Found Comput Math (2017) 17:359–422
DOI 10.1007/s10208-015-9290-8

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

# Steepest Descent Method with Random Step Lengths

**Zdeněk Kalousek**[1]

Received: 23 December 2013 / Revised: 26 March 2015 / Accepted: 20 August 2015 /
Published online: 23 November 2015

**Abstract** The paper studies the steepest descent method applied to the minimization of a twice continuously differentiable function. Under certain conditions, the random choice of the step length parameter, independent of the actual iteration, generates a process that is almost surely $R$-convergent for quadratic functions. The convergence properties of this random procedure are characterized based on the mean value function related to the distribution of the step length parameter. The distribution of the random step length, which guarantees the maximum asymptotic convergence rate independent of the detailed properties of the Hessian matrix of the minimized function, is found, and its uniqueness is proved. The asymptotic convergence rate of this optimally created random procedure is equal to the convergence rate of the Chebyshev polynomials method. Under practical conditions, the efficiency of the suggested random steepest descent method is degraded by numeric noise, particularly for ill-conditioned problems; furthermore, the asymptotic convergence rate is not achieved due to the finiteness of the realized calculations. The suggested random procedure is also applied to the minimization of a general non-quadratic function. An algorithm needed to estimate relevant bounds for the Hessian matrix spectrum is created. In certain cases, the random procedure may surpass the conjugate gradient method. Interesting results are achieved when minimizing functions having a large number of local minima. Preliminary results of numerical experiments show that some modifications of the presented basic method may significantly improve its properties.

Communicated by Michael Overton.

✉ Zdeněk Kalousek
zdenek.kalousek@tul.cz

[1] Department of Applied Mathematics, Faculty of Education, Technical University of Liberec, Liberec, Czech Republic

**Keywords** Optimization · Gradient methods · Stochastic process · Algorithms · Convergence · Computational experiments

**Mathematics Subject Classification** 49M30 · 65C05 · 65F10 · 90C26

## 1 Introduction

We consider a function $V(\mathsf{x}) : \mathcal{R}^M \to \mathcal{R}$ bounded from bellow and twice continuously differentiable on $\mathcal{R}^M$. The problem to solve is to find its minimum.

The steepest descent method, proposed by Cauchy [3], is a natural procedure creating a sequence of iterates $\{\mathsf{x}_j\}_{j\geq 0}$ using the formula

$$\mathsf{x}_j = \mathsf{x}_{j-1} - \gamma_j \mathsf{g}_{j-1} \quad \text{for } j \geq 1\,, \tag{1.1}$$

where $\mathsf{g}_j = \nabla V(\mathsf{x}_j)$ and $\gamma_j > 0$ is the step length parameter; the point $\mathsf{x}_0$ is selected arbitrarily. Denote $\mathsf{H}(\mathsf{x})$ the Hessian matrix of the function $V$ in the point $\mathsf{x} \in \mathcal{R}^M$. The Taylor's theorem implies

$$\begin{aligned} V(\mathsf{x}_j) &= V(\mathsf{x}_{j-1}) + \langle \mathsf{g}_{j-1} | \mathsf{x}_j - \mathsf{x}_{j-1} \rangle + \frac{1}{2} \langle \mathsf{x}_j - \mathsf{x}_{j-1} | \mathsf{H}(\xi_j)[\mathsf{x}_j - \mathsf{x}_{j-1}] \rangle \\ &= V(\mathsf{x}_{j-1}) - \gamma_j \|\mathsf{g}_{j-1}\|^2 + \frac{1}{2}\gamma_j^2 \langle \mathsf{g}_{j-1} | \mathsf{H}(\xi_j) \mathsf{g}_{j-1} \rangle \end{aligned} \tag{1.2}$$

for a $\xi_j$ between $\mathsf{x}_{j-1}$ and $\mathsf{x}_j$. The symbol $\langle\cdot|\cdot\rangle$ denotes the standard scalar product on $\mathcal{R}^M$ here. If $\gamma_j$ are sufficiently small, the relation (1.2) guarantees $V(\mathsf{x}_j) \leq V(\mathsf{x}_{j-1})$ for any $j \in \mathcal{N}$; therefore, the sequence $\{V(\mathsf{x}_j)\}_{j\geq 0}$ is convergent.

This result does not guarantee either the convergence of the sequence $\{\mathsf{x}_j\}_{j\geq 0}$ or the convergence of the sequence $\{V(\mathsf{x}_j)\}_{j\geq 0}$ to the minimal value of the function $V$. Yet, the presented idea is well applicable to many situations. The most elementary case occurs when the Hessian matrix $\mathsf{H}(\mathsf{x})$ has all its eigenvalues $\lambda_i \geq \lambda_0$ for any $\mathsf{x} \in \mathcal{R}^M$ and a fixed $\lambda_0 > 0$. Then, the function $V(\mathsf{x})$ has just one local minimum on $\mathcal{R}^M$ in a point $\underline{\mathsf{x}}$ and $\nabla V(\mathsf{x}) \neq \mathsf{0}$ for any $\mathsf{x} \neq \underline{\mathsf{x}}$; the parameters $\gamma_j$ may be selected so that the sequence $\{V(\mathsf{x}_j)\}_{j\geq 0}$ is decreasing, until the condition $\mathsf{x}_j = \underline{\mathsf{x}}$ is satisfied; moreover, the choice

$$\gamma_j = \arg\min_{\gamma\in\mathcal{R}} \big(V(\mathsf{x}_{j-1} - \gamma \mathsf{g}_{j-1})\big) \tag{1.3}$$

creates a sequence $\{\mathsf{x}_j\}_{j\geq 0}$ which converges to the point $\underline{\mathsf{x}}$.

The problem is further simplified, when the function $V(\mathsf{x})$ is quadratic; then, its Hessian matrix $\mathsf{H}(\mathsf{x}) = \mathsf{A}$ for any $\mathsf{x} \in \mathcal{R}^M$ and a positive symmetric matrix $\mathsf{A}$. Denote $\mathsf{b} = \mathsf{A}\underline{\mathsf{x}}$. In this case, the function

$$V(\mathsf{x}) = \frac{1}{2}\langle \mathsf{x} | \mathsf{A}\mathsf{x} \rangle - \langle \mathsf{b} | \mathsf{x} \rangle + V_0, \tag{1.4}$$

and its minimization is equivalent to the solution of the equation

$$\mathsf{A}\mathsf{x} = \mathsf{b}\,. \tag{1.5}$$

Since the function $V(\mathsf{x})$ may be approximated by an appropriate function (1.4) in a neighborhood of any isolated local minimum of $V(\mathsf{x})$,[1], the convergence properties of the created iterative process minimizing the function (1.4) are important in general. So, we will first consider the problem (1.5).

Denote $\mathsf{r}_j = \mathsf{b} - \mathsf{A}\mathsf{x}_j = -\nabla V(\mathsf{x}_j) = -\mathsf{g}_j$ the residual corresponding to the iteration $\mathsf{x}_j$. Given $\mathsf{x}_{j-1}$, the relation (1.3) leads to the locally optimal step length parameter (also called the *exact step length parameter*)

$$\hat{\gamma}_j = \frac{\|\mathsf{r}_{j-1}\|^2}{\langle \mathsf{r}_{j-1}|\mathsf{A}\mathsf{r}_{j-1}\rangle}. \tag{1.6}$$

As mentioned above, the steepest descent method using this locally optimal parameter in all the iteration steps converges to the solution to the Eq. (1.5). However, the convergence properties are very poor; an explanation of this fact was given by Akaike [1].

If we want to minimize the function $V(\mathsf{x}_{j+s})$ at given $\mathsf{x}_j$ and $s \in \mathcal{N}$, we obtain the *optimum s-gradient methods*; the detailed analysis of the asymptotic behavior of these methods can be found in [7]. The rate of convergence is significantly improved, increasing the value $s$ (the computation time decreases approximately $s$-times compared to the method using the exact step lengths); however, in practice, round-off errors do not allow the use of large values $s$, since in these situations it is necessary to work with extremely ill-conditioned matrices.[2]

A sophisticated algorithm based on the combination of the two-gradient optimal method and the use of the exact step length is presented in [12]. Its efficiency seems to be comparable with the efficiency of the optimal nine-gradient algorithm. Nevertheless, this result is still unsatisfactory for ill-conditioned problems.

The following text shows one possible way to achieve a significantly better convergence behavior of the method by using a selection of the step length parameter, giving up any direct form of optimization.

Denote different eigenvalues of the matrix $\mathsf{A}$ as $\lambda_1 < \lambda_2 < \cdots < \lambda_N$, $N \le M$ (the multiplicity of the eigenvalues is admitted). Furthermore, denote $\lambda = \lambda_1$, $\Lambda = \lambda_N$, and $\kappa(\mathsf{A}) = \frac{\Lambda}{\lambda}$ the condition number of the matrix $\mathsf{A}$.

Given the iteration $\mathsf{x}_{j-1}$; this point can be unequivocally characterized by the corresponding residual $\mathsf{r}_{j-1}$. By employing one step of the steepest descent method, we get

$$\mathsf{r}_j = \mathsf{b} - \mathsf{A}(\mathsf{x}_{j-1} + \gamma_j \mathsf{r}_{j-1}) = \left(\mathsf{I} - \gamma_j \mathsf{A}\right) \mathsf{r}_{j-1}\ , \tag{1.7}$$

and analogously

$$\mathsf{r}_{j+n} = \left(\mathsf{I} - \gamma_{j+n}\mathsf{A}\right)\left(\mathsf{I} - \gamma_{j+n-1}\mathsf{A}\right)\ldots\left(\mathsf{I} - \gamma_{j+1}\mathsf{A}\right)\mathsf{r}_j\ . \tag{1.8}$$

[1] except for the cases when the Hessian matrix in the local minimum of $V(\mathsf{x})$ is singular

[2] The indirect way to minimize $V(\mathsf{x}_{j+s})$ using $s$ steps of the conjugate gradient method is not considered now because it differs from the steepest descent method (1.1).

Setting $\mathsf{x}_0 = \mathsf{0}$, the residual $\mathsf{r}_0 = \mathsf{b}$. All the matrices in the product on the right-hand side of (1.8) commute; so, using inverse step length parameters $l_i = \frac{1}{\gamma_i}$, we can write

$$\mathsf{r}_n = \prod_{i=j+1}^{n} \left( \mathsf{I} - \frac{\mathsf{A}}{l_i} \right) \cdot \mathsf{r}_j = \prod_{i=1}^{n} \left( \mathsf{I} - \frac{\mathsf{A}}{l_i} \right) \cdot \mathsf{b} \tag{1.9}$$

independently of the order of the terms in the product on the right-hand side.

Let $\mathcal{S}(\mathsf{A}, \mathsf{b})$ be the set of all the eigenvalues of the matrix $\mathsf{A}$, for which the orthogonal projections $\mathsf{b}_j$ of the vector $\mathsf{b}$ onto the corresponding eigenspaces with the eigenvalue $\lambda_j$ are nonzero. When the set $\{l_i\}_{i=1}^{n} = \mathcal{S}(\mathsf{A}, \mathsf{b})$, the residual

$$\mathsf{r}_n = \prod_{i=1}^{n} \left( \mathsf{I} - \frac{\mathsf{A}}{l_i} \right) \cdot \mathsf{b} = \sum_{j=1}^{N} \prod_{i=1}^{n} \left( 1 - \frac{\lambda_j}{l_i} \right) \cdot \mathsf{b}_j = \mathsf{0} \tag{1.10}$$

, and the corresponding iterative $\mathsf{x}_n$ solves the Eq. (1.5). The matrix polynomial in (1.10) can be generated by the conjugate gradient method (which is not the aim of this paper), or by using the eigenvalues of $\mathsf{A}$ directly. The last idea is difficult to realize, but alternative (although less efficient) methods based on (1.9) exist. The idea resides in an appropriate choice of the values $l_i \in \langle \lambda, \Lambda \rangle$ guaranteeing

$$\lim_{n \to \infty} \mathsf{r}_n = \mathsf{0} \, .$$

For instance, the Barzilai–Borwein algorithm [2] is very efficient. Its favorable behavior is not yet clearly explained; the proved $R$-convergence [5] guarantees the asymptotic logarithmic convergence rate having the order $[\kappa(\mathsf{A}) \ln(\kappa(\mathsf{A}))]^{-N}$, which is extremely small in comparison with the common logarithmic convergence rate of the order $[\kappa(\mathsf{A})]^{-\frac{1}{2}}$. Similarly, the proved $Q$-superlinear convergence [11] is also of little practical significance.[3] Another idea is a controlled, “sufficiently dense” filling of the interval $\langle \lambda, \Lambda \rangle$ with the values $l_i$. An efficient algorithm based on this idea is described in [13], and its properties are well clarified. Our suggestion with the same aim is to choose the values $l_i$ in (1.9) randomly from $\langle \lambda, \Lambda \rangle$, using an appropriate distribution of the values $l_i$.

An attempt of this kind has been realized in [15]. The uniform distribution of the step length parameter $\gamma_j \in \langle 0, 2\hat{\gamma}_j \rangle$, which was used, seems to be not very effective because the useless values of $\gamma_j < \frac{1}{\Lambda}$ corresponding to $l_j > \Lambda$ and $\gamma_j > \frac{1}{\lambda}$ related to $l_j < \lambda$ can appear. The achieved results still outperform the classic method, which uses the exact step length in all the steps; nevertheless, the convergence properties lag behind other standard methods significantly. The advantage of the used step length parameter distribution lies in the $Q$-convergence of the method.

[3] The possibility that the Barzilai–Borwein algorithm can converge linearly with the logarithmic convergence rate $\sim \frac{2}{\kappa(\mathsf{A})+1}$, when an inappropriate starting point is chosen, is mentioned already in the paper [2]. Nevertheless, this situation occurs with zero probability.

Random elements can also be found in the algorithm suggested in [10]. The obtained results are very satisfactory, but the randomness of the process is strongly restricted by the requirements of the quasi-local convergence. Moreover, the distributions of the random variables used therein are not specified, so we cannot judge their suitability.

A more consistent application of the random step length can be found in [9]. However, the reported results are rather poor (although considerably better than the results obtained via systematic work with the exact step length). This is a result of the used uniform distribution of the parameter $l_i \in \langle \lambda, \Lambda \rangle$.

The possible use of the stochastic step length parameters' choice is also mentioned in [13]; nevertheless, it is not developed any further. The results published therein show that the distribution of the inverse step lengths $l_i$, presented in Theorem 4 below, should create a sequence of iterations characterized by highly satisfactory convergence properties.

The distribution of the inverse step length $l_i$ has a crucial influence on the behavior of the entire process. In this context, there are three problems related to the distribution of the step length to solve:

1. Define the conditions in which the random process can give the solution to problem (1.5) with any prescribed accuracy,
2. find the distribution of the step length that assures the best convergence properties of the random process, without knowing the spectrum of matrix $\mathsf{A}$,
3. express the rate of convergence of the process with the optimally distributed step lengths and compare it with the convergence properties of other computational techniques.

These questions are answered in this paper. The basic algorithm and statements concerning the convergence properties of the created sequence of iterations for a general distribution of the inverse step length parameter are formulated in Sect. 2. The conditions necessary for the best convergence properties of the random process (without knowing the spectrum of matrix $\mathsf{A}$) are formulated in Sect. 3. The only distribution of the inverse step length parameter satisfying these conditions is calculated, and the asymptotic convergence rate of the process with this optimal step length parameter distribution is computed; the result corresponds to the one cited in [13], but the used computational method slightly differs.

Some results of the proposed method tests are presented in Sect. 4. The algorithm is modified to account for the fact that the values $\lambda$, $\Lambda$ are unknown in general. It is shown that the method gives satisfactory results only for well-conditioned problems; otherwise, round-off errors impede the convergence of the process.

Even in the case of well-conditioned problems, the achieved convergence rate does not surpass the convergence rate of the Chebyshev polynomials method and—consequently—the convergence rate of the conjugate gradient method. Nevertheless, the robustness of the random steepest descent procedure, namely its very weak dependence on the history of the iterative process, allows its use for the minimization of sufficiently smooth non-quadratic functions. The algorithm is modified to account for the variability of the Hessian matrix of the minimized function. The results are presented in Sect. 5.

The random character of the created optimization process may be advantageous when solving problems requiring randomized processing. A typical representative of such a problem is searching for the global minimum of a function with several local minima. Sect. 6 documents the capabilities of the suggested method. Finally, possible improvements of the basic method are proposed. Some proofs and calculations which could distract the readers from focusing on basic ideas of the paper are placed in appendices.

## 2 Convergence of the Algorithm with Totally Random Step Length for Strictly Convex Quadratic Functions

In the following part of the paper, we will deal with the solution of the problem (1.5) by using

**Algorithm 1** Let $\lambda$, $\Lambda$ be the minimal and maximum eigenvalues of the symmetric positive definite matrix $\mathsf{A}$, respectively; $L$ be a random variable with the values in $\langle\lambda, \Lambda\rangle$ and the distribution function $F_L(l)$. The system of Eq. (1.5) is solved as follows:

1. $\mathsf{x}_0 = 0$,
2. $l_j$ is a value of the random variable $L$,
3. $\mathsf{x}_j = \mathsf{x}_{j-1} + \frac{1}{l_j}\mathsf{r}_{j-1}$ for $j > 0$,
4. if $\|\mathsf{r}_j\| <$given $\varepsilon$, then stop
   else go to step 2 for the next $j$.

