q^k \neq 0$, $m < n - 1$, and the iterations be completed as follows:*

$$QN(x^k, H^k, x^{k+1}, H^{k+1}, \gamma_k),\ \gamma_k = 1, \tag{113}$$

$$QN(x^{k+1}, H^k, x^{k+2}, H^{k+2}, \gamma_k),\ \gamma_k \neq 1. \tag{114}$$

Then,

$$H^{k+2} \in I_{m+1},\ r_Q^{k+2} \neq 0. \tag{115}$$

Proof of Theorem 10. For the iteration of (113), we apply the result of Lemma 3 (111), and for the iteration of (114), we apply the result of Lemma 2 (110). As a result, we obtain (115). □

Theorem 10 says that individual iterations with an exact one-dimensional descent make it possible to increase by one the dimension of the space where the quasi-Newton relation is satisfied. This means that after a finite number of such iterations, the matrix $H_k = I$ will be obtained.

Let us consider another way of increasing the dimension of the quasi-Newton relation. It consists of using, after iterations of QNMs, an additional iteration of descent along the orthogonal vector v^k defined in (67), and according to (91), with an exact one-dimensional descent coinciding, up to a scalar factor, with the descent direction $s^{k+1} = \left\langle H^k z^k, z^k \right\rangle^{-1/2} v^k$ of the BFGS method:

$$QN\left(x^k, H^k, x^{k+1/2}, H^{k+1/2}, \gamma_k\right),\ \gamma_k \in (0,2), \tag{116}$$

$$x^{k+1} = x^{k+1/2} + \beta_{k+1/2} v^k,\quad \gamma_k \in (0,2), \tag{117}$$

$$v^k = \langle z^k, H^k z^k \rangle^{\frac{1}{2}} \left[\frac{z^k}{\langle z^k, z^k \rangle} - \frac{H^k z^k}{\langle z^k, H^k z^k \rangle} \right], \tag{118}$$

$$H^{k+1} = H\left(H^{k+1/2}, \Delta x^{k+1/2}, y^{k+1/2}\right). \tag{119}$$

Let us denote the iterations in (116)–(119) by

$$VQN\left(x^k, H^k, x^{k+1}, H^{k+1}, \gamma_k, \gamma_{k+1/2}\right),\ \gamma_k \in (0,2),\ \gamma_{k+1/2} \in (0,2). \tag{120}$$

Lemma 4. *Let $H^k > 0$, $H^k \in I_m$, $r_Q^k \neq 0$, $r_D^k \neq 0$, $m \leq n - 1$, and the iteration of (120) be completed. Then,*

$$H^{k+1} \in I_{m+1},\ r_Q^{k+1} \neq 0. \tag{121}$$

Proof of Lemma 4. For the iteration of (116), as in the proof of Lemma 3, since $r_Q^k \neq 0$, we take this as an orthogonal system of eigenvectors in Q_m, where one of the eigenvectors is the vector r_Q^k. From the remaining eigenvectors, we form a subspace Q_{m-1} in which there is no residual, and for this subspace, $H^k \in I_{m-1}$ holds. As a result of (116), according to the results of Theorem 8, an eigenvector $z^k \notin Q_{m-1}$ is formed. It is a derivative of vector $s^k = -H^k r_Q^k \notin Q_{m-1}$, which, due to multiplication by a matrix $H^k \in I_{m-1}$ with residual $r_Q^k \notin Q_{m-1}$, according to the results of Lemma 1, does not belong to the subspace Q_{m-1}. For this

reason, the vector $v^k \notin Q_{m-1}$ obtained by Formula (118), orthogonal to z^k, because of (117)–(119), becomes an eigenvector of the matrix H^{k+1}. Thus, the subspace Q_{m-1} is replenished with two eigenvectors of the matrix H^{k+1}, resulting in (121). □

Theorem 11. *To obtain $H^k \in I_n$, it is necessary to perform the iteration of (120) $(n-1)$ times.*

Proof of Theorem 11. In the first iteration of (120), we obtain $H^{k+1} \in I_2$. In the next $(n-2)$ iterations of (120), according to the results of Lemma 4, we obtain $H^{k+n-1} \in I_n$. □

The results of Theorem 11 and Lemma 5 indicate the possibility of using techniques for increasing the dimension of the subspace of quasi-Newton relations' execution at arbitrary moments, which enables us, as will be shown below, to develop QNMs that are resistant to the inaccuracies of a one-dimensional search.

In summary, the following conclusions can be drawn about properties of QNMs on quadratic functions without the condition of an exact one-dimensional descent.

1. The dimension of the minimization subspace decreases as the dimension of the subspace of fulfillment of the quasi-Newton relation increases (Lemma 2).
2. The dimension of the subspace of fulfillment of the quasi-Newton relation does not decrease during the execution of the QNM (Lemmas 2–5).
3. Individual iterations with an exact one-dimensional descent increase the dimension of the subspace of the quasi-Newton relation (Lemma 4).
4. Separate inclusions of iterations with the transformation of matrices for pairs of conjugate vectors increase the dimension of the subspace of the quasi-Newton relation (Lemma 5).
5. It is sufficient to perform at most the $(n-1)$ inclusion of an exact one-dimensional descent (113) in arbitrary iterations to solve the problem of minimizing a quadratic function in a finite number of steps in the QNM (Lemma 4 and Theorem 10).
6. To solve the problem of minimizing a quadratic function in a finite number of steps in the QNM, it is sufficient to perform in arbitrary iterations no more than $(n-1)$ inclusions of matrix transformations for pairs of descent vectors obtained as a result of the transformations in (118) and (119) (Lemma 5 and Theorem 11).

7. Methods for Increasing the Orthogonality of Learning Vectors in Quasi-Newton Methods

The term "degree of orthogonality" refers to the type of function (71a). For the type of function (6), this term means the degree of conjugacy of the vectors. Several conclusions can be drawn from our considerations.

Firstly, it is preferable to use the BFGS method. With imprecise one-dimensional descent in the DFP method, successive descent directions are less orthogonal than in the BFGS method (79).

Secondly, it makes sense to increase the degree of accuracy of the one-dimensional search, since individual iterations with an exact one-dimensional descent increase the dimension of the subspace of the quasi-Newton relation (Theorem 10), which reduces the dimension of the minimum search region.

Thirdly, separate inclusions of iterations with matrix transformation for pairs of conjugate vectors increase the dimension of the subspace of the quasi-Newton relation (Lemma 4). This requires applying a sequence of descent iterations for pairs of conjugate vectors (120).

On the other hand, it is important to correctly select the scaling factor ω of the initial matrix $H^0 = \omega I$ from (1) in the QNM. Let us consider an example of a function of the form expressed in (6):

$$f(x) = \frac{1}{2}\sum_{i=1}^{n} x_i^2 / i \cdot \tag{122}$$

The eigenvalues of the matrix of second derivatives A and its inverse A^{-1} are $\lambda_i = \frac{1}{i}$ *and* $\lambda_i^{-1} = i$, respectively. The gradient of the quadratic function in (122) is $\nabla f(x) = \sum_{i=1}^{n} i\, x_i$. In the first stages of the search for $H^0 = I$, in the gradients $\nabla f(x) = A(x - x^*)$ and gradient differences, components of eigenvectors with large eigenvalues of matrix A and, accordingly, small eigenvalues of the matrix $A^{-1} = H$ prevail. Let us calculate an approximation of the eigenvalues for scaling the initial matrix using data from (3) of the first iteration of the methods in (1)–(4):

$$\lambda_{min}^H \le \omega = \frac{\langle \Delta x^0, \Delta x^0 \rangle}{\langle y^0, \Delta x^0 \rangle} = \frac{\langle A^{-1} y^0, A^{-1} y^0 \rangle}{\langle y^0, A^{-1} y^0 \rangle} \le \lambda_{max}^H, \tag{123}$$

where $\lambda_{\min}^H$, $\lambda_{\max}^H$ are the minimum and maximum eigenvalues of the matrix $A^{-1} = H$, respectively. To scale the initial matrix H^0, consider the following:

$$H^0 = K\omega I = K\frac{\langle \Delta x^0, \Delta x^0 \rangle}{\langle y^0, \Delta x^0 \rangle} I, K \ge 1. \tag{124}$$

Let us qualitatively investigate the operation of the quasi-Newton BFGS method (71). Taking into account the predominance of eigenvectors with large eigenvalues of the matrix A and, accordingly, small eigenvalues of the matrix $A^{-1} = H$, it is possible to qualitatively display the picture of the reconstruction of the matrix A^{-1} eigenvectors for different values of K, making a rough assumption that small eigenvalues are sequentially restored. A rough diagram of the process of reconstructing the spectrum of matrix eigenvalues is shown in Figure 2.

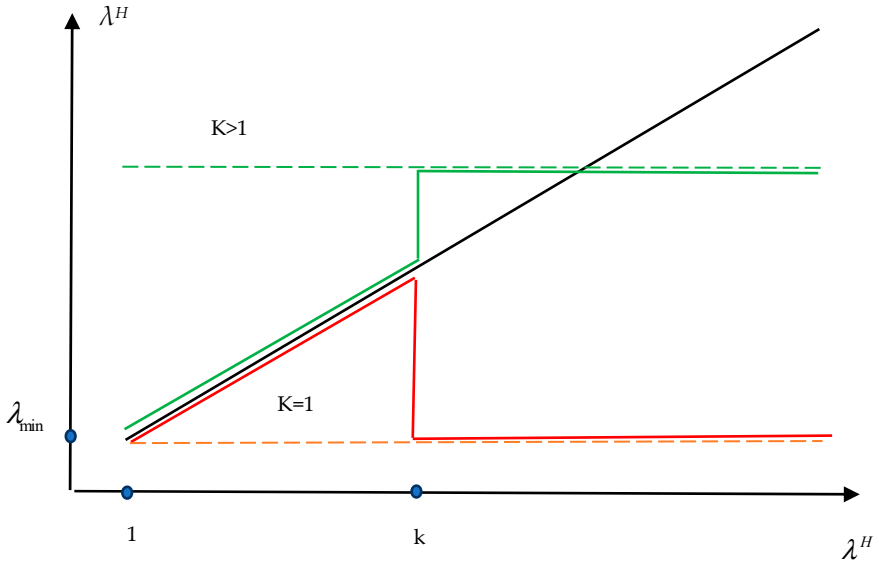

Figure 2. Qualitative behavior of the spectrum of matrix H^k eigenvalues for cases of scaling (124) for various values of K.