We will investigate the properties of the sequence of residuals $\{\mathsf{r}_j\}_{j\geq 0}$ created using this procedure. Let $\mathcal{E} = \{\mathsf{e}_i\}_{i=1}^N$ be such a set of the eigenvectors of the matrix $\mathsf{A}$ belonging to different eigenspaces that

$$\mathsf{b} = \sum_{i=1}^{N} \beta_i^{(0)} \mathsf{e}_i \ . \tag{2.1}$$

Denoting the coordinates of the residual $\mathsf{r}_j$ with respect to the basis $\mathcal{E}$ by $\beta_i^{(j)}$, we get, according to (1.9),

$$\mathsf{r}_j = \sum_{i=1}^{N} \beta_i^{(j)} \mathsf{e}_i \ = \left(\mathsf{I} - \frac{\mathsf{A}}{l_j}\right) \sum_{i=1}^{N} \beta_i^{(j-1)} \mathsf{e}_i \ = \sum_{i=1}^{N} \left(1 - \frac{\lambda_i}{l_j}\right) \beta_i^{(j-1)} \mathsf{e}_i \ ,$$
$$\beta_i^{(j)} = \left(1 - \frac{\lambda_i}{l_j}\right) \beta_i^{(j-1)} \tag{2.2}$$

for any $j \in \mathcal{N}$. Since $\|r_j\|^2$ is the sum of squares $[\beta_i^{(j)}]^2$, we require the sequences $\{\beta_i^{(j)}\}_{j\geq 0}$ to converge to zero for all $i = 1, \ldots, N$. If $\beta_i^{(0)} \neq 0$, then

$$p_i^{(n)} = \ln \frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|} = \ln \prod_{j=1}^{n} \frac{|\beta_i^{(j)}|}{|\beta_i^{(j-1)}|} = \sum_{j=1}^{n} \ln \frac{|\beta_i^{(j)}|}{|\beta_i^{(j-1)}|} = \sum_{j=1}^{n} \ln \left|1 - \frac{\lambda_i}{l_j}\right| \tag{2.3}$$

is a random value depending on the random values $l_j$ for $j = 1, \ldots, n$; the corresponding random variable $P_i^{(n)}$ is a sum of $n$ independent identically distributed elementary random variables

$$\widetilde{R}_i^{(j)} = \ln \left| 1 - \frac{\lambda_i}{L} \right| , \quad j = 1, 2, \ldots, n \tag{2.4}$$

In connection with (2.3), it will be useful to deal with the variables $\widetilde{R}_i$ more in detail.

The random variable $\widetilde{R}_i$ may be considered to be a special case of the random variable

$$R_x = \ln \left| 1 - \frac{x}{L} \right| .$$

We define the mean value function $\widehat{E}_R(x) = E(R_x)$ for all the values $x \in \langle \lambda, \Lambda \rangle$ for which the mean value $E(R_x)$ exists; for the other values $x \in \langle \lambda, \Lambda \rangle$, the function $\widehat{E}_R(x)$ remains undefined. Let $\phi$ be the probability measure on the set of subintervals of $\langle \lambda, \Lambda \rangle$ generated by the distribution function $F_L(l)$. The values $x \in \langle \lambda, \Lambda \rangle$ may be separated into three categories:

**Definition 1** Let $0 < \lambda \leq \Lambda$, $\phi$ be the probability measure on the subintervals of $\langle \lambda, \Lambda \rangle$. We denote the value $x \in \langle \lambda, \Lambda \rangle$ as

1. the point of the first type relative to the measure $\phi$, if the value $\widehat{E}_R(x)$ exists,
2. the point of the second type relative to the measure $\phi$, if the value $\widehat{E}_R(x)$ does not exist and $\phi(\{x\}) = 0$,
3. the point of the third type relative to the measure $\phi$, if $\phi(\{x\}) \neq 0$.

If $\hat{x}$ is the point of the third type relative to the measure $\phi$, then the value $\widehat{E}_R(\hat{x}) = E(R_{\hat{x}})$ does not exist, since in the definition of the mean value

$$\widehat{E}_R(\hat{x}) = \int\limits_{\langle \lambda, \Lambda \rangle} \ln \left| 1 - \frac{\hat{x}}{l} \right| \mathrm{d}\phi(l),$$

the integrated function is undefined on the set $\{\hat{x}\}$ the measure of which is $\phi(\{\hat{x}\}) \neq 0$. So each point of the interval $\langle \lambda, \Lambda \rangle$ belongs to just one of the sets cited in Definition 1.

The set $\mathcal{S}(\mathsf{A}, \mathsf{b})$ defined in Sect. 1 may be separated into three disjoint subsets $\mathcal{S}_j^{(\phi)}(\mathsf{A}, \mathsf{b})$, $j = 1, 2, 3$, such that the subset $\mathcal{S}_j^{(\phi)}(\mathsf{A}, \mathsf{b})$ is the set of all the points of the $j$th type relative to the measure $\phi$ belonging to the set $\mathcal{S}(\mathsf{A}, \mathsf{b})$. The properties of Algorithm 1 are characterized by the following statement:

**Theorem 1** *Let the Eq.* (1.5) *be solved by Algorithm* 1.

1. *If* $\mathcal{S}(\mathsf{A}, \mathsf{b}) = \mathcal{S}_3^{(\phi)}(\mathsf{A}, \mathsf{b})$*, then the random process gives the solution to the problem after a finite number of steps with probability one.*
2. *If* $\mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b}) \neq \emptyset$ *and*

$$E_{\phi,\mathsf{A},\mathsf{b}} = \max_{\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A},\mathsf{b})} \left( \widehat{E}_R(\lambda_i) \right) , \tag{2.5}$$

*then*

(a) *for* $E_{\phi,\mathsf{A},\mathsf{b}} > 0$*, the random process diverges with probability one,*
(b) *for* $E_{\phi,\mathsf{A},\mathsf{b}} < 0$*, the random process converges to the solution of the Eq.* (1.5) *with probability one; the convergence is almost surely R-linear, and the asymptotic logarithmic rate of convergence*

$$\Theta = \lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\mathsf{r}_0\|}{\|\mathsf{r}_n\|} = -E_{\phi,\mathsf{A},\mathsf{b}} \tag{2.6}$$

*with probability one,*
(c) *for* $E_{\phi,\mathsf{A},\mathsf{b}} = 0$*, the random process either does not converge, or its convergence is R-sublinear with probability one.*

3. *If* $\mathcal{S}_1^{(\phi)}(\mathsf{A},\mathsf{b}) = \emptyset$*, then the random process converges to the solution of the Eq.* (1.5) *R-superlinearly with probability one.*

*Proof* It is based on the strong law of large numbers, and the technical details are presented in “Appendix 1.” □

Theorem 1 indicates the possibility of the applicability of the Algorithm 1 to the solution of the Problem (1.5). However, it has not been verified whether the required conditions of the theorem can be met. The results [9] show that for the uniform distribution of the parameter $l_j$ on the interval $\langle\lambda, \Lambda\rangle$, the condition $E_{\phi,\mathsf{A},\mathsf{b}} < 0$ is satisfied independently of the location of the eigenvalues of matrix $\mathsf{A}$ in the interval $\langle\lambda, \Lambda\rangle$; so at least one distribution complying with our requirements exists.

The spectrum of matrix $\mathsf{A}$ is unknown in general. From Theorem 1, it then follows that the best universally applicable distribution of step lengths is such that the value

$$\overline{E}_\phi = \sup_{x\in\langle\lambda,\Lambda\rangle\wedge\widehat{E}_R(x)\ \text{exists}} \left(\widehat{E}_R(x)\right) \tag{2.7}$$

is the minimum possible (as shown below, the function $\widehat{E}_R(x)$ is defined almost everywhere on $\langle\lambda, \Lambda\rangle$ and the favorable statements (1) and (3) of Theorem 1 remain unused for almost all the matrices $\mathsf{A}$). The aim of the next section was to find such a distribution.

## 3 Optimal Step Length Distribution

Before the further calculations, we will formulate two useful lemmas:

**Lemma 1** *Let* $\lambda < \Lambda$*. Then, the value of the integral*

$$I_c = \int_\lambda^\Lambda \frac{\ln\left|1 - \frac{q}{x}\right|}{\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}x \tag{3.1}$$

*does not depend on the parameter* $q \in \langle\lambda, \Lambda\rangle$*.*

*Proof* The statement follows immediately from the results [13]. In "Appendix 2," we show a different form of the proof. □

**Lemma 2** *Let* $0 < \lambda < \Lambda$. *Then, the value of the integral*

$$I_d = \int\limits_{\lambda}^{\Lambda} \frac{\ln\left|1 - \frac{x}{q}\right|}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}x \tag{3.2}$$

*does not depend on the parameter* $q \in \langle\lambda, \Lambda\rangle$.

*Proof* It is described in "Appendix 2." □

Now, we can formulate the key statement of this section:

**Theorem 2** *Let* $0 < \lambda < \Lambda$. *Then, the integral*

$$I = \int\limits_{\lambda}^{\Lambda} \frac{\widehat{E}_R(x)}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}x$$

*exists, and its value does not depend on the choice of the distribution of the random variable L.*

*Proof* Let us consider the function of two variables $x, l$

$$f(x,l) = \frac{\ln\left|1 - \frac{x}{l}\right|}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}} \tag{3.3}$$

defined on the set $[x, l] \in \Omega = (\lambda, \Lambda) \times \langle\lambda, \Lambda\rangle \setminus \{x = l\}$. We define the measure on the set $\langle\lambda, \Lambda\rangle \times \langle\lambda, \Lambda\rangle$ based on the measure of the rectangles: if $A$, $B$ are intervals in $\langle\lambda, \Lambda\rangle$, $a_1 \leq a_2$ being the limits of the interval $A$, then

$$\mu(A \times B) = (a_2 - a_1) \cdot \phi(B)\ .$$

The integral

$$J = \int\limits_{\langle\lambda,\Lambda\rangle\times\langle\lambda,\Lambda\rangle} f(x,l)\,\mathrm{d}\mu = \int\limits_{\Omega} f(x,l)\,\mathrm{d}\mu\ . \tag{3.4}$$

exists, as proved in "Appendix 3." According to Fubini's theorem,

$$\begin{aligned} J &= \int\limits_{\lambda}^{\Lambda} \left[\int\limits_{\langle\lambda,\Lambda\rangle} \frac{\ln\left|1 - \frac{x}{l}\right|}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}\phi(l)\right]\mathrm{d}x \\ &= \int\limits_{\lambda}^{\Lambda} \frac{\widehat{E}_R(x)}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}x = I\ , \end{aligned} \tag{3.5}$$

this result verifies the existence of the integral $I$. The integration using the opposite order of variables gives

$$I = J = \int\limits_{\langle\lambda,\Lambda\rangle} \left[ \int\limits_{\lambda}^{\Lambda} \frac{\ln\left|1 - \frac{x}{l}\right|}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}x \right] \mathrm{d}\phi(l) = \int\limits_{\langle\lambda,\Lambda\rangle} I_d\,\mathrm{d}\phi(l) = I_d$$

as a consequence of Lemma 2. □

We can define the measure $\nu$ on $\langle\lambda, \Lambda\rangle$ using the rule

$$\nu(\Gamma) = \int\limits_{\Gamma} \frac{\mathrm{d}x}{x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(x - \frac{\Lambda+\lambda}{2}\right)^2}}$$

for any measurable set $\Gamma \subset \langle\lambda, \Lambda\rangle$. The proved theorem implies that the function $\widehat{E}_R(x)$ is defined $\nu$-almost everywhere on $\langle\lambda, \Lambda\rangle$. The measure $\nu$ and the standard Lebesgue measure on $\langle\lambda, \Lambda\rangle$ have the same system of zero sets, so almost all the points of the interval $\langle\lambda, \Lambda\rangle$ are the points of the first type relative to the measure $\phi$ for the arbitrarily chosen probability measure $\phi$, as mentioned at the end of Sect. 2.

Theorem 2 enables the formulation of a condition guaranteeing the maximum possible rate of convergence independent of the spectrum of matrix $\mathsf{A}$.

**Theorem 3** *The value $\overline{E}_\phi$ defined by* (2.7) *acquires its minimal possible value if and only if the distribution of the inverse step length parameter generates the function $\widehat{E}_R(x)$ satisfying*

$$\widehat{E}_R(x) = \overline{E}_\phi \tag{3.6}$$

*for $\nu$-almost all the values $x \in \langle\lambda, \Lambda\rangle$.*

*Proof* For $x \in \langle\lambda, \Lambda\rangle$, we can define the function

$$\widetilde{E}_R(x) = \begin{cases} \widehat{E}_R(x) \text{ if } \widehat{E}_R(x) \text{ is defined,} \\ \overline{E}_\phi \quad\;\; \text{otherwise} \end{cases}$$

This function is defined for all $x \in \langle\lambda, \Lambda\rangle$; further, $\widetilde{E}_R(x) = \widehat{E}_R(x)$ $\nu$-almost everywhere on $\langle\lambda, \Lambda\rangle$ and $\widetilde{E}_R(x) < \overline{E}_\phi$ if and only if $\widehat{E}_R(x) < \overline{E}_\phi$. Let $\varepsilon > 0$ and $\widetilde{E}_R(x) = \widehat{E}_R(x) \le \overline{E}_\phi - \varepsilon$ on some set $\Delta \subset \langle\lambda, \Lambda\rangle$ with the measure $\nu(\Delta) = \delta$. Due to (8.33), the random variable $\widetilde{R}_i$ is bounded from above; consequently, $\overline{E}_\phi \le \ln(\kappa(\mathsf{A}) - 1)$, and both the integrals

$$\int\limits_{\langle\lambda,\Lambda\rangle\setminus\Delta} \widetilde{E}_R(x)\,\mathrm{d}\nu(x)\,, \quad \int\limits_{\Delta} \widetilde{E}_R(x)\,\mathrm{d}\nu(x)$$

are finite as a consequence of (3.5). We estimate

$$\int\limits_{\Delta} \widetilde{E}_R(x)\,\mathrm{d}\nu(x) \le \int\limits_{\Delta} (\overline{E}_\phi - \varepsilon)\,\mathrm{d}\nu(x) = (\overline{E}_\phi - \varepsilon)\delta, \tag{3.7}$$

and since the integral on the left-hand side of (3.7) is finite, we can find a constant $k \geq 1$ satisfying

$$\int\limits_{\Delta} \widetilde{E}_R(x)\, \mathrm{d}\nu(x) = (\overline{E}_\phi - k\varepsilon)\delta \,.$$

According to Theorem 2

$$\begin{aligned} I_d = \int\limits_{\langle\lambda,\Lambda\rangle} \widehat{E}_R(x)\, \mathrm{d}\nu(x) \; &= \int\limits_{\langle\lambda,\Lambda\rangle\setminus\Delta} \widetilde{E}_R(x)\, \mathrm{d}\nu(x) + \int\limits_{\Delta} \widetilde{E}_R(x)\, \mathrm{d}\nu(x) \\ &\leq \int\limits_{\langle\lambda,\Lambda\rangle\setminus\Delta} \overline{E}_\phi\, \mathrm{d}\nu(x) + (\overline{E}_\phi - k\varepsilon)\delta \\ &= \overline{E}_\phi(\nu(\langle\lambda,\Lambda\rangle) - \delta) + (\overline{E}_\phi - k\varepsilon)\delta \; = \; \overline{E}_\phi \nu(\langle\lambda,\Lambda\rangle) - k\varepsilon\delta \,, \end{aligned} \tag{3.8}$$

analogously we can obtain the lower estimate

$$\begin{aligned} I_d > \int\limits_{\langle\lambda,\Lambda\rangle\setminus\Delta} (\overline{E}_\phi - \varepsilon)\, \mathrm{d}\nu(x) + (\overline{E}_\phi - k\varepsilon)\delta \\ = (\overline{E}_\phi - \varepsilon)(\nu(\langle\lambda,\Lambda\rangle) - \delta) + (\overline{E}_\phi - k\varepsilon)\delta \\ = (\overline{E}_\phi - \varepsilon)\nu(\langle\lambda,\Lambda\rangle) - (k-1)\varepsilon\delta \,. \end{aligned} \tag{3.9}$$

The estimate (3.8) implies

$$\overline{E}_\phi \geq \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)} \tag{3.10}$$

in any case. If $\delta > 0$ for some $\varepsilon > 0$, then (3.8) leads to

$$\overline{E}_\phi > \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)} \,. \tag{3.11}$$

On the other hand, if $\delta = 0$ for any $\varepsilon > 0$, then (3.9) gives

$$\overline{E}_\phi < \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)} + \varepsilon;$$

thus, regarding (3.10), we get the lowest possible value

$$\overline{E}_\phi = \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)} \tag{3.12}$$

in this case. This result, together with (3.11), validates the statement of the theorem. □

If we find the distribution of inverse step length parameter satisfying (3.6)—if it exists, then all we need to do is to guess it correctly—the goal of this section will be achieved. A result of this kind follows immediately from Lemma 1; however, we can formulate a much stronger statement:

**Theorem 4** *Let the Problem* (1.5) *be solved by Algorithm* 1*. Then, the only distribution of the inverse step length parameter $l$, which generates almost surely the iterative process with the maximum asymptotic convergence rate independently of the properties of matrix* $\mathsf{A}$ *and vector* $\mathsf{b}$*, is described by the probability density*

$$f(l) = \frac{1}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(l - \frac{\Lambda+\lambda}{2}\right)^2}} = \frac{1}{\pi\sqrt{(\Lambda - l)(l-\lambda)}} \tag{3.13}$$

*for $l \in (\lambda, \Lambda)$.*

*Proof* The only distribution of the inverse step length parameter, generating the function $\widehat{E}_R(x)$ which is constant almost everywhere on the interval $\langle\lambda, \Lambda\rangle$, is constructed in “Appendix 4.” However, this result does not guarantee that the assumptions of Theorem 3 are satisfied—the possibility $\widehat{E}_R(x) > \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)}$ for $x$ belonging to some set with zero measure is not yet excluded. Nevertheless, the found distribution function of the inverse step length parameter (8.53) is associated with the probability density (3.13), and as a consequence of Lemma 1, it holds

$$\widehat{E}_R(x) = \int\limits_{\lambda}^{\Lambda} \frac{\ln\left|1 - \frac{x}{l}\right|}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(l - \frac{\Lambda+\lambda}{2}\right)^2}} = \frac{I_c}{\pi} \tag{3.14}$$

for any $x \in \langle\lambda, \Lambda\rangle$, which means the function $\widehat{E}_R(x)$ is constant on the whole interval $\langle\lambda, \Lambda\rangle$ and satisfies condition (3.6) for all $x \in \langle\lambda, \Lambda\rangle$. That is why the distribution of the random variable $L$ with the probability density (3.13) provides the lowest achievable value $\overline{E}_\phi$ guaranteeing $E_{\phi,\mathsf{A},\mathsf{b}} \leq \overline{E}_\phi$ for all couples $\mathsf{A}$, $\mathsf{b}$. □

The last problem to solve is the calculation of the value $\overline{E}_\phi$ for the found optimal distribution. The relations (3.12) or (3.14) might be used, but direct computation is impossible because neither integral (3.1) nor integral (3.2) can be calculated using an antiderivative for all $q \in \langle\lambda, \Lambda\rangle$. Nevertheless, the way to express the desired value exists. The needed result may be derived based on the calculations carried out in [13]; another kind of the calculation uses the results from “Appendix 5.” The relations (8.57), (8.62) give

$$\overline{E}_\phi = -\operatorname{argcosh}\frac{\kappa(\mathsf{A})+1}{\kappa(\mathsf{A})-1} = -\ln\frac{\sqrt{\kappa(\mathsf{A})}+1}{\sqrt{\kappa(\mathsf{A})}-1}\,. \tag{3.15}$$

Note that value (3.15) of the almost surely achievable asymptotic rate of convergence of the random process is equal to the convergence rate of the Chebyshev polynomials method. This result is very favorable; in fact, we could not expect better. Indeed, the possibility that random polynomials (1.9) give systematically better results than the optimally constructed polynomials (the Chebyshev ones) seems to be extremely improbable at first sight.