One of the components of increasing the degree of orthogonality of learning vectors in QNMs is the normalization of the initial metric matrix (124). In Section 8, we will consider the impact of the methods noted in this section on increasing the efficiency of QNMs.

8. Numerical Study of Ways to Increase the Orthogonality of Learning Vectors in Quasi-Newton Methods

We implemented and compared quasi-Newtonian BFGS and DFP methods. A one-dimensional search procedure with cubic interpolation [41] (exact one-dimensional descent) and a one-dimensional minimization procedure [34] (inexact one-dimensional descent) were used. We used both the classical QNM with the iterations of (1)–(4) (denoted as BFGS and DFP) and the QNM including iterations with additional orthogonalization (116)–(119) in the form of a sequence of iterations (120) (denoted as BFGS_V and DFP_V). The experiments were carried out by varying the coefficients of the initial normalization of the matrices of the QNM metric.

Since the use of quasi-Newtonian methods is justified primarily based on functions with a high degree of conditionality where conjugate gradient methods do not work efficiently, the test functions were selected based on this principle. Since the QNM is based on a quadratic model of a function, its local convergence rate in a certain neighborhood of the current minimum is largely determined by the efficiency of minimizing the ill-conditioned quadratic functions. The test functions are as follows:

(1) $f_1(x) = \sum_{i=1}^{n} x_i^2 i^6, x_0 = (10/1, 10/2, \ldots, 10/n)$.

The optimal value and minimum point are $f_1^* = 0\ and\ x^* = (0,0,\ldots,0)$. The condition number of the matrix of second derivatives for some n is $cond\left(\nabla^2 f_1(x)\right) = \lambda_{max}/\lambda_{min} = n^6$. When n=1000, the condition number will be $cond\left(\nabla^2 f_1(x)\right) = 1000^6 = 10^{18}$.

(2) $f_2(x) = \sum_{i=1}^{n} x_i^2 \left(\frac{n}{i}\right)^6, x_0 = (10,10,\ldots,10)$.

The optimal value and minimum point are $f_2^* = 0\ and\ x^* = (0,0,\ldots,0)$. The condition number of the matrix of second derivatives for some n is $cond\left(\nabla^2 f_2(x)\right) = \lambda_{max}/\lambda_{min} = n^6$. When n = 1000, the condition number will be $cond\left(\nabla^2 f_2(x)\right) = 1000^6 = 10^{18}$.

(3) $f_3(x) = (\sum_{i=1}^{n} x_i^2 i)^r, x_0 = (1,1,\ldots,1), r = 2$.

The optimal value and minimum point are $f_3^* = 0\ and\ x^* = (0,0,\ldots,0)$. The function f_3 is based on a quadratic function with the condition number of the matrix of second derivatives for some $n\ cond\left(\nabla^2 f_3(x)\right) = \lambda_{max}/\lambda_{min} = n$. When n = 1000, the condition number will be $cond\left(\nabla^2 f_3(x)\right) = 1000$. The topology of the level surfaces of the function f_3 is identical to the topology of the level surfaces of the basic quadratic function. The matrix of second derivatives of a function tends to zero as it approaches the minimum. Consequently, the inverse matrix tends to infinity. The approximation pattern for the matrix of second derivatives in the QNM will correspond to K = 1 in Figure 2. This case makes it difficult to enter a new subspace due to the significant predominance of eigenvalues in the metric matrix in the already surveyed part of the subspace compared to the eigenvalues of the metric matrix in the unsurveyed area.

(4) $f_4(x) = \sum_{i=1}^{n/2} [10^8 \cdot (x_{2i-1}^2 - x_{2i})^2 + (x_{2i-1} - 1)^2], x_0 = (1.2,1,-1.2,1,\ldots,-1.2,1)$.

The optimal value and minimum point of rescaled multidimensional Rosenbrock function [42] are $f_4^* = 0\ and\ x^* = (1,1,\ldots,1)$. This function has a curved ravine with small values of the second derivative in the direction of the bottom of the ravine and large values of the second derivative in the direction of the normal to the bottom of the ravine. The ratio of second derivatives along such directions is approximately 10^8.

The stopping criterion is

$$f(x^k) - f^* \le \varepsilon = 10^{-10}.$$

The results of minimizing the presented functions are given in Tables 1 and 2 for n = 1000. The problem was considered solved if the method, within the allotted number of iterations and calculations of the function and gradient, reached a function value that satisfied the stopping criterion. The cell indicates the number of iterations (one-dimensional searches along a direction), and below is the number of calls to the function procedure, where the function and gradient are calculated simultaneously. The number

of iterations in all tests were limited to 40,000. If the costs of the method exceeded the specified number of iterations, the method was stopped. It was believed that no solution had been found by this method. The dash sign indicates options where a solution could not be obtained. In cases where there was no solution, looping of methods occurred due to the smallness of the minimization steps and, as a consequence, large errors in the gradient differences used in the transformation operations of metric matrices.

Let us consider the effects of reducing the convergence rate of the method. For example, for the function f_3, the matrix of second derivatives tends to zero as it approaches the minimum. Consequently, the inverse matrix tends to infinity. The approximation pattern for the matrix of second derivatives in the QNM will correspond to K = 1 in Figure 2. In the explored part of the subspace, the matrix of the QNM grows. Therefore, the slight presence of residuals in this part of the subspace is greatly amplified. In the unexplored part of the space, the eigenvalues are fixed. This case makes it difficult to enter a new subspace due to the significant predominance of eigenvalues in the metric matrix in the explored part of the subspace compared to the eigenvalues of the metric matrix in the unexplored area. In order to enter the unexplored part of the subspace, it is necessary to eliminate the discrepancy in the explored part of the space. As a consequence, when minimizing functions with a high degree of conditionality, the search steps become smaller, the errors in the gradient differences increase, and the minimization method becomes loopy.

Table 1. Results of minimization with normalization of matrix (124) at $K = 1$ and $n = 1000$.

	Exact Descent				**Inexact Descent**			
	BFGS	**BFGS_V**	**DFP**	**DFP_V**	**BFGS**	**BFGS_V**	**DFP**	**DFP_V**
$f_1(x)$	1157	1157	1228	1211	1854	1648	-	1762
	2526	2523	2712	2667	3980	3413		3750
$f_2(x)$	2400	2370	-	-	4351	3218	-	-
	5663	5560			9908	7242		
$f_3(x)$	1404	1396	-	1643	1905	1508	5837	2497
	3206	3190		3743	4286	3394	13,362	5686
$f_4(x)$	3328	2964	-	-	-	-	-	-
	7455	6668						

For exact descent, there are practically no differences between the BFGS and BFGS_V methods. In exact descent, successive descent vectors for quadratic functions are conjugated, and matrix learning, considered in a coordinate system with an identity matrix of second derivatives, is carried out using an orthogonal system of vectors. Minor errors lead to the fact that this orthogonality is violated, which affects the DFP method.

For inexact descent, the BFGS_V method significantly outperforms the BFGS method. The DFP and DFP_V methods are practically ineffective on these tests, although the DFP_V method shows better results.

Thus, with one-dimensional search errors, the BFGS_V algorithm is significantly more effective than the BFGS method. The DFP method is practically not applicable when the problem is highly conditioned.

Table 2 shows the experimental data with normalization of the matrix (124) at $K > 1$. For the functions $f_3(x)$ and $f_4(x)$, the coefficient K had to be reduced to obtain a more effective result.

Table 2. Results of minimization with normalization of matrix (124) at $K = 10{,}000$ and $n = 1000$. For results marked with an asterisk, $K = 100$.

	Exact Descent				**Inexact Descent**			
	BFGS	**BFGS_V**	**DFP**	**DFP_V**	**BFGS**	**BFGS_V**	**DFP**	**DFP_V**

$f_1(x)$	1038 2194	1038 2193	1041 2197	1041 2195	1221 2190	1189 2116	1307 2431	1260 2343
$f_2(x)$	791 1863	795 1874	1091 2560	852 2028	1386 3129	1012 2159	2524 5794	1509 3341
$f_3(x)$	1082 * 2436	1090 * 2454	8977 * 20,201	1343 * 3055	1281 2802	1129 2453	4281 9742	1845 4183
$f_4(x)$	4062 * 9135	3850* 8686	-	-	-	-	-	-

The initial normalization of the metric matrices, as follows from the results of Tables 1 and 2, significantly improves the convergence of QNMs. The situation corresponds to the case in Figure 2 for K > 1. Large eigenvalues in the unexplored part of the subspace make it easy to find new conjugate directions and efficiently train metric matrices with almost orthogonal training vectors.

For exact descent, there are practically no differences between the BFGS and BFGS_V methods. For inexact descent, the BFGS_V method significantly outperforms the BFGS method. The DFP and DFP_V methods are efficient for functions $f_1(x) - f_3(x)$, while for inexact descent, the DFP_V method significantly outperforms the DFP method.

Thus, in the case of one-dimensional search errors, the BFGS_V algorithm is significantly more efficient than the BFGS method and correct initial normalization of metric matrices can significantly increase the convergence rate of the method.

For the purpose of giving a visual demonstration of the method, we minimize a two-dimensional function as follows:

$$f_5(x) = (x_1^2 + 100x_2^2)^2, x_0 = (1,1).$$

To test the idea of the efficiency of orthogonalization to increase the performance of the quasi-Newton method, to adversely affect the minimization conditions, the initial matrix was normalized at K = 0.000001, which should significantly complicate the solution of the problem and reveal the effect of the advantages of the degree of orthogonality of the learning vectors of the BFGS and BFGS_V methods over the DFP method.