## 4 Tests of the Optimal Random Algorithm

Algorithm 1 was tested for various problems constructed generally as the finite differences scheme for the homogenous Dirichlet's problem for the equation

$$-\Delta u(\mathsf{x}) + \alpha u(\mathsf{x}) = f(\mathsf{x})$$

on a rectangular $k$-dimensional domain, $k \leq 4$. The domain was discretized to $(n_1 + 1) \times (n_2 + 1) \times \cdots \times (n_k + 1)$ elements with the sizes $d_1 \times d_2 \times \cdots \times d_k$; the vertices of the elementary $k$-dimensional rectangles are used as the mesh points. The different values $d_i$ enable the generation of various structures of the spectrum of the obtained matrix $\mathsf{A}$, while the implementation of the shift $\alpha$ makes the arbitrary prescribed condition number $\kappa(\mathsf{A})$ possible. The problems were solved in $C$ `long double` precision, i.e., 63 significant binary digits.

Since the eigenvalues of the matrix $\mathsf{A}$ are not known through practical use of the iterative methods, the estimations of the boundary eigenvalues $\lambda$, $\Lambda$ extracted from the information obtained during the previous course of the solution were used in our tests:

$$\lambda \leq \overline{\lambda}^{(n)} = \min_{j<n} \frac{\langle \mathsf{r}_{j-1} | \mathsf{A}\mathsf{r}_{j-1} \rangle}{\|\mathsf{r}_{j-1}\|^2} = \min_{j<n} \frac{\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle}{\gamma_j \|\mathsf{r}_{j-1}\|^2}, \tag{4.1}$$

$$\Lambda \geq \underline{\Lambda}^{(n)} = \max_{j<n} \frac{\|\mathsf{A}\mathsf{r}_{j-1}\|^2}{\langle \mathsf{r}_{j-1} | \mathsf{A}\mathsf{r}_{j-1} \rangle} = \max_{j<n} \frac{\|\mathsf{r}_{j-1} - \mathsf{r}_j\|^2}{\gamma_j \langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle}; \tag{4.2}$$

in the $n$-th step, the value $l_n$ was chosen from the interval $\langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$ according to distribution (3.13) where $\lambda = \overline{\lambda}^{(n)}$, $\Lambda = \underline{\Lambda}^{(n)}$ were used.[4] This technique makes the process a little bit slower in comparison with the optimal value (3.15) because the distribution of the inverse step length created in this way suppresses the values $l_j$ in the boundary areas around the eigenvalues $\lambda$, $\Lambda$. In the end, the condition (3.6) is not satisfied in general.

The scalar products $\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle$ used in (4.1), (4.2) were applied at the same time as the criterion for the termination of the process—the iterations were stopped when the value

$$\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle \leq 0 \tag{4.3}$$

was obtained. The relation (4.3) may be satisfied in three different situations, but each one of them constitutes a reason for the termination of the process:

[4] Only in the first step, no vectors $\mathsf{r}_{j-1} - \mathsf{r}_j$ and the values $\gamma_j$ for $j \leq 0$ are defined, so the values $\overline{\lambda}^{(1)}$, $\underline{\Lambda}^{(1)}$ were expressed directly using the vector $\mathsf{A}\mathsf{r}_0$.

1. $\mathsf{r}_{j-1} = 0$ means that the solution is achieved,
2. $\mathsf{r}_j = \mathsf{r}_{j-1}$ may be considered equivalent to $\mathsf{x}_{j-1} = \mathsf{x}_j = \mathsf{x}_{j-1} - \gamma_j \mathsf{r}_{j-1}$. In this case, the residual $\mathsf{r}_{j-1}$ is so small that the continuation of the process does not deliver any new results,
3. $\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle < 0$ signifies that the numerical errors are so important that the continuation of the process would be produced largely by the numeric noise with dubious information value.

The criterion (4.3) for terminating the process was used (instead of the one mentioned in step 4 of the algorithm) with the aim to determine the available accuracy of the method.

Utilizing all the ideas mentioned above, Algorithm 1 was modified:

**Algorithm 2** Let $\{F_{L_n}(l)\}_{n\geq 1}$ be a set of distribution functions of random variables $L_n$ having their values in $\langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$. The system of Eq. (1.5) is solved as follows:

1. $\mathsf{x}_0 = 0,\ j = 1,$
2. $\overline{\lambda}^{(1)} = \frac{\langle \mathsf{r}_0 | \mathsf{A}\mathsf{r}_0 \rangle}{\|\mathsf{r}_0\|^2},\ \underline{\Lambda}^{(1)} = \frac{\|\mathsf{A}\mathsf{r}_0\|^2}{\langle \mathsf{r}_0 | \mathsf{A}\mathsf{r}_0 \rangle},$
3. $l_j$ is a value of the random variable $L_j$,
4. $\mathsf{x}_j = \mathsf{x}_{j-1} + \frac{1}{l_j}\mathsf{r}_{j-1},\ \mathsf{r}_j = \mathsf{b} - \mathsf{A}\mathsf{x}_j,$
5. if $\|\mathsf{r}_j\| <$given $\varepsilon$ or $\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle \leq 0$, then stop
6. $$\hat{\lambda}^{(j+1)} = \frac{l_j \langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle}{\|\mathsf{r}_{j-1}\|^2}, \quad \hat{\Lambda}^{(j+1)} = \frac{l_j \|\mathsf{r}_j - \mathsf{r}_{j-1}\|^2}{\langle \mathsf{r}_{j-1} | \mathsf{r}_{j-1} - \mathsf{r}_j \rangle}, \tag{4.4}$$
   $\overline{\lambda}^{(j+1)} = \min\left(\overline{\lambda}^{(j)}, \hat{\lambda}^{(j+1)}\right),\ \underline{\Lambda}^{(j+1)} = \max\left(\underline{\Lambda}^{(j)}, \hat{\Lambda}^{(j+1)}\right),$
7. go to step 3 for the next $j$.

The distribution functions $F_{L_n}(l)$ were defined—according to (3.13)—by the probability density of the inverse step length parameter $l$

$$f_{L_n}(l) = \frac{1}{\pi\sqrt{\left(\underline{\Lambda}^{(n)} - l\right)\left(l - \overline{\lambda}^{(n)}\right)}}. \tag{4.5}$$

This process is not exactly stationary because the distribution functions in the particular steps can differ from each other; due to this effect, the process does not correspond to the Theorem 1 exactly. Nevertheless, the sequences $\{\overline{\lambda}^{(n)}\}_{n\geq 1}, \{\underline{\Lambda}^{(n)}\}_{n\geq 1}$ often become stationary after a relatively small number of steps, and thereafter, the assumptions of Theorem 1 are satisfied, and its statement is applicable.

The tests were realized for the problems with the dimension $M_k = n_1 \times n_2 \times \cdots \times n_k$, where $n_1 = 8$, $n_2 = 22$, $n_3 = 13$, $n_4 = 18$, and $k = 2, 3, 4$. The problems with the condition numbers $\kappa(\mathsf{A}) = 90.51$, $\kappa(\mathsf{A}) = 1024$, $\kappa(\mathsf{A}) = 1.16 \times 10^4$ and $\kappa(\mathsf{A}) = 1.31 \times 10^5$ were solved. Using the parameters $d_i$, we created two kinds of matrices $\mathsf{A}$—the high and low densities of the eigenvalues in the boundary regions of the interval $\langle \lambda, \Lambda \rangle$ were set.

In order to judge the influence of the truncation errors, we transformed all the solved problems into the basis, in which the matrix $\mathsf{A}$ has the diagonal form, and we compared the solutions of the original problem and its diagonalized form.

Since in practical computing the first criterion of the process termination formulated by step 5 of Algorithm 2 is used, the achievement of the prescribed small value of $\|\mathsf{r}_n\|$ is sufficient for process interruption; the eventual subsequent larger values $\|\mathsf{r}_j\| > \|\mathsf{r}_n\|$ for $j > n$ are insignificant. Therefore, we monitored, apart from the residuals themselves, the "best previous iterations," i.e., the values

$$\|\mathsf{r}_n\|^* = \min_{i \le n} (\|\mathsf{r}_i\|) \ ;$$

these values are preferable for the characterization of the algorithm efficiency.

Approximately two hundred random iterative processes for each monitored configuration of the general problem were realized. The obtained data were statistically treated.

In the graphs in Fig. 1, the typical courses of the solution of the problem with $\kappa(\mathsf{A}) = 90.51$ for various dimensions of the diagonalized problem and different structures of the spectrum of matrix $\mathsf{A}$ are displayed. For the individual samples, there is impossible to state any dependence of the computation course on the problem dimension or on the spectrum distribution.

Practically, all the residuals were larger than the values corresponding to the Q-linearly convergent process with the rate of convergence (3.15)—these hypothetical

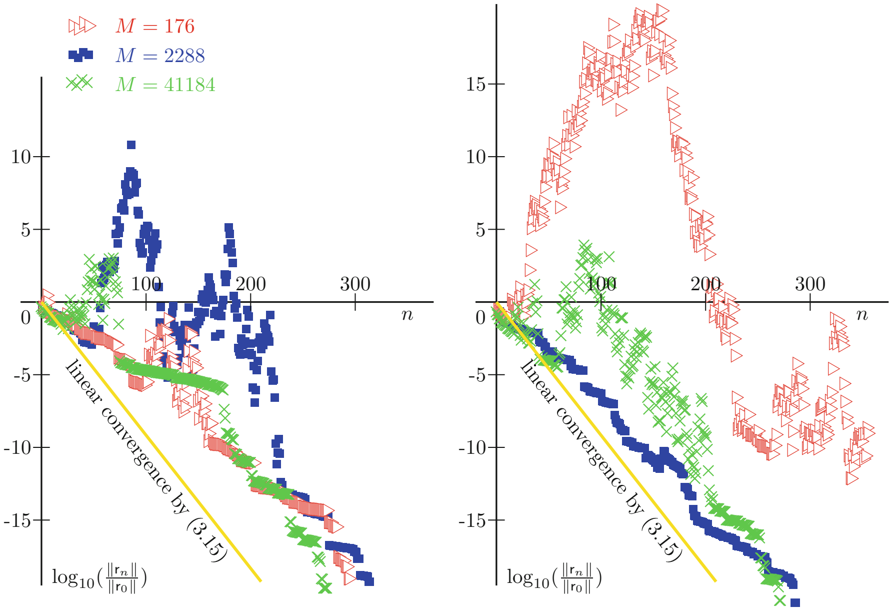

**Fig. 1** Demonstration of the solution course of the problem with $\kappa(\mathsf{A}) = 90.51$ (*left*—high density of the eigenvalues close to $\lambda$, $\Lambda$, *right*—low density of the eigenvalues close to $\lambda$, $\Lambda$)

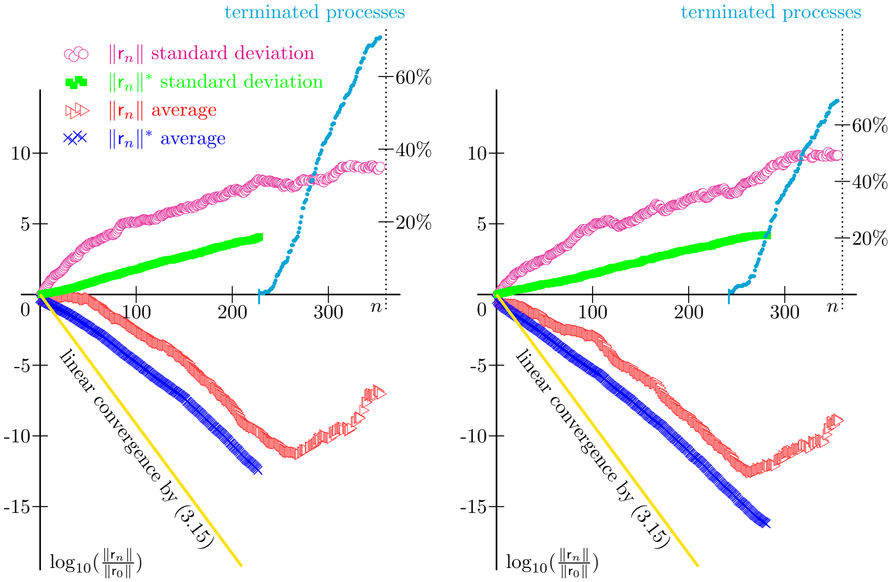

**Fig. 2** Statistical treatment of the results for the diagonalized problem with $\kappa(\mathsf{A}) = 90.51$ and high density of the eigenvalues close to $\lambda$, $\Lambda$ (*left*—$M = 176$, *right*—$M = 41184$)

values lie on the displayed half line from $[0, 0]$ in the direction down to the right. This effect may be easily explained proceeding from the properties of the random residual coordinates $\beta_i^{(n)}$. Since $\|r_n\|^2 = \sum(\beta_i^{(n)})^2$, the norm of the residual is determined by the largest values $|\beta_i^{(n)}|$. Even if only one coordinate $|\beta_i^{(n)}|$ exceeds its value corresponding to (3.15) substantially, the residual norm increases significantly too. Contrarily, the small residual norm requires all the coordinates $|\beta_i^{(n)}|$ to be sufficiently small simultaneously. This event is extremely improbable (despite the generally nonzero correlations between the random variables $R_x$, $R_y$ at $x \neq y$), particularly for the large dimensions $M$ of the problem. Nevertheless, in spite of the described effect (which is not contrary to the proved $R$-linearity),[5] the predicted convergence properties are apparent.

The statistical treatment of the realized tests provides much more transparent results. In Fig. 2, the average values and standard deviations related to $\|r_n\|$ and $\|r_n\|^*$ for the well-conditioned problem with $\kappa(\mathsf{A}) = 90.51$ and diagonalized matrix $\mathsf{A}$ are displayed. Many processes were terminated according to (4.3) relatively early; information about these events is also provided. The statistics for the values $\|r_n\|$ were based on the data from the not yet terminated processes. The information value of the results at $n >\sim 250$ steps is very low therefore (the "good" results originating from the terminated processes are not included). In the statistics concerning the values $\|r_n\|^*$, the last magnitudes $\|r_n\|^*$ before the eventual process interruption were used for all

[5] The importance of the deviations in comparison with the $R$-linear descent of the residual norm decreases by increasing the number of steps, but the limited accuracy of the computation does not allow us realize the related observation.

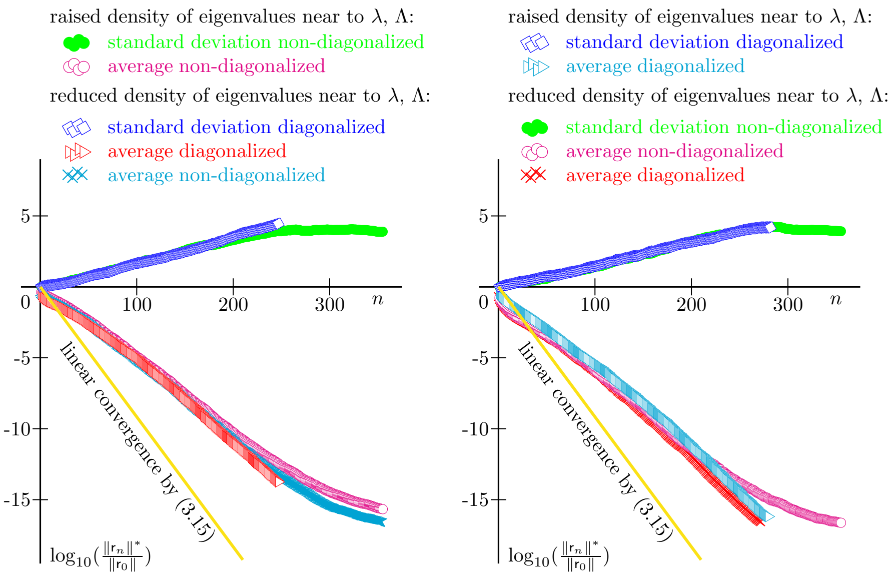

**Fig. 3** Comparison of the statistical characteristics for the problem with $\kappa(\mathsf{A}) = 90.51$ and different spectral structures, including the influence of diagonalization (*left*—$M = 176$, *right*—$M = 41184$)

following step numbers $m > n$. If the zero residual was reached after $n$ steps in an experiment, then the statistical characteristics were not further calculated for $m > n$.

The results for the problems with very diverse dimensions are rather similar to one another. The difference between the results for $\|\mathsf{r}_n\|$ and $\|\mathsf{r}_n\|^*$ is apparent. Since the importance of the random variable $\|\mathsf{r}_n\|^*$ for the evaluation of the process convergence exceeds the importance of $\|\mathsf{r}_n\|$, we left the observation of $\|\mathsf{r}_n\|$ from this point on.

The influence of the spectral structure and the diagonalization were observed. The results are displayed in Fig. 3. All four possible combinations were realized; if the corresponding results are missing in Fig. 3, then they were very similar to the ones obtained using the complementary spectral structure. The high concentration of the eigenvalues of matrix $\mathsf{A}$ in the boundary areas of the interval $\langle\lambda, \Lambda\rangle$ sometimes caused the rather slower convergence of the process. The truncation errors in the non-diagonalized problems became evident in the final phases of the process; they are more considerable for the problems with high dimension $M$.