The stopping criterion was

$$f(x^k) - f^* \le \varepsilon = 10^{-2}.$$

The results are shown in Table 3. The row with $f_5(x)$ shows the number of iterations, while the row with f_{min} shows the minimal function value achieved.

Table 3. Results of minimization with normalization of matrix (124) at $K = 0.000001$ and $n = 2$.

	Exact Descent		
	BFGS	**BFGS_V**	**DFP_V**
Number of iterations	13	5	3733
F_{min}	2.5862×10^{-3}	7.5003×10^{-4}	7.7552×10^{-3}

The path of three considered algorithms is shown in Figure 3.

Here, theoretical results of the influence of the orthogonality degree of matrix learning vectors on the convergence rate of the method are confirmed. The BFGS_V method performs forced orthogonalization, which improves the result of the BFGS method. The trajectories of the methods are listed in Tables A1–A3 of Appendix A (the trajectory of the DFP method is shown partially).

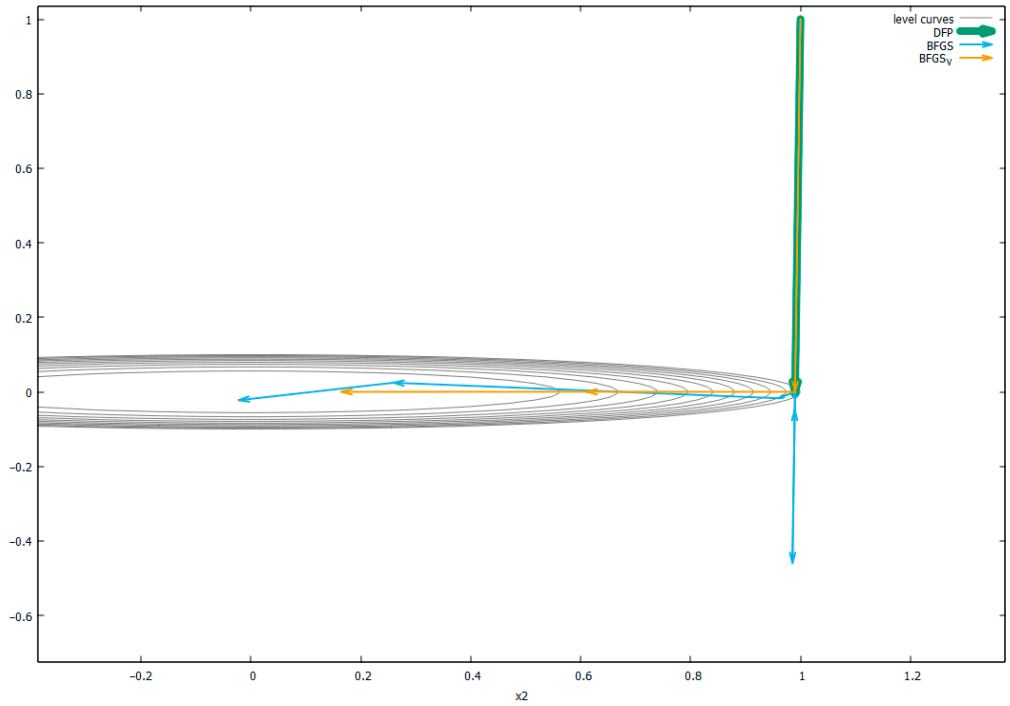

Figure 3. Level curves and paths of the optimization algorithms for function f_5.

9. Conclusions

This paper presents methods for converting metric matrices in quasi-Newton methods based on gradient learning algorithms. As a result, it is possible to represent the system of learning steps in the form of an algorithm for minimizing a certain objective function along a system of directions and to draw conclusions about the convergence rate of the learning process based on the properties of this system of directions. The main conclusion is that the convergence rate is directly dependent on the degree of orthogonality of the learning vectors.

Based on the study of learning algorithms in the DFP and BFGS methods, it is possible to show that the degree of orthogonality of the learning vectors in the BFGS method is higher than that in the DFP method. This means that entering the unexplored region of the minimization space due to the noise and inaccuracies of one-dimensional descent in the DFP method is more difficult than in the BFGS method, which explains why the BFGS updating formula has the best results.

As a result of studies on quadratic functions, it has been revealed that the dimension of the minimization space is reduced when iterations with an exact one-dimensional descent or iterations with additional orthogonalization are included in the quasi-Newton method. It is shown that it is also possible to increase the orthogonality of the learning vectors and thereby increase the convergence rate of the method through special normalization of the initial metric matrix. The theoretically predicted effects of increasing the efficiency of quasi-Newton methods were confirmed as a result of a computational experiment on complex ill-conditioned minimization problems. In future work, we plan to study minimization methods under the conditions of a linear background that adversely affects the convergence.

Author Contributions: Conceptualization, V.K. and E.T.; methodology, V.K., E.T. and P.S.; software, V.K.; validation, L.K., E.T., P.S. and D.K.; formal analysis, P.S., E.T. and D.K.; investigation, E.T.; resources, L.K.; data curation, P.S. and D.K.; writing—original draft preparation, V.K.; writing—review and editing, E.T., P.S. and L.K.; visualization, V.K.; supervision, V.K. and L.K.; project administration, L.K. and D.K.; funding acquisition, D.K. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the Ministry of Science and Higher Education of the Russian Federation (Grant No. 075-15-2022-1121). Predrag Stanimirović is supported by the Science Fund of the Republic of Serbia (No. 7750185, Quantitative Automata Models: Fundamental Problems and Applications—QUAM).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Data are contained within the article.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Appendix A

Table A1. Trajectory of the BFGS_V method moving.

Iteration	$f_5(x)$	x_1	x_2
0	0.0	1.0	1.0
1	9.605985×10^{-1}	9.900005×10^{-1}	5.073008×10^{-5}
2	9.605986×10^{-1}	9.900005×10^{-1}	5.332304×10^{-5}
3	9.605988×10^{-1}	9.900006×10^{-1}	5.659179×10^{-5}
4	1.396454×10^{-1}	6.112994×10^{-1}	2.122674×10^{-4}
5	7.500359×10^{-4}	1.654885×10^{-1}	5.745936×10^{-5}

Table A2. Trajectory of the BFGS method moving.

Iteration	$f_5(x)$	x_1	x_2
0	0.0	1.0	1.0
1	4.865982×10^{2}	9.854078×10^{-1}	-4.592161×10^{-1}
2	1.489919	9.895086×10^{-1}	-4.914216×10^{-2}
3	9.608406×10^{-1}	9.899878×10^{-1}	-1.220497×10^{-3}
4	9.605957×10^{-1}	9.899999×10^{-1}	-6.968405×10^{-6}
5	9.605941×10^{-1}	9.899990×10^{-1}	-1.031223×10^{-4}
6	9.605941×10^{-1}	9.899990×10^{-1}	-9.891454×10^{-5}
7	9.605941×10^{-1}	9.899990×10^{-1}	-9.945802×10^{-5}
8	9.612290×10^{-1}	9.899966×10^{-1}	1.815384×10^{-3}
9	9.641387×10^{-1}	9.899999×10^{-1}	-4.249478×10^{-3}
10	9.605818×10^{-1}	9.899963×10^{-1}	9.036299×10^{-6}
11	9.059276×10^{-1}	9.603298×10^{-1}	-1.719565×10^{-2}
12	1.634266×10^{-2}	2.596599×10^{-1}	2.457950×10^{-2}
13	2.586155×10^{-3}	-2.187391×10^{-2}	-2.244455×10^{-2}

Table A3. Trajectory of the DFP method moving.

Iteration	$f_5(x)$	x_1	x_2
0	0.0	1.0	1.0
1	9.605985×10^{-1}	9.900005×10^{-1}	5.073008×10^{-5}
2	9.605986×10^{-1}	9.900005×10^{-1}	5.332304×10^{-5}
3	9.605988×10^{-1}	9.900006×10^{-1}	5.659179×10^{-5}
4	9.605991×10^{-1}	9.900006×10^{-1}	6.073300×10^{-5}
5	9.605994×10^{-1}	9.900007×10^{-1}	6.601199×10^{-5}
6	9.605942×10^{-1}	9.899988×10^{-1}	-1.191852×10^{-4}
7	9.605942×10^{-1}	9.899988×10^{-1}	-1.202172×10^{-4}

8	9.605942×10^{-1}	9.899988×10^{-1}	-1.215724×10^{-4}
9	9.605942×10^{-1}	9.899988×10^{-1}	-1.233791×10^{-4}
10	9.605942×10^{-1}	9.899987×10^{-1}	-1.258332×10^{-4}
11	9.605957×10^{-1}	9.899999×10^{-1}	-8.005638×10^{-6}
12	9.605974×10^{-1}	9.899977×10^{-1}	-2.294417×10^{-4}
13	9.605962×10^{-1}	9.900000×10^{-1}	4.809401×10^{-6}
14	9.605946×10^{-1}	9.899985×10^{-1}	-1.494894×10^{-4}
15	9.605941×10^{-1}	9.899992×10^{-1}	-8.356599×10^{-5}
16	9.605941×10^{-1}	9.899990×10^{-1}	-1.019703×10^{-4}
17	9.605941×10^{-1}	9.899990×10^{-1}	-9.866443×10^{-5}
18	9.605941×10^{-1}	9.899990×10^{-1}	-9.826577×10^{-5}
19	9.605941×10^{-1}	9.899990×10^{-1}	-9.914318×10^{-5}
20	9.605941×10^{-1}	9.899990×10^{-1}	-9.806531×10^{-5}
21	9.639558×10^{-1}	9.899569×10^{-1}	-4.240006×10^{-3}
22	9.605941×10^{-1}	9.899991×10^{-1}	-9.292366×10^{-5}
23	9.605942×10^{-1}	9.899988×10^{-1}	-1.237791×10^{-4}
24	9.605943×10^{-1}	9.899994×10^{-1}	-6.417724×10^{-5}
25	9.605943×10^{-1}	9.899986×10^{-1}	-1.365365×10^{-4}
26	9.605942×10^{-1}	9.899992×10^{-1}	-7.609652×10^{-5}
27	9.605941×10^{-1}	9.899989×10^{-1}	-1.126456×10^{-4}
28	9.605941×10^{-1}	9.899990×10^{-1}	-9.613542×10^{-5}
29	9.605941×10^{-1}	9.899990×10^{-1}	-1.018252×10^{-4}
30	9.605941×10^{-1}	9.899990×10^{-1}	-1.002864×10^{-4}
31	9.605941×10^{-1}	9.899990×10^{-1}	-1.007205×10^{-4}
32	9.605941×10^{-1}	9.899990×10^{-1}	-1.001624×10^{-4}
33	9.605941×10^{-1}	9.899990×10^{-1}	-1.006906×10^{-4}
34	9.605941×10^{-1}	9.899990×10^{-1}	-1.000533×10^{-4}
35	9.605941×10^{-1}	9.899990×10^{-1}	-1.006925×10^{-4}
36	9.605941×10^{-1}	9.899990×10^{-1}	-9.993761×10^{-5}
37	9.605941×10^{-1}	9.899990×10^{-1}	-1.007145×10^{-4}
38	9.605941×10^{-1}	9.899990×10^{-1}	-9.980530×10^{-5}
39	9.605941×10^{-1}	9.899990×10^{-1}	-1.007537×10^{-4}
40	9.605941×10^{-1}	9.899990×10^{-1}	-9.964890×10^{-5}
41	9.605941×10^{-1}	9.899990×10^{-1}	-1.008103×10^{-4}