Similar tests were implemented for $\kappa(\mathsf{A}) = 1024$. Some significant results are displayed in Fig. 4. By increasing the dimension of the problem, the convergence properties of the method improve in both the diagonalized and non-diagonalized cases. This result may be surprising—since the square of the residual norm $\|\mathsf{r}_j\|^2$ is a sum of $N$ nonnegative random variables $[\beta_i^{(j)}]^2$ and additional terms of this kind may increase the probability of the positive deviations of $\|\mathsf{r}_j\|^2$ from their expected values, we could anticipate a worsening of the convergence properties of the method by increasing the dimension of the problem. Nevertheless, the logarithms of the coordinates $|\beta_i^{(j)}|, |\beta_k^{(j)}|$ are very strongly positively correlated, when the eigenvalues $\lambda_i \approx \lambda_k$, as proved in “Appendix 5.” Increase in the density of the eigenvalues of the matrix $\mathsf{A}$ inside the

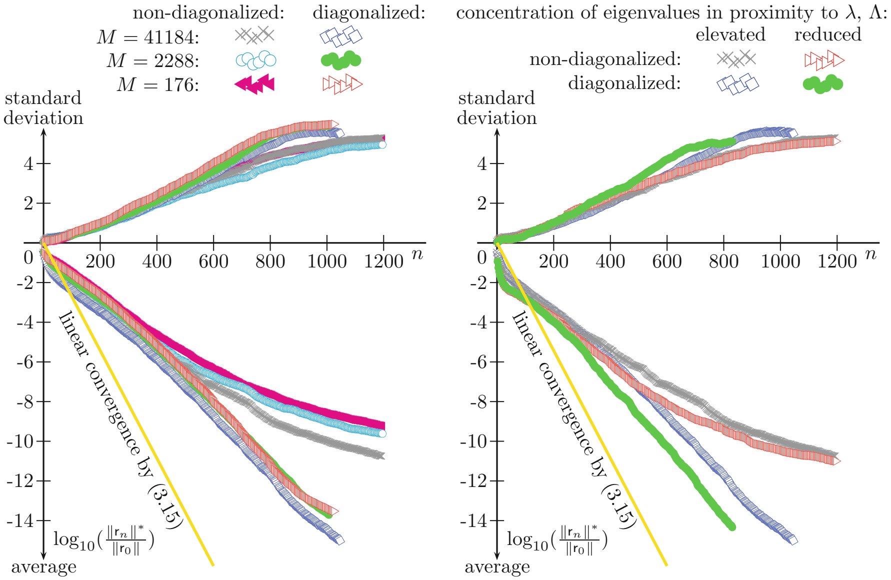

**Fig. 4** Statistical characteristics for the solutions of the problem with $\kappa(\mathsf{A}) = 1024$ and different dimensions (*left*) and spectral structures (*right*) regarding the influence of the matrix $\mathsf{A}$ diagonalization. Average values are negative, sample standard deviations positive.

interval $\langle\lambda, \Lambda\rangle$ creates new eigenvalues which are very close to the "old," already existing eigenvalues. On the newly created eigenspaces, the deviations of $|\beta_k^{(j)}|$ from their mean values are very similar to the deviations of the coordinates $|\beta_i^{(j)}|$ on the "old" eigenspaces, and therefore, it causes only negligible additional fluctuations of the residual norm.

The evident difference between all the results for the original problem and its diagonalized form at $n >\sim 700$ steps does not require a statistical test. The worsening of the convergence of the non-diagonalized problem by increasing the condition number is caused by the more frequent occurrence of the large coordinates $|\beta_i^{(j)}|$. In such situations, the truncation errors affect the other coordinates $\beta_{i'}^{(j+1)}$ for $i' \neq i$ negatively; this does not occur, when matrix $\mathsf{A}$ is diagonalized.

The statement concerning the elevated presence of the large coordinates $|\beta_i^{(j)}|$ is based on the following consideration: For a sufficiently large step number $j$, the distribution of the random value $\ln|\beta_i^{(j)}|$ is close to the normal distribution (due to the central limit theorem). The probability of occurring large values $\ln|\beta_i^{(j)}|$ is related to the variance of the random variable $P_i^{(j)}$ defined before (2.4); this variance is equal to $j \cdot \sigma^2(R_i)$. If the distribution (3.13) is used in Algorithm 1, the value $\sigma(R_i) \sim \pi$ except for the extremely small eigenvalues $\lambda_i \approx \lambda$, as calculated in "Appendix 5." When $\kappa(\mathsf{A})$ is large, the positive fluctuations of the random variable $P_i^{(j)}$ from its mean value, which are characterized by the standard deviation $\pi\sqrt{j}$, significantly outshine the negative mean value $E(P_i^{(j)}) = j\overline{E}_\phi \approx -\frac{2j}{\sqrt{\kappa(\mathsf{A})}}$ with probability close to $\frac{1}{2}$.

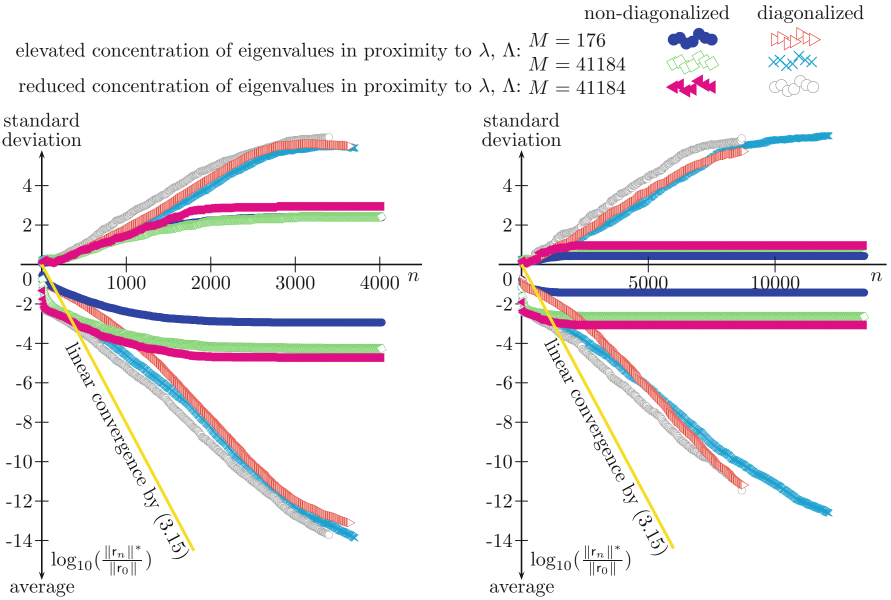

**Fig. 5** Statistical characteristics for the solutions of the problem with $\kappa(\mathsf{A}) = 1.16 \times 10^4$ (*left*) and $\kappa(\mathsf{A}) = 1.31 \times 10^5$ (*right*). Average values are negative, sample standard deviations positive

The dependence of the results on the spectral character of matrix A can be observed. The process converges more quickly for matrices with a reduced concentration of the eigenvalues in proximity to the bounds $\lambda$, $\Lambda$. This effect is the most distinct when solving larger problems (on the right part of Fig. 4. the results for $M = 41184$ are displayed). We did not test its statistical significance, but it is possible to explain it simply—the convergence factor of the process on the eigenspaces with $\lambda_i > \underline{\Lambda}^{(n)}$, $\lambda_i < \overline{\lambda}^{(n)}$ is worse than the value (3.15) during the initial phases of the process, until the estimates $\overline{\lambda}^{(n)}$, $\underline{\Lambda}^{(n)}$ are sufficiently precise.[6] The slower convergence of the residuals is caused by the higher portion of the eigenspaces of this kind, and therefore. we can suppose that the observed dependence is not accidental.

The mean values $\log \frac{\|\mathsf{r}_n\|^*}{\|\mathsf{r}_0\|}$ are visually distant from the values corresponding to (3.15) more than in the case of $\kappa(\mathsf{A}) = 90.51$. Nevertheless, the increase in the ratio of the steps number needed to achieve the given value $\log \frac{\|\mathsf{r}_n\|^*}{\|\mathsf{r}_0\|}$ by the realized process to the number of steps corresponding to the relation (3.15) is of little significance.

The results for the problems with $\kappa(\mathsf{A}) = 1.16 \times 10^4$ and $\kappa(\mathsf{A}) = 1.31 \times 10^5$ are displayed in Fig. 5. All the tendencies mentioned in connection with the results for $\kappa(\mathsf{A}) = 1024$ are once again observable; some of them are still more distinct, mainly the pitiful properties of the processes with the non-diagonalized matrices—these processes diverged (after a favorable initial part) without exception (the values $\|\mathsf{r}_n\|^*$ remain constant for large $n$ in this case).

[6] The convergence factor (3.15) is applicable only to the eigenspaces with $\lambda_i \in \langle\overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)}\rangle$.

In the initial phases of the process, the values $\|\mathsf{r}_n\|^*$ are more favorable than forecast (3.15); this effect is more conspicuous at the larger problem dimension $M$. We suppose it may be generated by the use of inexact estimates (4.1), (4.2) of the extremal eigenvalues of matrix $\mathsf{A}$ at the beginning of the process.[7] As follows from the comparison of the results for the diagonalized problem with $\kappa(\mathsf{A}) = 1.31 \times 10^5$, the positive influence of the elevated problem dimension may fade out at a larger steps number.

## 5 Minimization of Non-quadratic Functions

All the examples presented in Sect. 4 confirm that the process generated by Algorithm 2 using distribution (4.5) behaves in accordance with the results of Theorem 1. Nevertheless, the asymptotic convergence rate cannot be achieved during any finite part of the iterative process, so the mean convergence rate achieved using Algorithm 2 is poorer than the convergence rate of the Chebyshev polynomials method. Since the convergence of the Chebyshev polynomials method cannot be faster than the convergence of the conjugate gradient method, the conjugate gradient method is preferable compared to the random steepest descent method solving the problem (1.5).

On the other hand, the convergence of the proposed method compared to the conjugate gradient method is not too slow, and the method might be used to solve the original problem, i.e., the minimization of a general twice differentiable function of $M$ variables $V(\mathsf{x})$. The Hessian matrix is variable in this case, and the methods of higher order, including the conjugate gradient method, lose partially their efficiency due to the dependence of the calculation of the next iterative $\mathsf{x}_{n+1}$ on the actually useless Hessian matrices $\mathsf{H}(\mathsf{x}_j)$ for $j < n$. In contrast, the impact of the history of the calculations on the variables used in the steepest descent Algorithm 2 is only secondary, through the values $\overline{\lambda}^{(n)}$, $\underline{\Lambda}^{(n)}$. It elevates its robustness in comparison with other methods.

The main problem associated with the non-constantness of the Hessian matrix is the suitable choice of the boundary values $\overline{\lambda}^{(n)}$, $\underline{\Lambda}^{(n)}$ in the step 6 of Algorithm 2. Our considerations are as follows: If, even though once only, the estimate $\hat{\lambda}^{(n)}$ is chosen too small, the caused mistake cannot be corrected during the subsequent course of the process, since the values $\overline{\lambda}^{(j)}$ cannot exceed the value $\hat{\lambda}^{(n)}$ for any $j \geq n$. It produces an overestimate of the condition number of the problem and—consequently—the deterioration of the convergence rate of the procedure. In contrast to this, a too large estimate $\hat{\lambda}^{(n)}$ does not exclude the possible improvement of the value $\overline{\lambda}^{(j)}$ at the subsequent steps of the process. Therefore, if we have a set $\overline{\mathcal{L}}^{(n)}$ of the possible estimates $\hat{\lambda}^{(n)}$, we always prefer the selection $\hat{\lambda}^{(n)} = \sup(\overline{\mathcal{L}}^{(n)})$. Naturally, the values $\hat{\lambda}^{(n)} \leq 0$ are unacceptable (this may occur, when the function $V$ is not convex); such values must not be considered. Analogously, we always prefer the choice of the smallest possible (and positive) estimate $\hat{\Lambda}^{(n)}$.

[7] The narrowed interval of the possible values $l_j$ causes the more efficient suppression of the coordinates $\beta_i^{(n)}$ for $\lambda_i$ from the actual interval $\langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$; the number of the eigenvalues $\lambda_i \notin \langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$ is much lower than the problem dimension $M$, and the insufficient efficiency of the process on the related eigenspaces does not significantly affect the total value $\|\mathsf{r}_n\|$, while the decrease in the norm $\|\mathsf{r}_n\|$ is small.

A detailed analysis of this problem is presented in "Appendix 6." Based on the results derived there, we can formulate the complete random steepest descent algorithm designated for the minimization of a continuously differentiable function of several variables.[8]

**Algorithm 3** Let $\{F_{L_n}(l)\}_{n\geq 1}$ be a set of distribution functions of random variables $L_n$ having their values in $\langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$. The continuously differentiable function $V(\mathsf{x})$ bounded from below is minimized as follows:

1. $\mathsf{x}_0 = \mathsf{0}$, $\underline{\Lambda}^{(1)} = 0$, $\overline{\lambda}^{(1)} = \infty$, $\mathsf{g}_0 = \nabla V(\mathsf{0})$, $j = 1$, set $\gamma_1 > 0$ adequately small,
2. increase $\gamma_1$ to its double, while $4[V(-\gamma_1 \mathsf{g}_0) - V(\mathsf{0})] + \gamma_1 \|\mathsf{g}_0\|^2 < 0$,
3. $\mathsf{p}_j = -\gamma_j \mathsf{g}_{j-1}$, $\mathsf{x}_j = \mathsf{x}_{j-1} + \mathsf{p}_j$, $\mathsf{g}_j = \nabla V(\mathsf{x}_j)$,
4. if $\|\mathsf{g}_j\| <$given $\varepsilon$, then stop
5. calculate

$$\begin{aligned} a &= \langle \mathsf{g}_j | \mathsf{p}_j \rangle + \langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle - 2\left[V(\mathsf{x}_j) - V(\mathsf{x}_{j-1})\right] \,, \\ b &= 3\left[V(\mathsf{x}_j) - V(\mathsf{x}_{j-1})\right] - 2\langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle - \langle \mathsf{g}_j | \mathsf{p}_j \rangle \,, \\ t_0 &= \frac{1}{6a}\left[\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2} - 2b\right] \,, \quad \text{if } a \neq 0 \end{aligned}$$

   (a) if $a \leq 0$ and $b \leq 0$, or $b^2 \leq 3a\langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle$, then $\overline{\lambda}^{(j+1)} = \overline{\lambda}^{(j)}$, $\underline{\Lambda}^{(j+1)} = \underline{\Lambda}^{(j)}$ and go to step 8,
   (b) if $a \leq 0$ and $b > 0$, then $\hat{\lambda}^{(j+1)} = \frac{2b}{\|\mathsf{p}_j\|^2}$,
   (c) if $a > 0$ and $\langle \mathsf{g}_j | \mathsf{p}_j \rangle \geq 0$, then $\hat{\lambda}^{(j+1)} = \frac{6a+2b}{\|\mathsf{p}_j\|^2}$,
   (d) if $a > 0$ and $\langle \mathsf{g}_j | \mathsf{p}_j \rangle < 0$, then $\hat{\lambda}^{(j+1)} = \frac{2\sqrt{b^2 - 3a\langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle}}{\|\mathsf{p}_j\|^2}$,
6. (a) if $a \leq 0$ and $b \in (0, \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2} \rangle$, or $a \geq 0$ and $b \geq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}$, then $\hat{\Lambda}^{(j+1)} = \frac{4b^2 + \|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}{2b\|\mathsf{p}_j\|^2}$,
   (b) if $a > 0$ and $b \leq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}$, or $a < 0$ and $b \geq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}$, then
      i. if $t_0 \leq 1$, then $\hat{\Lambda}^{(j+1)} = 2\frac{\sqrt{\|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}}{\|\mathsf{p}_j\|^2}$,
      ii. if $t_0 > 1$ and ($a < 0$ or $\langle \mathsf{g}_j | \mathsf{p}_j \rangle \geq 0$), then $\hat{\Lambda}^{(j+1)} = \frac{(6a+2b)^2 + \|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}{(6a+2b)\|\mathsf{p}_j\|^2}$,
      iii. if $t_0 > 1$ and $a > 0$ and $\langle \mathsf{g}_j | \mathsf{p}_j \rangle < 0$, then $\hat{\Lambda}^{(j+1)} = \frac{(6a\tilde{t}+2b)^2 + \|\mathsf{g}_j\|^2 \|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j | \mathsf{p}_j \rangle^2}{(6a\tilde{t}+2b)\|\mathsf{p}_j\|^2}$,
      where $\tilde{t} = \min(t_0, \frac{-b + \sqrt{b^2 - 3a\langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle}}{3a})$,
7. $\overline{\lambda}^{(j+1)} = \min(\overline{\lambda}^{(j)}, \hat{\lambda}^{(j+1)})$, $\underline{\Lambda}^{(j+1)} = \max(\underline{\Lambda}^{(j)}, \hat{\Lambda}^{(j+1)})$,
8. if $\overline{\lambda}^{(j+1)} < \underline{\Lambda}^{(j+1)}$, then $l_{j+1}$ is a value of the random variable $L_{j+1}$,
   else $l_{j+1} = \overline{\lambda}^{(j+1)}$

[8] The algorithm itself does not need the second derivatives; nevertheless, they are hidden in the considerations in "Appendix 6."

9. $\gamma_{j+1} = \frac{1}{l_{j+1}}$,
10. go to step 3 for the next $j$.

We applied this algorithm to the function

$$V(\mathsf{x}) = \left(1 + \frac{1}{2}\langle \xi(\mathsf{x}, \underline{\mathsf{x}}) | \mathsf{A}\xi(\mathsf{x}, \underline{\mathsf{x}})\rangle\right)^m \tag{5.1}$$

where $\underline{\mathsf{x}}$ is the solution to the problem solved in Sect. 4, $\mathsf{A}$ is the related matrix, $m$ is a real positive parameter, and the transformed coordinates $\xi$ are defined as

$$\begin{aligned} \xi_1 &= x_1 - \underline{x}_1 + \alpha\left(\sqrt{1 + \sum_{j=2}^{M}(x_j - \underline{x}_j)^2} - 1\right), \\ \xi_j &= x_j - \underline{x}_j \quad \text{for } j \geq 2 \end{aligned}$$

for an $\alpha \in \mathcal{R}$. The function $V(\mathsf{x})$ has the only one minimum in the point $\underline{\mathsf{x}}$, and its Hessian matrix in this point is the matrix $m \cdot \mathsf{A}$. The transformation of coordinates and the level sets of the function $V(\mathsf{x})$ in two-dimensional case at $m = 1$ are illustrated on the left part of Fig. 6 (the non-convex character of the function $V(\mathsf{x})$ is apparent); the right part of Fig. 6 contains the results of processing the 2288-dimensional problem with $\kappa(\mathsf{A}) = 90.51$, $\alpha = 1$, $m = 1$. The value $\varepsilon = 10^{-12}\|\mathsf{g}_0\|$ was used in the step 4 of the algorithm.

The results of the procedure generated by the Algorithm 3 with the distribution functions $F_{L_n}(l)$ associated with the densities (4.5) were compared with various concurrent methods. The first one was the Polak–Ribière conjugate gradient method; the

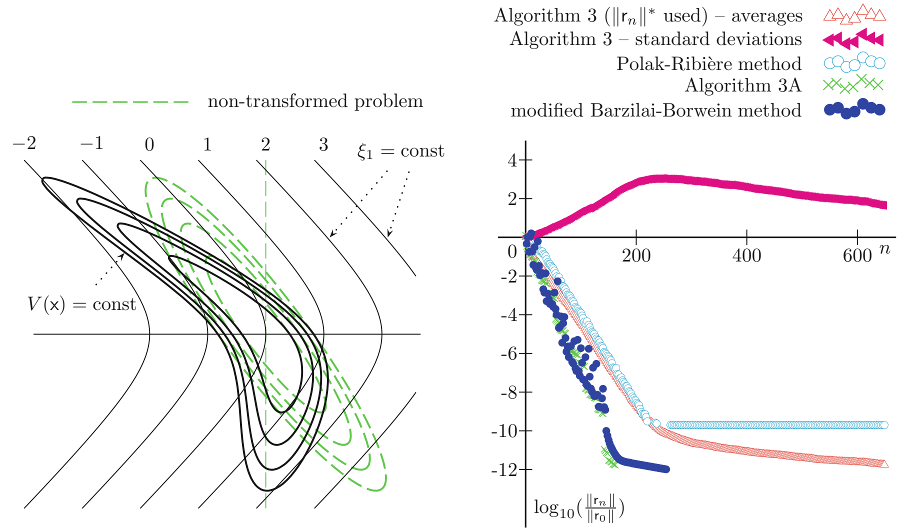

**Fig. 6** Illustration of the transformed problem in two dimensions for $\alpha = 0.8$, $m = 1$ (*left*) and results of the solution for $\alpha = 1$, $m = 1$ at $\kappa(\mathsf{A}) = 90.51$ (*right*)

linear search was realized using the cubic polynomial interpolation proposed in [6]. Since the calculation of gradients is the most expensive part of any gradient method, the number of calculated gradients is used as the measure of time; usually, when the conjugate gradient method is used, the number of calculated gradients exceeds a little bit the doubled number of steps.[9]

Further, we have modified the deterministic algorithm proposed in [13], which is also based on the density (3.13). The new Algorithm 3A coincides with Algorithm 3, except for the step 8, which is replaced by

$$\begin{aligned}
&8\text{A.if } \overline{\lambda}^{(j+1)} < \underline{\Lambda}^{(j+1)}, \text{ then}\\
&l_{j+1} = \overline{\lambda}^{(j+1)} + \frac{1}{2}\left[1 + \cos\left(\left\{\frac{(j+1)(\sqrt{5}-1)}{2}\right\}\pi\right)\right](\underline{\Lambda}^{(j+1)} - \overline{\lambda}^{(j+1)}),\\
&\quad \text{else } l_{j+1} = \overline{\lambda}^{(j+1)},
\end{aligned}$$

the symbol $\{x\}$ denotes the fractional part of $x$ here. The formula corresponds to the distribution generated by the density (3.13); a modification using general distribution functions $\{F_{L_n}(l)\}_{n\geq 1}$ is also possible. Contrary to [13], we use the whole interval $\langle \overline{\lambda}^{(n)}, \underline{\Lambda}^{(n)} \rangle$ without any restriction parameter $\tau$ introduced there,[10], and we do not work with the symmetric pairs $l_{2k-1}, l_{2k}$ satisfying the relation $l_{2k-1} + l_{2k} = \underline{\Lambda}^{(2k)} + \overline{\lambda}^{(2k)}$.

The last algorithm applied to the solution of the given problem was a modification of the Barzilai–Borwein algorithm. Since this algorithm uses the exact step length parameter belonging to the preceding step, and this parameter may be undefined for non-convex functions, we used in (1.1) the step length parameter

$$\gamma_j = \frac{1}{\text{the last available } \hat{\lambda}^{(i)}\ ,\quad i \leq j}\ . \tag{5.2}$$

We did not apply the method [14] which is usually cited in this context; the algorithm proposed therein converges, but it suppresses one of the characteristic properties of the Barzilai–Borwein method—the possibility that there exist some pairs $[i, j]$, $i < j$ for which the iterates $\mathsf{x}_j$ are significantly more distant from the solution $\underline{\mathsf{x}}$ to the problem than the preceding iterates $\mathsf{x}_i$, and the corresponding values $V(\mathsf{x}_j) \gg V(\mathsf{x}_i)$. However, as a consequence of the choice (5.2), the convergence of the process is not guaranteed.

Surprisingly, when $\alpha = 1$ and $m = 1$, the worst results were obtained for the conjugate gradient method; moreover, its achievable accuracy is apparently limited. Nevertheless, the mean values obtained by random Algorithm 3 are only slightly better than the results of Polak–Ribière method. Apparently better were the results achieved by the Algorithm 3A and by the modified Barzilai–Borwein method. These two methods may be rated as approximately equal.

[9] The number of steps is equal to the number of calculated gradients for other studied methods.

[10] A “safety” parameter of this kind is implicitly present in the algorithm as a consequence of the inexact estimates $\overline{\lambda}^{(n)} > \lambda$, $\underline{\Lambda}^{(n)} < \Lambda$.

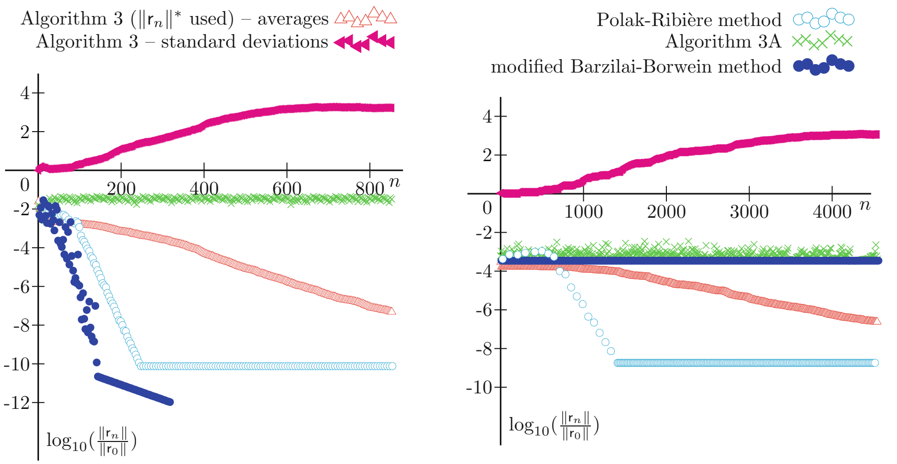

**Fig. 7** Minimization of the function (5.1) for $\alpha = 1$, $m = 0.6$ (*left*) and $m = \frac{1}{3}$ (*right*) with $\kappa(\mathsf{A}) = 90.51$

The growth of the function $V(\mathsf{x})$ for $\|\mathsf{x} - \underline{\mathsf{x}}\| \to \infty$ may be reduced using the parameter $m < 1$; for $m < \frac{1}{2}$, the function $V$ totally loses its convex character for large $\mathsf{x}$. We tested the values $m = 0.6$, $m = \frac{1}{3}$ with $\alpha = 1$. The results are displayed in Fig. 7. Algorithm 3A produces very bad results in both the studied cases. The modified Barzilai–Borwein method is very efficient at $m = 0.6$, but for $m = \frac{1}{3}$, it totally loses its applicability. The conjugate gradient method is reliable in both the cases, although its results lag behind the Barzilai–Borwein method for $m = 0.6$. The results of the random process outperform the Pronzato–Zhigljavsky Algorithm 3A, and for $m = \frac{1}{3}$ also the Barzilai–Borwein process, but in comparison with the Polak–Ribière method, they are relatively poor.

We can explain the unsatisfactory results of the steepest descent methods (except for the Barzilai–Borwein method for $m = 0.6$) based on the results for $m = \frac{1}{3}$. Evidently, the iterative processes get very quickly to an area characterized by small residuals (for $m < \frac{1}{2}$, this occurs at large $\|\mathsf{x}\|$), and the preceding calculations do not provide data enabling appropriately large step length parameters $\gamma_j$. If, in addition, the behavior of the function $V$ along the gradient direction is concave, then the algorithms in question do not allow for computing new characteristics $\overline{\lambda}$, $\underline{\Lambda}$, and the iterative process consists of very short shifts $\mathsf{x}_{j+1} - \mathsf{x}_j = -\gamma_{j+1}\mathsf{g}_j$, and its convergence is extremely slowed down. The only possibility to break out of such an area is an accumulation of relatively large steps; this may occur only using the random step length parameters, and therefore Algorithm 3 outperforms the remaining two steepest descent processes.

Opposite conditions occur, when the parameter $m > 1$. Figure 8 contains the results for $m = 1.3$ and $m = 1.38$ The modified Barzilai–Borwein algorithm diverged extremely quickly (after a few dozens steps), and the results are not displayed. This divergence may be caused by the fact that the residuals at the points far away from the solution $\underline{\mathsf{x}}$ increase much more than the distances from the solution (the increase in the residuals is documented by the values for Algorithm 3A), and the parameter $\hat{\lambda}^{(j)}$ calculated based on the preceding steps is too small. Therefore, the following shift

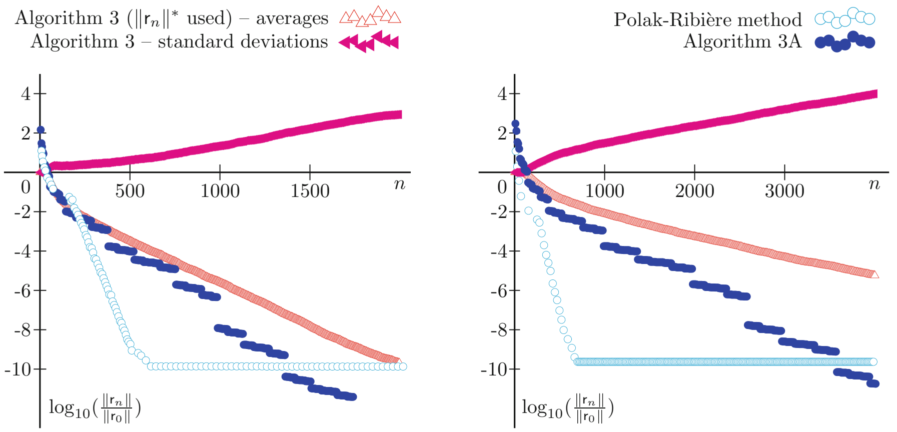

**Fig. 8** Minimization of the function (5.1) for $\alpha = 1$, $m = 1.3$ (*left*) and $m = 1.38$ (*right*) when $\kappa(A) = 90.51$

$-\frac{g_j}{\hat{\lambda}^{(j)}}$ is too long, the process gets to an even more distant point on the "opposite side" of the solution, and the situation reproduces itself.

The algorithms 3 and 3A work with the estimates $\underline{\Lambda}^{(j)}$ which are usually greater than $\overline{\lambda}^{(j)}$ used in the modified Barzilai–Borwein algorithm; therefore, shorter steps are possible, and the processes can converge. Nevertheless, the random process is, cumulating small values of the parameters $l_j$, put in a danger analogous to the one mentioned above; really, many random processes diverged (and some of them very quickly)—at $m = 1.38$ almost half of the processes diverged, while only 8 % of the processes achieved the required relative accuracy $10^{-12}$ during less than 5000 steps.

In any case, the conjugate gradient method gave evidently the best results. The slow convergence of the process generated by Algorithm 3A is probably caused by an overestimate of the parameters $\underline{\Lambda}^{(j)}$ related to the steps of the procedure far away of the solution. This overestimate creates a virtual condition number $\widetilde{\kappa}(A) = \frac{\underline{\Lambda}^{(j)}}{\overline{\lambda}^{(j)}} \gg \kappa(A)$, and the convergence rate (3.15) corresponds to this improper value. In connection with this effect, the creation of an algorithm having shorter memory should be worthy of our attention.

All the outlined phenomena can be observed when solving the problem characterized by $\kappa(A) = 1024$, as shown in Fig. 9; the unfavorable effects are usually stronger than in the case $\kappa(A) = 90.51$. For example, at $m = 1.2$ approximately 82 % of runs of Algorithm 3 diverged, and only less than 7 % of runs reached the needed accuracy during the first 4000 steps. The result for $m = 0.8$ showing that the modified Pronzato–Zhigljavsky Algorithm 3A can give results after sufficiently large number of steps also for $m < 1$ is remarkable.

## 6 Searching for the Global Minimum of a Function

Based on the results presented in the preceding section, we can state that the steepest descent algorithm with random step lengths is relatively reliable in comparison

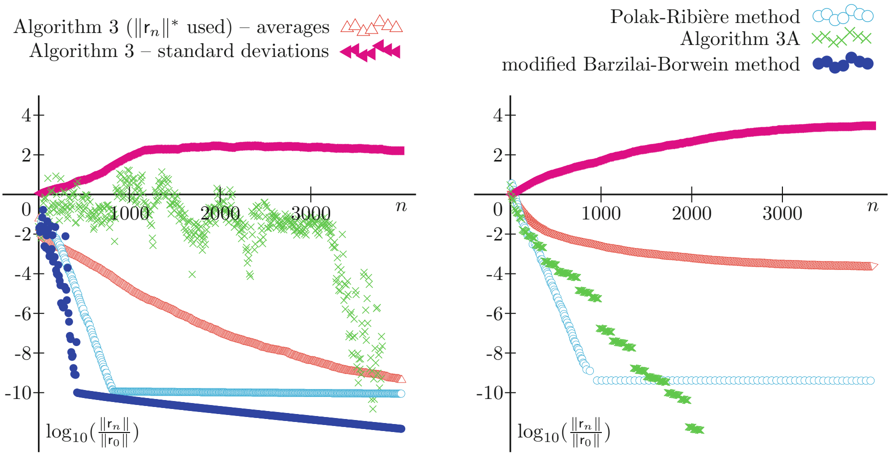

**Fig. 9** Minimization of the function (5.1) at $\kappa(\mathsf{A}) = 1024$ for $\alpha = 0.3$, $m = 0.8$ (*left*) and $m = 1.2$ (*right*)

with other steepest descent methods (the modified Barzilai–Borwein method and Pronzato–Zhigljavsky method give sometimes excellent results and sometimes they are unserviceable). However, the conjugate gradient method is commonly—with a few exceptions—preferable.

Nevertheless, a class of problems solvable using random Algorithm 3 exists. It is the minimization of functions having several local minima. For such functions, the deterministic methods give always one local minimum depending on the initial point $\mathsf{x}_0$; a suitable choice of this initial point is aggravated by the fact that the set of initial points, creating the process which converges to a fixed local minimum of the given function, is usually fractal. The proposed Algorithm 3 combines the relatively fast convergence to the local minimum of a function in its proximity and the possibility to get "almost anywhere" in the space. Thus, the repeated application of the process (a common starting point for all the realized experiments is possible) may result in various local minima of the given function. As follows from the preceding results, the approximately quadratic behavior of the function $V(\mathsf{x})$ for $\|\mathsf{x}\| \to \infty$ increases the applicability of the method, and the well-conditioned Hessian matrices of the minimized function in the local minima increase the convergence rate.

Moreover, the repeated use of the random algorithm suppresses its main disadvantage—the possibility that just the currently running process does not converge sufficiently quickly; the multiple run of the stochastic process compensates a slow convergence of one experiment by a quick convergence of another one.

We tested this idea for the function

$$V(\mathsf{x}) = \frac{1}{2}\langle \mathsf{x}|\mathsf{A}\mathsf{x}\rangle - \langle \mathsf{b}|\mathsf{x}\rangle - \alpha \sum_{i=1}^{k} \cos \frac{x_i - \underline{x}_i}{T} \; ; \qquad (6.1)$$

the matrix $\mathsf{A}$ and vector $\mathsf{b}$ come from the problem solved in Sect. 4 when $\kappa(\mathsf{A}) = 90.51$ and the dimension $M = 176$. The parameters $\alpha$ and $T$ determine the amplitude and the

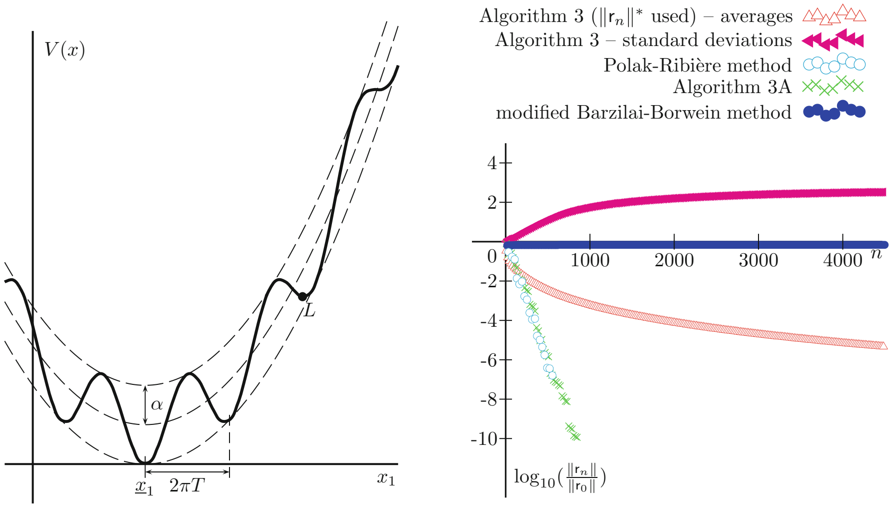

**Fig. 10** Illustration of the function (6.1) in one dimension (*left*) and characteristics of the course of the function (6.1) minimization (*right*)

period of the harmonic waves added to the basic quadratic function (one-dimensional example is displayed in the left part of Fig. 10). The function (6.1) has its only global minimum in the point $\underline{\mathsf{x}}$, which is the solution of problem (1.5), and several other local minima depending on the parameters $k$, $\alpha$, $T$. The added terms in (6.1) significantly change the Hessian matrix in the global minimum of the function $V(\mathsf{x})$—the condition number of the Hessian matrix in the point $\underline{\mathsf{x}}$ was approximately equal to 1024 for the parameters used in our tests. The total amount of the local minima of the tested function was 2197.