References

1. Polyak, B.T. *Introduction to Optimization*; Translated from Russian; Optimization Software Inc., Publ. Division: New York, NY, USA, 1987.
2. Nocedal, J.; Wright, S. *Numerical Optimization, Series in Operations Research and Financial Engineering*; Springer: New York, NY, USA, 2006.
3. Bertsekas, D.P. *Constrained Optimization and Lagrange Multiplier Methods*; Academic Press: New York, NY, USA, 1982.
4. Gill, P.E.; Murray, W.; Wright, M.H. *Practical Optimization*; SIAM: Philadelphia, PE, USA, 2020.
5. Dennis, J.E.; Schnabel, R.B. *Numerical Methods for Unconstrained Optimization and Nonlinear Equations*; SIAM: Philadelphia, PE, USA, 1996.
6. Evtushenko, Y.G. *Methods for Solving Extremal Problems and Their Application in Optimization Systems*; Nauka: Moscow, Russia, 1982. (In Russian)
7. Polak, E. *Computational Methods in Optimization: A Unified Approach*; Academic Press: New York, NY, USA, 1971.
8. Kokurin, M.M.; Kokurin, M.Y.; Semenova, A.V. Iteratively regularized Gauss–Newton type methods for approximating quasi–solutions of irregular nonlinear operator equations in Hilbert space with an application to COVID-19 epidemic dynamics. *Appl. Math. Comput.* **2022**, *431*, 127312.
9. Zhang, J.; Tao, X.; Sun, P.; Zheng, Z. A positional misalignment correction method for Fourier ptychographic microscopy based on the quasi-Newton method with a global optimization module. *Opt. Commun.* **2019**, *452*, 296–305.