The iterative processes were terminated either at $\|\mathsf{r}_n\| < 10^{-10}\|\mathsf{r}_0\|$ or after 10 000 steps. The right part of Fig. 10 documents some characteristics of the results obtained using all the methods tested in the preceding section. The results for the random process are undermined by the fact that the possible values $\|\mathsf{r}_n\| < 10^{-10}\|\mathsf{r}_0\|$ are inaccessible due to the early termination of the process. The modified Barzilai–Borwein process did not converge sufficiently quickly; the explanation is probably the same as in the preceding section for $m = \frac{1}{3}$. The next two processes achieved different local minima, but not the global minimum of the function (6.1). A fact is also noticeable that the conjugate gradient process was terminated before achieving the required accuracy.

The assessment of the random process is possible using the following data: The total number of runs was 19748, including 269 divergent runs and 4165 convergent runs (other runs did not reach any result within the first 10000 steps). 991 local minima of the function $V(\mathsf{x})$ were found, and the global minimum was achieved 68 times. The mean duration of the successfully terminated run was 4827 steps with the standard deviation of 2745 steps.

Frequent convergence to the global minimum of the function (6.1)—more than 1 % of the successfully terminated runs, although the number of all the local minima exceeds 2000—may be explained by the structure of the minimized function (see Fig. 10). The probability that a stochastic process leaves the vicinity of a local minimum different from the global minimum is greater than the probability of this event in the vicinity of the global minimum: For example, the "domain of attraction" of the local minimum $L$ in Fig. 10 on the left of this point is very small, and on the right, the gradients are large, and long steps getting away from the proximity of $L$ are probable.

## 7 Possible Improvements of the Random Procedure

Although the results presented in the preceding two sections are satisfactory, the application of the proposed method to the problems, characterized by ill-conditioned Hessian matrices of the minimized function in their local minima, is limited. The common origin of the most of these problems is described in the part dealing with the experimental results for the quadratic problem with $\kappa(\mathsf{A}) = 1024$. More ways to suppress the probability of occurring the large coordinates $|\beta_i^{(j)}|$ exist, for example:

1. The distribution of the random variable $L$ can be modified with the aim to improve the relation between the standard deviation $\sigma(R_i)$ and the mean value $E(R_i)$ for the eigenvalues $\lambda_i \gg \lambda$. In this case, the asymptotic rate of convergence does not reach the optimal value (3.15), but the proper modification may cause only a small reduction in the asymptotic rate of convergence, accompanied by a significant improvement in the properties of the random procedure in real time.
2. The asymptotic properties of an infinite procedure usually are not affected by a finite number of steps. Therefore, we can use a distribution of the inverse step length different from (3.13), which causes a suppression of the coordinates $|\beta_i^{(j)}|$ related to large eigenvalues $\lambda_i$, during a finite initial phase of the process. Subsequently, the distribution (3.13) may be applied; the asymptotic convergence rate (3.15) is preserved in this case.
3. A similar idea is to use an appropriate sequence of distribution functions $\{F_{L_j}(l)\}_{j\geq 1}$. If the functions $F_{L_j}(l)$ converge for $j \to \infty$ to the distribution function related to the probability density (3.13), then the resulting procedure may provide the asymptotic convergence rate (3.15).[11]
4. The exact step length parameter (1.6) corresponds to the exact inverse step length

$$\hat{l}_j = \frac{\langle \mathsf{r}_{j-1} | \mathsf{A} \mathsf{r}_{j-1} \rangle}{\|\mathsf{r}_{j-1}\|^2} = \frac{\sum_{i=1}^{N} \lambda_i \left(\beta_i^{(j-1)}\right)^2}{\sum_{i=1}^{N} \left(\beta_i^{(j-1)}\right)^2} ,$$

   i.e., $\hat{l}_j$ represents the weighted arithmetic average of the eigenvalues $\lambda_i$ with the weights $(\beta_i^{(j-1)})^2$. If the eigenvalues $\lambda_i$ corresponding to the large coordinates

[11] Algorithm 2 is a special realization of this idea.

$|\beta_i^{(j)}|$ are concentrated in some part of the interval $\langle\lambda, \Lambda\rangle$ around the value $\widetilde{\lambda}$, then $\hat{l}_j \approx \widetilde{\lambda}$ and if in the $j$th step the value $\hat{l}_j$ is used, then for $\lambda_i \approx \widetilde{\lambda} \approx \hat{l}_j$

$$\left|\beta_i^{(j)}\right| = \frac{|\hat{l}_j - \lambda_i|}{\hat{l}_j}\left|\beta_i^{(j-1)}\right| \ll \left|\beta_i^{(j-1)}\right| ,$$

so the extremely large coordinates $|\beta_i|$ decrease significantly.
The described technique loses its efficiency if the eigenvalues corresponding to the extraordinary large coordinates $|\beta_i^{(j-1)}|$ are located in distant parts of the interval $\langle\lambda, \Lambda\rangle$, but we can suppose that the use of the exact value $\hat{l}_j$ in such cases at least causes no damage.
Therefore, we can suggest sporadic application of the exact step length during the iterative process. The exact description of the properties of this semi-random process is complicated, but—according to the performed tests—the results are favorable. In this context, we can note that the Barzilai–Borwein method and its forms [8], using older exact step lengths, may be considered as particular forms of some semi-random process of this kind, as reported in [9].[12]
The sporadic use of the exact step length may be applied when the modified distribution of the step length parameter, as described in proposition 1, is used. Then, usually, the function $\widehat{E}_R(x)$ is decreasing, $\overline{E}_\phi = \widehat{E}_R(\lambda)$, and the large coordinates $|\beta_i^{(j)}|$ are typical for the eigenvalues $\lambda_i \approx \lambda$ due to (8.4). In this case, the application of the exact step length causes a reduction in the mentioned large values $|\beta_i^{(j)}|$ and, consequently, improves the asymptotic convergence rate, according to (8.14).

5. We can avoid extremely large values $|\beta_i^{(j-1)}|$ when the random process with the optimal distribution of the inverse step length parameter (3.13) is adapted to the $Q$-convergent random relaxed method suggested in Part 2 of [15], allowing only the values $l_j > \frac{\hat{l}_j}{2}$. This method may be considered a specific implementation of the general proposition 3.

We can present the results obtained for some realizations of the propositions 1, 4. As for the last mentioned suggestion, the calculation of the actual exact step length is complicated. However, as shown in [9], the change in the step length parameters' order does not affect the final result, if the minimized function is quadratic (it follows from (1.9) due to the commutativity of the matrices in the product); for other functions, this holds only approximately, but the deviation is usually small. Therefore, we can substitute the actual exact step length parameter by the step length parameter which had to be exact in the preceding step, analogously to the Barzilai–Borwein algorithm; the "substituting" exact step length parameter is then given by the relation (5.2). The resulting iterative $\mathsf{x}_{j+1}$ is then the same (for quadratic functions) as when using the exact step length in the $j$th step and the random value $l_j$ in the $(j+1)$th step.

[12] In fact, the attempt to interpret the Barzilai–Borwein method in a nonstandard way was the origin of the work resulting in this article.

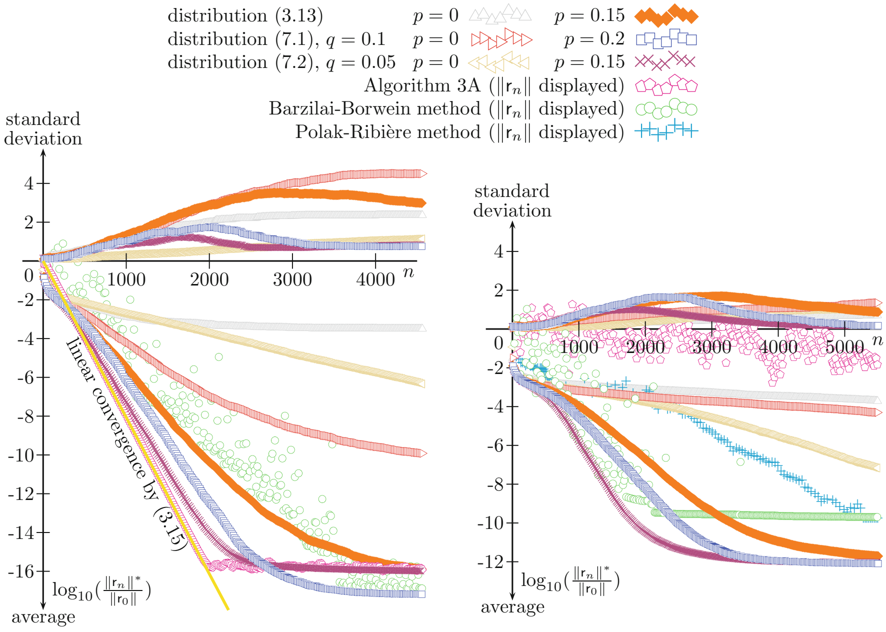

**Fig. 11** Solution of the Eq. (1.5) at $\kappa(\mathsf{A}) = 1.16 \times 10^4$ (*left*) and minimization of the function (5.1) at $\kappa(\mathsf{A}) = 1.16 \times 10^4$ for $\alpha = 0.3$, $m = 0.8$ (*right*) using improved stochastic processes

We tested three kinds of the distribution functions—the basic distribution function (8.53), the distribution with the “added shortest step lengths”

$$F_L(l) = \begin{cases} \frac{1-q}{\pi} \arccos \frac{\Lambda+\lambda-2l}{\Lambda-\lambda} & \text{for } l \in \langle \lambda, \Lambda) \,, \\ 1 & \text{for } l = \Lambda \,, \end{cases} \quad q \in \langle 0, 1) \,, \tag{7.1}$$

giving the inverse step length parameter $l_j = \underline{\Lambda}^{(j)}$ with probability $q$, and the distribution with “suppressed largest step lengths”

$$F_L(l) = \frac{\arccos \frac{\lambda+\Lambda\cos(q\pi)-l(1+\cos(q\pi))}{\Lambda-\lambda} - q\pi}{(1-q)\pi} \,, \quad q \in \langle 0, 1) \,. \tag{7.2}$$

We further tested these three kinds of distribution functions using the previous exact step length parameter (5.2) with probability $p \in \langle 0, 1)$—thus, the distribution function $F_L$ was applied with probability $1 - p$. Some results are displayed in Fig. 11. The characteristics of the solution of the quadratic problem (1.5) with the dimension $M = 2288$ and $\kappa(\mathsf{A}) = 1.16 \times 10^4$ are presented in the left part of Fig. 11. The Barzilai–Borwein method and the simplified Pronzato–Zhigljavsky Algorithm 3A were also tested.

The isolated modifications of the basic random process, realized by employing the distributions (7.1) and (7.2), improve the results noticeably, but still unsatisfactorily; in this context, the application of the distribution (7.2) is the poorest one. The application of the step length parameter (5.2) with suitable probability $p$ improves the results significantly, and the synergy of both the tested modifications produces processes behaving very well, namely the process generated by the distribution (7.2)—its average convergence rate is very close to the value (3.15).

The only advantage of Barzilai–Borwein algorithm is its high accuracy—after $\sim$ 3400 steps, it is surpassed only by the process generated by the distribution (7.1) with previous exact step lengths applied with probability $p = 0.2$. The fastest calculation was achieved using the simplified Pronzato–Zhigljavsky process—the convergence rate corresponds almost exactly to the theoretical value (3.15).

We applied the improved distribution also to the problem solved in Sect. 5. The testing parameters of the function (5.1) were $\alpha = 0.3$, $m = 0.8$, when $\kappa(\mathsf{A}) = 1.16 \times 10^4$. Due to the positive experience with the Barzilai–Borwein procedure for $m \in (\frac{1}{2}, 1)$, we can expect an improvement of the results mainly in the processes employing the previous exact step lengths. Really, all the processes containing the casual application of the step length parameter (5.2) outperformed the Polak–Ribière method; the relevance of this result is backed by the small standard deviations of the values $\|\mathsf{r}_n\|^*$ up. The best results were gained applying the distribution (7.2); only the results of the Barzilai–Borwein process can compete with them during the initial part of the procedure, but the Barzilai–Borwein process did not achieve the required accuracy.

The procedure generated by the simplified Pronzato–Zhigljavsky algorithm did not converge during the first 8500 steps, analogously to the results obtained in Sect. 5.

The use of the step length parameter (5.2) for the minimization of the function (5.1) when $m > 1$ could elevate the tendency of the process to diverge, analogously to the behavior of the Barzilai–Borwein process in this case (see the end of Sect. 5). This assumption was vindicated in general, but the process generated by the distribution (7.2) supplemented by a careful use of the previous exact step length still gave acceptable results. Concretely, by the minimization of function (5.1) with the parameters $\alpha = 0.3$, $m = 1.2$, and $\kappa(\mathsf{A}) = 1024$ using the distribution (7.2) for $q = 0.04$ and the application of previous exact step length with probability $p = 0.04$, only 15 % of the processes diverged, and the average values of $\|\mathsf{r}_n\|^*$ were only little inferior to the results for the simplified Pronzato–Zhigljavsky Algorithm 3A (displayed in the right part of Fig. 9).

The comparison of the improved methods applied to the minimization of the function (6.1) is possible studying the results in Table 1. All the variants were tested during the same time on the same computer as the only running process. The most important value in Table 1 is the amount of successfully terminated runs. The best results were achieved using the distribution (7.2) with careful application of the previous exact step length; however, the isolated distribution (7.2) also gives good results (the total amount of the found local minima is even greater in this case). What is surprising is the apparent worsening of the results by applying the previous exact step lengths to the distribution (7.1). Interesting information is also given by the geometric means of the remaining residual norms —for the distribution (7.2), the tendency of the processes to converge is evident.

**Table 1** Results of the minimization of the function (6.1)

| Distribution | $p$ | Runs | | | Local minima found | Global minimum achieved | $\log_{10}\frac{\lVert \mathsf{r}_{10000}\rVert}{\lVert \mathsf{r}_0\rVert}$ (unterminated runs) |
|---|---|---|---|---|---|---|---|
| | | Total amount | Successfully terminated | Divergent | | | |
| (3.13) | 0 | 19,748 | **4165** | 269 | 991 | 68 | 4.37 |
| | 0.2 | 20,449 | **6505** | 0 | 1008 | 58 | −0.45 |
| (7.1), $q = 0.1$ | 0 | 28,651 | **17,468** | 141 | 1352 | 205 | 4.46 |
| | 0.1 | 22,994 | **9536** | 0 | 1117 | 99 | −0.31 |
| | 0.2 | 23,159 | **11,041** | 0 | 1 137 | 80 | −0.64 |
| (7.2), $q = 0.04$ | 0 | 70,435 | **70,391** | 0 | 1643 | 449 | −6.99 |
| | 0.05 | 96,724 | **96,722** | 0 | 1630 | 652 | −8.85 |
| | 0.1 | 85,161 | **85,152** | 0 | 1586 | 631 | −6.78 |

Most important values are given in bold

## 8 Conclusions

The application of the random steepest descent method suggested in this paper is possible in various cases, although it is not destined for universal use, and many problems are solvable by other methods more comfortably and quickly. Its use for the minimization of non-quadratic functions is possible, namely when the minimized function behaves similarly to $\lVert \mathsf{x}\rVert^m$ for an $m \in (1, 2)$ at $\lVert \mathsf{x}\rVert \to \infty$. The repeated use of the random algorithm can successfully serve for the search for the global minimum of the functions with several local minima. Another, in this paper not studied problem, may be the minimization of a quadratic function of more variables on the constrained boxes (see, e.g., [4] and the references therein). The use of non-optimal distributions accompanied by occasional application of the previous exact step length can significantly improve the practical applicability of the method.

## Appendix 1: Proof of Theorem 1

First, we prove an important property of the points of the second type:

**Lemma 3** *If $\hat{x} \in \langle \lambda, \Lambda\rangle$ is a point of the second type relative to the measure $\phi$, then*

$$\lim_{\delta\to 0^+} \int\limits_{\langle\lambda,\Lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1-\frac{\hat{x}}{l}\right| \mathrm{d}\phi(l) = -\infty\,.$$

*Proof* Since $\phi(\{\hat{x}\}) = 0$, it holds

$$\widehat{E}_R(\hat{x}) = \int\limits_{\langle\lambda,\Lambda\rangle} \ln\left|1 - \frac{\hat{x}}{l}\right| \mathrm{d}\phi(l) = \int\limits_{\langle\lambda,\hat{x})\cup(\hat{x},\Lambda\rangle} \ln\left|1 - \frac{\hat{x}}{l}\right| \mathrm{d}\phi(l)$$

$$= \lim_{\delta\to 0^+} \int\limits_{\langle\lambda,\Lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1 - \frac{\hat{x}}{l}\right| \mathrm{d}\phi(l)\,. \tag{8.1}$$

The value $\widehat{E}_R(\hat{x})$ is undefined; therefore, the limit on the right-hand side of (8.1) is either infinite or it does not exist at all. The function

$$g(\delta) = \int\limits_{\langle\lambda,\Lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1 - \frac{\hat{x}}{l}\right| \mathrm{d}\phi(l)$$

is non-decreasing on $(0, \frac{\hat{x}}{2})$, as follows from

$$g(\delta_1) - g(\delta_2) = \int\limits_{\langle\lambda,\Lambda\rangle\cap[(\hat{x}-\delta_2,\hat{x}-\delta_1\rangle\cup\langle\hat{x}+\delta_1,\hat{x}+\delta_2)]} \ln\left|1 - \frac{\hat{x}}{l}\right| \mathrm{d}\phi(l) \le 0 \tag{8.2}$$

for $0 < \delta_1 < \delta_2 < \frac{\hat{x}}{2}$, because the argument of the integrated logarithm is

$$\left|1 - \frac{\hat{x}}{l}\right| = \frac{|l - \hat{x}|}{l} < \frac{\frac{\hat{x}}{2}}{\frac{\hat{x}}{2}} = 1$$

and a non-positive function is integrated in (8.2). The limit of the non-decreasing function $g(\delta)$ in (8.1) exists, and since it is not finite according to the assumptions, it must be equal to $-\infty$. □

Now, we can prove the particular statements of Theorem 1. The convergence of the process to the solution is equivalent to the convergence of the residuals $\mathsf{r}_n$ to zero, so we will deal with the sequence of the residuals.

1. We will observe the coordinates $\beta_i^{(j)}$ of the residual $\mathsf{r}_j$ according to (2.1). The set $\mathcal{S}_3^{(\phi)}(\mathsf{A}, \mathsf{b})$ is finite; therefore, the value

   $$s_{\phi,\mathsf{A},\mathsf{b}} = \min_{\lambda_i\in\mathcal{S}_3^{(\phi)}(\mathsf{A},\mathsf{b})} \phi\left(\{\lambda_i\}\right) > 0$$

   exists. This result implies the inequality

   $$\Pr\left(l_j \ne \lambda_i\right) \le 1 - s_{\phi,\mathsf{A},\mathsf{b}}$$

   for any eigenvalue $\lambda_i \in \mathcal{S}(\mathsf{A}, \mathsf{b})$. If at least one value $l_j = \lambda_i$ for $j = 1, 2, \ldots, n$, then as a consequence of (2.2), $\beta_i^{(n)} = 0$. Thus, the independence of the values $l_j$ for different $j$ leads to

$$\Pr\left(\beta_i^{(n)} \neq 0\right) = \Pr\left(\bigcap_{j=1}^{n}\left[l_j \neq \lambda_i\right]\right) = \prod_{j=1}^{n}\Pr\left(l_j \neq \lambda_i\right) \leq (1 - s_{\phi,\mathsf{A},\mathsf{b}})^n \,. \quad (8.3)$$

The probability that the residual $\mathsf{r}_n \neq 0$ is the same as the probability of at least one coordinate $\beta_i^{(n)}$ of the residual $\mathsf{r}_n$ being nonzero, so

$$\Pr\left(\mathsf{r}_n \neq 0\right) = \Pr\left(\bigcup_{i=1}^{N}\left[\beta_i^{(n)} \neq 0\right]\right) \leq \sum_{i=1}^{N}\Pr\left(\beta_i^{(n)} \neq 0\right) \leq N(1 - s_{\phi,\mathsf{A},\mathsf{b}})^n$$

, and this probability converges to zero at $n \to \infty$.

2. In the case $\mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b}) \neq \emptyset$, we get, according to (2.3) and (2.4) and due to the validity of the strong law of large numbers,

$$\lim_{n\to\infty}\frac{1}{n}\ln\frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|} = \lim_{n\to\infty}\frac{1}{n}\sum_{j=1}^{n}\widetilde{R}_i^{(j)} = \widehat{E}_R(\lambda_i) \quad (8.4)$$

for any $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$ with probability one.[13]

(a) Let $E_{\phi,\mathsf{A},\mathsf{b}} > 0$. Then, there exists an eigenvalue $\lambda_{\hat{\imath}} \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$ such that $\widehat{E}_R(\lambda_{\hat{\imath}}) = E_{\phi,\mathsf{A},\mathsf{b}} > 0$. We express the probability that there exists an infinite bounded subsequence of coordinates $\{\beta_{\hat{\imath}}^{(n_k)}\}_{k>0}$. This event occurs, when for some $C \in \mathcal{R}$ and all $k \in \mathcal{N}$

$$\frac{1}{n_k}\ln\frac{|\beta_{\hat{\imath}}^{(n_k)}|}{|\beta_{\hat{\imath}}^{(0)}|} < \frac{1}{n_k}\ln\frac{C}{|\beta_{\hat{\imath}}^{(0)}|}\;; \quad (8.5)$$

The necessary condition to satisfy (8.5) is

$$\limsup_{k\to\infty}\frac{1}{n_k}\ln\frac{|\beta_{\hat{\imath}}^{(n_k)}|}{|\beta_{\hat{\imath}}^{(0)}|} \leq \limsup_{k\to\infty}\frac{1}{n_k}\ln\frac{C}{|\beta_{\hat{\imath}}^{(0)}|} = 0\,. \quad (8.6)$$

The relation (8.4) applied for $\lambda_{\hat{\imath}}$

$$\lim_{n\to\infty}\frac{1}{n}\ln\frac{|\beta_{\hat{\imath}}^{(n)}|}{|\beta_{\hat{\imath}}^{(0)}|} = \widehat{E}_R(\lambda_{\hat{\imath}}) > 0$$

characterizes an event, which is incompatible with the event characterized by (8.6). Therefore, the probability of the existence of an infinite bounded subsequence of coordinates $\{\beta_{\hat{\imath}}^{(n_k)}\}_{k>0}$ is

[13] The mean value $\widehat{E}_R(\lambda_i)$ exists by the definition of the point of the first type relative to the measure $\phi$ and the condition $\beta_i^{(0)} \neq 0$ is included in the assumption $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$.

$$\Pr\left(\forall k \in \mathcal{N} : |\beta_{\hat{\imath}}^{(n_k)}| < C\right) \leq \Pr\left(\limsup_{k\to\infty} \frac{1}{n_k} \ln \frac{|\beta_{\hat{\imath}}^{(n_k)}|}{|\beta_{\hat{\imath}}^{(0)}|} \leq 0\right) \leq$$

$$\leq 1 - \Pr\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{|\beta_{\hat{\imath}}^{(n)}|}{|\beta_{\hat{\imath}}^{(0)}|} > 0\right) = 0,$$

and it implies

$$\lim_{n\to\infty} |\beta_{\hat{\imath}}^{(n)}| = \infty$$

with probability one. Since

$$\|\mathsf{r}_n\| \geq \max_{\lambda_i \in \mathcal{S}^{(\phi)}(\mathsf{A},\mathsf{b})} |\beta_i^{(n)}| \,,$$

the sequence of residuals and consequently the sequence of the iterations are almost surely divergent.

(b) We prove the convergence of the process in both cases 2b) and 3) together. We will study the convergence of the sequence $\{\beta_i^{(n)}\}_{n\geq 0}$ for general $i = 1, 2, \ldots, N$ such that $\lambda_i \in \mathcal{S}^{(\phi)}(\mathsf{A}, \mathsf{b})$.
If $\lambda_i \in \mathcal{S}_3^{(\phi)}(\mathsf{A}, \mathsf{b})$, the relation (8.3) implies

$$\Pr\left(\lim_{n\to\infty} \beta_i^{(n)} = 0\right) = 1 \,. \tag{8.7}$$

For $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$, the mean value $\widehat{E}_R(\lambda_i)$ exists. Analogously to (8.4)

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|} = \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \widetilde{R}_i^{(j)} = \widehat{E}_R(\lambda_i) \leq E_{\phi,\mathsf{A},\mathsf{b}} < 0 \tag{8.8}$$

with probability one. Let us suppose

$$\lim_{n\to\infty} \beta_i^{(n)} \neq 0 \,.$$

We formulate this assumption now in the standard way: There exists $\varepsilon > 0$ and some infinite subsequence $\{n_k\}_{k\geq 1} \subset \mathcal{N}$ such that $|\beta_i^{(n_k)}| \geq \varepsilon$ for all $k \in \mathcal{N}$. The technique applied above gives

$$\Pr\left(\lim_{n\to\infty} \beta_i^{(n)} \neq 0\right) = \Pr\left(\forall k \in \mathcal{N} : |\beta_i^{(n_k)}| \geq \varepsilon\right)$$

$$\leq \Pr\left(\liminf_{k\to\infty} \frac{1}{n_k} \ln \frac{|\beta_i^{(n_k)}|}{|\beta_i^{(0)}|} \geq 0\right)$$

$$\leq 1 - \Pr\left(\lim_{n\to\infty} \frac{1}{n} \ln \frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|} < 0\right) = 0 \,.$$

Thus, the relation (8.7) holds for $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$ too.

Let $\lambda_i \in \mathcal{S}_2^{(\phi)}(\mathsf{A}, \mathsf{b})$. Then, the statement of Lemma 3. implies for any $K < 0$ the existence of such $\delta_K > 0$ that

$$\int\limits_{\langle\lambda,\Lambda\rangle\setminus(\lambda_i-\delta_K,\lambda_i+\delta_K)} \ln\left|1-\frac{\lambda_i}{l}\right| \mathrm{d}\phi(l) \le K\ , \tag{8.9}$$

this integral being finite.

We define the random variable

$$Q_i^{(j)} = \begin{cases} \ln\left|1-\frac{\lambda_i}{l_j}\right| & \text{for} \quad l_j \in \langle\lambda, \Lambda\rangle \setminus (\lambda_i - \delta_K, \lambda_i + \delta_K)\ , \\ 0 & \text{for} \quad l_j \in \langle\lambda, \Lambda\rangle \cap (\lambda_i - \delta_K, \lambda_i + \delta_K) \end{cases}$$

The variable $Q_i^{(j)}$ has its mean value $E(Q_i) \le K$ due to (8.9), and at the given value $l_j$ of the random variable $L$ always holds $\widetilde{R}_i^{(j)} \le Q_i^{(j)}$. Therefore,

$$\limsup_{n\to\infty} \frac{1}{n} \ln \frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|} = \limsup_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \widetilde{R}_i^{(j)} \le \lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \widetilde{Q}_i^{(j)} = E(Q_i) \le K < 0 \tag{8.10}$$

almost surely, and the relation (8.7) may be proved using the same method as for $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$.

We define the vectors

$$\mathsf{r}_n^{(k)} = \sum_{i=1}^{k} \beta_i^{(n)} \mathsf{e}_i$$

for $k = 0, 1, \ldots, N$; thus, $\mathsf{r}_n^{(N)} = \mathsf{r}_n$, $\mathsf{r}_n^{(0)} = \mathsf{0}$. We express the probabilities

$$\Pr\left(\lim_{n\to\infty} \|\mathsf{r}_n^{(k)} - \mathsf{r}_n^{(k-1)}\| = 0\right) = \Pr\left(\lim_{n\to\infty} |\beta_k^{(n)}| = 0\right) = 1$$

as a consequence of (8.7) for any $k = 1, 2, \ldots, N$. It implies

$$\lim_{n\to\infty} \mathsf{r}_n = \lim_{n\to\infty} \mathsf{r}_n^{(N)} = \lim_{n\to\infty} \mathsf{r}_n^{(N-1)} = \cdots = \lim_{n\to\infty} \mathsf{r}_n^{(0)} = 0$$

with probability one.

As for the rate of convergence, we express the values of $\lim\limits_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|}$. If $\beta_i^{(0)} = 0$, all the members of the considered sequence are zero, otherwise

$$\lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} = \lim_{n\to\infty} \sqrt[n]{\frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|}}\ . \tag{8.11}$$

If $\lambda_i \in \mathcal{S}_3^{(\phi)}(\mathsf{A}, \mathsf{b})$, then the relation (8.3) implies the limit (8.11) is almost surely equal to zero. For the values $\lambda_i \in \mathcal{S}_2^{(\phi)}(\mathsf{A}, \mathsf{b})$, we get, based on (8.10),

$$\limsup_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} = e^{\limsup_{n\to\infty} \frac{1}{n} \ln \frac{|\beta_i^{(n)}|}{|\beta_i^{(0)}|}} \le e^K \tag{8.12}$$

with probability one for arbitrary $K < 0$; therefore, the limit (8.11) is almost surely zero in this case. Analogously, the relation (8.8) implies for $\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b})$

$$\lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} = e^{\widehat{E}_R(\lambda_i)} .$$

with probability one.
Now, we can evaluate

$$\begin{aligned} \lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} &= \lim_{n\to\infty} \sqrt[2n]{\sum_{i=1}^{N} \left(\beta_i^{(n)}\right)^2} \\ &\ge \lim_{n\to\infty} \sqrt[2n]{\max_{i\le N} \left(\beta_i^{(n)}\right)^2} = \max_{i\le N} \lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} ; \end{aligned} \tag{8.13}$$

the last equality is valid, if all the limits on the right-hand side of (8.13) exist. However, their almost sure existence is confirmed by the preceding results. Analogously

$$\lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} \le \lim_{n\to\infty} \sqrt[2n]{N \cdot \max_{i\le N} \left(\beta_i^{(n)}\right)^2} = \max_{i\le N} \lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|}$$

with probability one, thus

$$\lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} = \max_{i\le N} \lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} \tag{8.14}$$

almost surely.
In the case $\mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b}) \neq \emptyset$ is by definition (2.5)

$$\lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} = \max_{\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A},\mathsf{b})} e^{\widehat{E}_R(\lambda_i)} = e^{E_{\phi,\mathsf{A},\mathsf{b}}} \tag{8.15}$$

with unit probability. The logarithm of the relation (8.15) results in (2.6).

(c) In the case $E_{\phi,\mathsf{A},\mathsf{b}} = 0$, the technique used in the proofs of the statements 2a) and 2b) cannot be applied; nevertheless, if we admit the possibility of the convergence of the process to the solution, then, according to (8.14),

$$\lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} = \max_{\lambda_i \in \mathcal{S}_1^{(\phi)}(\mathsf{A},\mathsf{b})} e^{\widehat{E}_R(\lambda_i)} = 1,$$

and the convergence is almost surely $R$-sublinear.

3. When $\mathcal{S}_1^{(\phi)}(\mathsf{A}, \mathsf{b}) = \emptyset$, the relations (8.3), (8.12) imply

$$\lim_{n\to\infty} \sqrt[n]{|\beta_i^{(n)}|} = 0$$

for all the eigenvalues $\lambda_i \in \mathcal{S}(\mathsf{A}, \mathsf{b})$. Therefore, according to (8.14),

$$\lim_{n\to\infty} \sqrt[n]{\frac{\|\mathsf{r}_n\|}{\|\mathsf{r}_0\|}} = \max_{i\le N}(0) = 0$$

almost surely, which proves statement 3.

## Appendix 2: Proofs of Lemma 1 and Lemma 2

The integrals $I_c$ and $I_d$ exist for any $q \in \langle\lambda, \Lambda\rangle$, as both the singularities of the integrands are integrable. After the standard substitution

$$x = \frac{\Lambda+\lambda}{2} + \frac{\Lambda-\lambda}{2}\sin\alpha\ , \quad \alpha \in \langle -\frac{\pi}{2}, \frac{\pi}{2}\rangle, \tag{8.16}$$

we get

$$\begin{aligned} I_c &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln\left|\frac{\frac{\Lambda+\lambda-2q}{\Lambda-\lambda}+\sin\alpha}{\frac{\Lambda+\lambda}{\Lambda-\lambda}+\sin\alpha}\right| \mathrm{d}\alpha \\ &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln\left|\frac{\Lambda+\lambda-2q}{\Lambda-\lambda}+\sin\alpha\right| \mathrm{d}\alpha - \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln\left|\frac{\Lambda+\lambda}{\Lambda-\lambda}+\sin\alpha\right| \mathrm{d}\alpha\ . \end{aligned} \tag{8.17}$$

The second integral on the right-hand side of (8.17) is finite and does not depend on the parameter $q$, so we can neglect it. The first integral (we will denote it as $I_{c1}$) can be expressed using the variable $\varphi$ (see Fig. 12) defined by:

$$v(\varphi)\cos\varphi = \frac{\Lambda-\lambda}{2}\cos\alpha\ , \tag{8.18}$$

$$v(\varphi)\sin\varphi = \frac{\Lambda+\lambda}{2} - q + \frac{\Lambda-\lambda}{2}\sin\alpha\ . \tag{8.19}$$

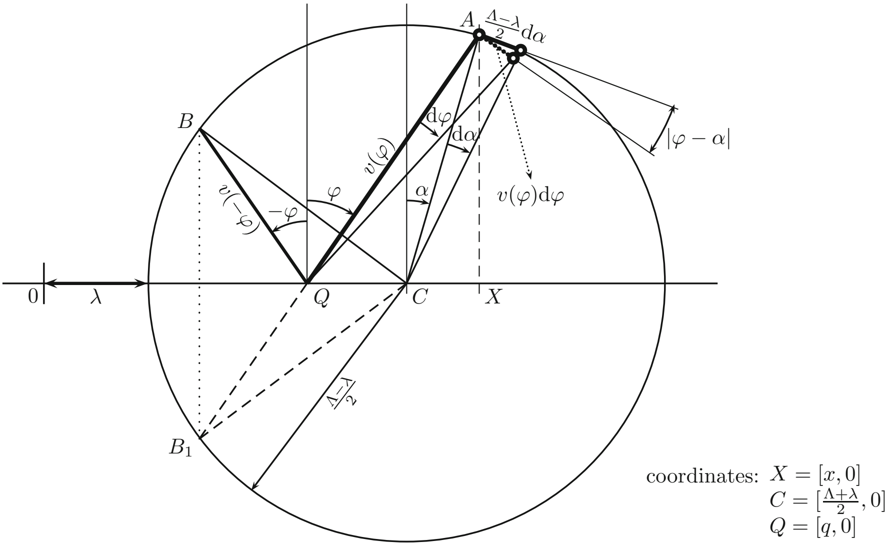
A
$\frac{\Lambda-\lambda}{2}\mathrm{d}\alpha$
$|\varphi-\alpha|$
B
$\mathrm{d}\varphi$
$\mathrm{d}\alpha$
$v(\varphi)$
$v(-\varphi)$
$\varphi$
$-\varphi$
$\alpha$
$v(\varphi)\mathrm{d}\varphi$
0
$\lambda$
Q
C
X
$\frac{\Lambda-\lambda}{2}$
$B_1$
coordinates: $X = [x, 0]$
$C = [\frac{\Lambda+\lambda}{2}, 0]$
$Q = [q, 0]$

**Fig. 12** Geometric interpretation of the substitution (8.18), (8.19)

As follows from the Fig. 12, the mapping $\varphi \longleftrightarrow \alpha$ is unequivocal for $\alpha \in \langle -\frac{\pi}{2}, \frac{\pi}{2} \rangle$, $\varphi \in \langle -\frac{\pi}{2}, \frac{\pi}{2} \rangle$ at $q \in (\lambda, \Lambda)$; for $q = \lambda$, the unequivocalness is ensured for the mapping $\alpha \in (-\frac{\pi}{2}, \frac{\pi}{2} \rangle \longleftrightarrow \varphi \in (0, \frac{\pi}{2} \rangle$ and $\alpha = -\frac{\pi}{2} \longleftrightarrow \varphi = 0$; the analogous unequivocal mapping exists for $q = \Lambda$. Therefore, we can define the functions $\varphi(\alpha)$ and $\alpha(\varphi)$ and use them as necessary.