10. Lampron, O.; Therriault, D.; Lévesque, M. An efficient and robust monolithic approach to phase-field quasi-static brittle fracture using a modified Newton method. *Comput. Methods Appl.* **2021**, *386*, 114091.
11. Spenke, T.; Hosters, N.; Behr, M. A multi-vector interface quasi-Newton method with linear complexity for partitioned fluid–structure interaction. *Comput. Methods Appl. Mech. Engrg.* **2020**, *361*, 112810.
12. Zorrilla, R.; Rossi, R. A memory-efficient MultiVector Quasi-Newton method for black-box Fluid-Structure Interaction coupling. *Comput. Struct.* **2023**, *275*, 106934.
13. Davis, K.; Schulte, M.; Uekermann, B. Enhancing Quasi-Newton Acceleration for Fluid-Structure Interaction. *Math. Comput. Appl.* **2022**, *27*, 40. https://doi.org/10.3390/mca27030040.
14. Tourn, B.; Hostos, J.; Fachinotti, V. Extending the inverse sequential quasi-Newton method for on-line monitoring and controlling of process conditions in the solidification of alloys. *Int. Commun. Heat Mass Transf.* **2023**, *142*, 1106647.
15. Hong, D.; Li, G.; Wei, L.; Li, D.; Li, P.; Yi, Z. A self-scaling sequential quasi-Newton method for estimating the heat transfer coefficient distribution in the air jet impingement. *Int. J. Therm. Sci*. **2023**, *185*, 108059.
16. Berahas, A.S.; Jahani, M.; Richtárik, P.; Takác, M. Quasi-Newton Methods for Machine Learning: Forget the Past, Just Sample. *Optim. Methods Softw*. **2022**, *37*, 1668-1704. https://doi.org/10.1080/10556788.2021.1977806.
17. Rafati, J. Quasi-Newton Optimization Methods For Deep Learning Applications, 2019. Available online: https://arxiv.org/abs/1909.01994.pdf (accessed on 11 January 2024).
18. Indrapriyadarsini, S.; Mahboubi, S.; Ninomiya, H.; Kamio, T.; Asai, H. Accelerating Symmetric Rank-1 Quasi-Newton Method with Nesterov's Gradient for Training Neural Networks. *Algorithms* **2022**, *15*, 6. https://doi.org/10.3390/a15010006.
19. Davidon, W.C. *Variable Metric Methods for Minimization*; A.E.C. Res. and Develop. Report ANL–5990; Argonne National Laboratory: Argonne, IL, USA, 1959.
20. Fletcher, R.; Powell, M.J.D. A rapidly convergent descent method for minimization. *Comput. J*. **1963**, *6*, 163–168.
21. Oren, S.S. Self-scaling variable metric (SSVM) algorithms I: Criteria and sufficient conditions for scaling a class of algorithms. *Manag. Sci*. **1974**, *20*, 845–862.
22. Oren, S.S. Self-scaling variable metric (SSVM) algorithms II: Implementation and experiments. *Manag. Sci.* **1974**, *20*, 863–874.
23. Powell, M.J.D. Convergence Properties of a Class of Minimization Algorithms. In *Nonlinear Programming*; Mangasarian, O.L., Meyer, R.R., Robinson, S.M., Eds.; Academic Press: New York, NY, USA, 1975; Volume 2, pp. 1–27.
24. Dixon, L.C. Quasi-Newton algorithms generate identical points. *Math. Program.* **1972**, *2*, 383–387.
25. Huynh, D.Q.; Hwang, F.-N. An accelerated structured quasi-Newton method with a diagonal second-order Hessian approximation for nonlinear least squares problems. *J. Comp. Appl. Math*. **2024**, *442*, 115718. https://doi.org/10.1016/j.cam.2023.115718.
26. Chai, W.H.; Ho, S.S.; Quek, H.C. A Novel Quasi-Newton Method for Composite Convex Minimization. *Pattern Recognit.* **2022**, *122*, 108281. https://doi.org/10.1016/j.patcog.2021.108281.
27. Fang, X.; Ni, Q.; Zeng, M. A modified quasi-Newton method for nonlinear equations. *J. Comp. Appl. Math*. **2018**, *328*, 44–58. https://doi.org/10.1016/j.cam.2017.06.024.
28. Zhou, W.; Zhang, L. A modified Broyden-like quasi-Newton method for nonlinear equations. *J. Comp. Appl. Math*. **2020**, *372*, 112744. https://doi.org/10.1016/j.cam.2020.112744.
29. Broyden, C.G. The convergence of a class of double–rank minimization algorithms. *J. Inst. Math. Appl*. **1970**, *6*, 76–79.
30. Fletcher, R. A new approach to variable metric algorithms. *Comput. J*. **1970**, *13*, 317–322.
31. Goldfarb, D. A family of variable metric methods derived by variational means. *Math. Comput*. **1970**, *24*, 23–26.
32. Liu, D.C.; Nocedal, J. On the limited memory BFGS method for large scale optimization. *Math. Program.* **1989**, *45*, 503–528.
33. Zhu, C.; Byrd, R.H.; Lu, P.; Nocedal, J. L-BFGS-B: Algorithm 778: L-BFGS-B, FORTRAN routines for large scale bound constrained optimization. *ACM Trans. Math. Softw.* **1997**, *23*, 550–560.
34. Tovbis, E.; Krutikov, V.; Stanimirović, P.; Meshechkin, V.; Popov, A.; Kazakovtsev, L. A Family of Multi-Step Subgradient Minimization Methods. *Mathematics* **2023**, *11*, 2264. https://doi.org/10.3390/math11102264.
35. Krutikov, V.; Gutova, S.; Tovbis, E.; Kazakovtsev, L.; Semenkin, E. Relaxation Subgradient Algorithms with Machine Learning Procedures. *Mathematics* **2022**, *10*, 3959. https://doi.org/10.3390/math10213959.
36. Feldbaum, A.A. On a class of dual control learning systems. *Avtomat. i Telemekh*. **1964**, *25*, 433–444. (In Russian)
37. Aizerman, M.A.; Braverman, E.M.; Rozonoer, L.I. *Method of Potential Functions in Machine Learning Theory*; Nauka: Moscow, Russia, 1970. (In Russian)
38. Tsypkin, Y.Z. *Foundations of the Theory of Learning Systems*; Academic Press: New York, NY, USA, 1973.
39. Kaczmarz, S. Approximate solution of systems of linear equations. *Internat. J. Control*. **1993**, *54*, 1239–1241.
40. Krutikov, V.N. On the convergence rate of minimization methods along vectors of a linearly independent system. *USSR Comput. Math. Math. Phys*. **1983**, *23*, 218–220.
41. Rao, S.S. *Engineering Optimization*; Wiley: Hoboken, NJ, USA, 2009.
42. Andrei, N. An Unconstrained Optimization Test Functions Collection. Available online: http://www.ici.ro/camo/journal/vol10/v10a10.pdf (accessed on 1 April 2024).

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.