The integral $I_{c1}$ may be expressed as

$$I_{c1} = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln \left| \frac{2v(\varphi(\alpha)) \sin \varphi(\alpha)}{\Lambda - \lambda} \right| \mathrm{d}\alpha = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln |\cos \alpha \tan \varphi(\alpha)| \, \mathrm{d}\alpha$$
$$= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln |\cos \alpha| \, \mathrm{d}\alpha + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \ln |\tan \varphi(\alpha)| \, \mathrm{d}\alpha \ . \tag{8.20}$$

The first integral on the right-hand side of (8.20) is finite and does not depend on the parameter $q$ again, so we neglect it. The remaining integral (we will further denote it as $I_{c2}$) can depend on the parameter $q$ through the function $\varphi(\alpha)$. We can express it by use of the variable $\varphi$ as the integration variable. The geometric interpretation of the substitution $\alpha \to \varphi$ gives simply (see Fig. 12—the angle $\angle QAC = |\varphi - \alpha|$)

$$v(\varphi)\mathrm{d}\varphi = \frac{\Lambda - \lambda}{2} \cos(\varphi - \alpha)\mathrm{d}\alpha \ . \tag{8.21}$$

We must distinguish three situations now. For $q \in (\lambda, \Lambda)$,

$$I_{c2} = \int\limits_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{2v(\varphi)\ln|\tan\varphi|}{(\Lambda-\lambda)\cos[\varphi-\alpha(\varphi)]}\mathrm{d}\varphi$$

$$= \int\limits_{0}^{\frac{\pi}{2}} \left[\frac{2v(\varphi)\ln\tan\varphi}{(\Lambda-\lambda)\cos[\varphi-\alpha(\varphi)]} + \frac{2v(-\varphi)\ln\tan\varphi}{(\Lambda-\lambda)\cos[-\varphi-\alpha(-\varphi)]}\right]\mathrm{d}\varphi.$$

From Fig. 12, it follows that the triangles $QCB$ and $QCB_1$ are identical. The points $B_1$, $Q$, and $A$ lie on the same straight line, and the triangle $CAB_1$ is isosceles; therefore, the angles $\angle QAC = \angle QBC$ and

$$\cos[-\varphi-\alpha(-\varphi)] = \cos[\varphi-\alpha(\varphi)] \tag{8.22}$$

Moreover, the base of the triangle $B_1CA$ has the length

$$AB_1 = v(\varphi) + v(-\varphi) = (\Lambda-\lambda)\cos(\varphi-\alpha)\ , \tag{8.23}$$

thus

$$I_{c2} = \int\limits_{0}^{\frac{\pi}{2}} \frac{[2v(\varphi)+2v(-\varphi)]\ln\tan\varphi}{(\Lambda-\lambda)\cos[\varphi-\alpha(\varphi)]}\mathrm{d}\varphi = \int\limits_{0}^{\frac{\pi}{2}} 2\ln\tan\varphi\,\mathrm{d}\varphi = 0\ ,$$

and the integral $I_c$ does not depend on the parameter $q$.

If $q = \lambda$, then

$$I_{c2} = \int\limits_{0}^{\frac{\pi}{2}} \frac{2v(\varphi)\ln\tan\varphi}{(\Lambda-\lambda)\cos[\varphi-\alpha(\varphi)]}\mathrm{d}\varphi\ .$$

In this case, the point $Q$ coincides with the point $B_1$, and we can express directly the value

$$v(\varphi) = AB_1 = (\Lambda-\lambda)\cos(\varphi-\alpha)\ ,$$

and we get

$$I_{c2} = \int\limits_{0}^{\frac{\pi}{2}} 2\ln\tan\varphi\mathrm{d}\varphi = 0$$

again. We can finish the proof of Lemma 1 for $q = \Lambda$ analogously.

As for the integral (3.2), we get after the implementation of the variable $\alpha$ according to (8.16)

$$I_d = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\ln \frac{\left|q - \frac{\Lambda+\lambda}{2} - \frac{\Lambda-\lambda}{2}\sin\alpha\right|}{q}}{\frac{\Lambda+\lambda}{2} + \frac{\Lambda-\lambda}{2}\sin\alpha}\, \mathrm{d}\alpha \,,$$

and the expression inside the absolute value in the numerator may be substituted using the variable $\varphi$ defined by (8.18), (8.19)

$$I_d = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\ln \frac{|(\Lambda-\lambda)\cos\alpha\tan\varphi(\alpha)|}{2q}}{\frac{\Lambda+\lambda}{2} + \frac{\Lambda-\lambda}{2}\sin\alpha}\, \mathrm{d}\alpha = \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\ln \frac{(\Lambda-\lambda)\cos\alpha}{2}}{\frac{\Lambda+\lambda}{2} + \frac{\Lambda-\lambda}{2}\sin\alpha}\, \mathrm{d}\alpha + \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{\ln \frac{|\tan\varphi(\alpha)|}{q}}{\frac{\Lambda+\lambda}{2} + \frac{\Lambda-\lambda}{2}\sin\alpha}\, \mathrm{d}\alpha \,.$$

The first integral on the right-hand side is finite and evidently does not depend on the parameter $q$. Therefore, we will deal with the second integral (denoted as $I_{d1}$) only. We change over the integration variable to $\varphi$. For $q \in (\lambda, \Lambda)$, we get, according to (8.21),

$$\begin{aligned} I_{d1} &= \int_{-\frac{\pi}{2}}^{\frac{\pi}{2}} \frac{2v(\varphi)\ln\frac{|\tan\varphi|}{q}}{(\Lambda-\lambda)\,[q+v(\varphi)\sin\varphi]\cos[\varphi-\alpha(\varphi)]}\,\mathrm{d}\varphi \\ &= \int_0^{\frac{\pi}{2}} \left[ \frac{2v(\varphi)\ln\frac{\tan\varphi}{q}}{(\Lambda-\lambda)\,[q+v(\varphi)\sin\varphi]\cos[\varphi-\alpha(\varphi)]} \right. \\ &\qquad \left. + \frac{2v(-\varphi)\ln\frac{\tan\varphi}{q}}{(\Lambda-\lambda)\,[q-v(-\varphi)\sin\varphi]\cos[\varphi-\alpha(\varphi)]} \right] \mathrm{d}\varphi \\ &= \int_0^{\frac{\pi}{2}} \frac{2q\ln\frac{\tan\varphi}{q}}{q^2+q[v(\varphi)-v(-\varphi)]\sin\varphi - v(\varphi)v(-\varphi)\sin^2\varphi}\,\mathrm{d}\varphi \,. \end{aligned} \tag{8.24}$$

(We have used the results (8.22), (8.23) during the calculations.)

If we use the law of cosines for the triangle $QCA$ in Fig. 12, we get the relation

$$v(\varphi)^2 - (\Lambda+\lambda-2q)v(\varphi)\sin\varphi + \left(\frac{\Lambda+\lambda}{2} - q\right)^2 = \left(\frac{\Lambda-\lambda}{2}\right)^2 ; \tag{8.25}$$

this relation also holds for $q > \frac{\Lambda+\lambda}{2}$ and may be interpreted as a quadratic equation for the unknown variable $v(\varphi)$. This quadratic equation always has two different solutions with opposite signs, the positive solution being the value $v(\varphi)$ itself.

The law of cosines applied to the triangle $QCB$ gives the equation

$$v(-\varphi)^2 + (\Lambda + \lambda - 2q)v(-\varphi)\sin\varphi + \Big(\frac{\Lambda+\lambda}{2} - q\Big)^2 = \Big(\frac{\Lambda-\lambda}{2}\Big)^2 , \tag{8.26}$$

and so $v$ solves the Eq. (8.25) if and only if $(-v)$ satisfies the Eq. (8.26). This implies that the negative solution to the Eq. (8.25) is the opposite value of $v(-\varphi)$. According to Vieta's theorem, we can write therefore

$$\begin{aligned} v(\varphi) - v(-\varphi) &= (\Lambda + \lambda - 2q)\sin\varphi , \\ -v(\varphi)v(-\varphi) &= \Big(\frac{\Lambda+\lambda}{2} - q\Big)^2 - \Big(\frac{\Lambda-\lambda}{2}\Big)^2 \; = \; (\Lambda - q)(\lambda - q) . \end{aligned}$$

After the substitution of these results into (8.24), we get

$$\begin{aligned} I_{d1} &= \int_0^{\frac{\pi}{2}} \frac{2q \ln \frac{\tan\varphi}{q}}{q^2 + q(\Lambda+\lambda-2q)\sin^2\varphi + (\Lambda - q)(\lambda - q)\sin^2\varphi}\,\mathrm{d}\varphi \\ &= \int_0^{\frac{\pi}{2}} \frac{2q \ln \frac{\tan\varphi}{q}}{q^2\cos^2\varphi + \Lambda\lambda\sin^2\varphi}\,\mathrm{d}\varphi . \end{aligned} \tag{8.27}$$

Let us define the variable $\psi \in \langle 0, \frac{\pi}{2})$ by the relation

$$\frac{\tan\psi}{\sqrt{\Lambda\lambda}} = \frac{\tan\varphi}{q}.$$

The integral (8.27) expressed using this substitution is

$$I_{d1} = \int_0^{\frac{\pi}{2}} \frac{2\ln\frac{\tan\psi}{\sqrt{\Lambda\lambda}}}{\sqrt{\Lambda\lambda}}\,\mathrm{d}\psi , \tag{8.28}$$

i.e., its value does not depend on the parameter $q$.

If $q = \lambda$, the relation (8.24) reduces to

$$I_{d1} = \int_0^{\frac{\pi}{2}} \frac{2v(\varphi)\ln\frac{\tan\varphi}{q}}{(\Lambda-\lambda)\,[p + v(\varphi)\sin\varphi]\cos[\varphi - \alpha(\varphi)]}\,\mathrm{d}\varphi = \int_0^{\frac{\pi}{2}} \frac{2\ln\frac{\tan\varphi}{\lambda}}{\lambda + v(\varphi)\sin\varphi}\,\mathrm{d}\varphi.$$

Since the angle $\angle PCA = 2\varphi$ in this case, the value $v(\varphi)$ can be expressed as

$$v(\varphi) = (\Lambda - \lambda)\sin\varphi$$

and

$$I_{d1} = \int\limits_0^{\frac{\pi}{2}} \frac{2\ln\frac{\tan\varphi}{\lambda}}{\lambda + (\Lambda-\lambda)\sin^2\varphi}\, \mathrm{d}\varphi = \int\limits_0^{\frac{\pi}{2}} \frac{2\ln\frac{\tan\varphi}{\lambda}}{\lambda\cos^2\varphi + \Lambda\sin^2\varphi}\, \mathrm{d}\varphi\ .$$

The substitution $\frac{\tan\varphi}{\sqrt{\lambda}} = \frac{\tan\psi}{\sqrt{\Lambda}}$ gives the result (8.28) again; similarly, we get the same result for $q = \Lambda$.

## Appendix 3: Existence of the Integral *J*

Let us define

$$\Omega_0 = \Omega \setminus \Big( (\lambda, \lambda + \tfrac{\Lambda}{\kappa(\mathsf{A})-1}\rangle \times \langle\lambda, \Lambda\rangle \cup \langle \Lambda - \tfrac{\Lambda}{\kappa(\mathsf{A})-1}, \Lambda) \times \langle\lambda, \Lambda\rangle \cup \{|x-l| < \tfrac{\Lambda}{\kappa(\mathsf{A})-1}\} \Big)\,,$$

$$\Omega_k = \Big[ \Omega \setminus \Big( (\lambda, \lambda + \tfrac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}\rangle \times \langle\lambda, \Lambda\rangle \cup \langle \Lambda - \tfrac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}, \Lambda) \times \langle\lambda, \Lambda\rangle$$
$$\cup \{|x-l| < \frac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}\} \Big) \Big] \setminus \bigcup_{j=0}^{k-1} \Omega_j \quad \text{for } k \in \mathcal{N}\ .$$

The sets $\Omega_k$ for $k \in \mathcal{N}$ consist of two V-shaped strips with the width $\frac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}$; these are shown in Fig. 13. Note, if the inequality

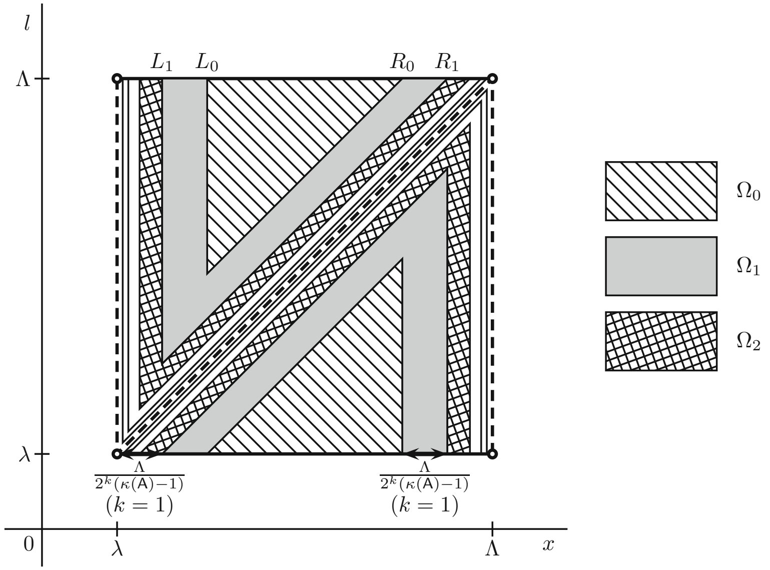

**Fig. 13** Splitting of the set $\Omega$ into subsets $\Omega_k$

$$2 \cdot \frac{\Lambda}{2^k(\kappa(\mathsf{A})-1)} > \Lambda - \lambda \tag{8.29}$$

is satisfied for a number $k$, then the set $\Omega_k$ is empty; geometrically it follows from this fact than the point $L_k$ in Fig. 13 lies to the right of the point $R_k$. Contrarily, when $\Omega_k \neq \emptyset$, then

$$2^k(\kappa(\mathsf{A})-1) \geq \frac{2\Lambda}{\Lambda - \lambda} > 2\,; \tag{8.30}$$

this information will be used later.

The intersection $\Omega_i \cap \Omega_j = \emptyset$ for $i \neq j$ and

$$\Omega = \bigcup_{k \geq 0} \Omega_k \subset \langle \lambda, \Lambda \rangle \times \langle \lambda, \Lambda \rangle\,. \tag{8.31}$$

Now, we find the bounds for the values $f(x,l)$, when $[x,l] \in \Omega_0$. The argument of the logarithm in the definition of the function $f(x,l)$ is

$$\left|1 - \frac{x}{l}\right| = \frac{|x-l|}{l} \geq \frac{\Lambda}{l(\kappa(\mathsf{A})-1)} \geq \frac{1}{\kappa(\mathsf{A})-1}\,, \tag{8.32}$$

at the same time

$$\left|1 - \frac{x}{l}\right| \leq \frac{\max\limits_{[x,l]\in\langle\lambda,\Lambda\rangle\times\langle\lambda,\Lambda\rangle} |\, l - x \,|}{\min\limits_{[x,l]\in\langle\lambda,\Lambda\rangle\times\langle\lambda,\Lambda\rangle} l} \leq \frac{\Lambda - \lambda}{\lambda} = \kappa(\mathsf{A}) - 1 \tag{8.33}$$

on the whole square $\langle \lambda, \Lambda \rangle \times \langle \lambda, \Lambda \rangle$, so

$$\left|\ln\left|1 - \frac{x}{l}\right|\right| \leq |\ln(\kappa(\mathsf{A})-1)|\,.$$

on $\Omega_0$. When $k > 0$, we can get by the way leading to (8.32)

$$\left|1 - \frac{x}{l}\right| > \frac{1}{2^k(\kappa(\mathsf{A})-1)}$$

on $\Omega_k$. If $\Omega_k \neq \emptyset$, then $\ln\left[2^k(\kappa(\mathsf{A}-1)\right] > 0$ due to (8.30), and we can estimate

$$\left|\ln\left|1 - \frac{x}{l}\right|\right| \leq \left|\ln\left[2^k(\kappa(\mathsf{A})-1)\right]\right| = \ln\left[2^k(\kappa(\mathsf{A})-1)\right] = k\ln 2 + \ln(\kappa(\mathsf{A})-1)\,. \tag{8.34}$$

The denominator of the expression (3.3) is bounded from below on $\Omega_k \neq \emptyset$ by the value

$$
\begin{aligned}
&x\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2-\left(x-\frac{\Lambda+\lambda}{2}\right)^2}\\
&\quad\ge\left(\lambda+\frac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}\right)\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2-\left(\lambda+\frac{\Lambda}{2^k(\kappa(\mathsf{A})-1)}-\frac{\Lambda+\lambda}{2}\right)^2}\\
&\quad=\lambda^2\frac{(2^k+1)\kappa(\mathsf{A})-2^k}{2^k(\kappa(\mathsf{A})-1)}\sqrt{\left[1-\frac{\kappa(\mathsf{A})}{2^k(\kappa(\mathsf{A})-1)^2}\right]\frac{\kappa(\mathsf{A})}{2^k}}\ge\lambda^2\sqrt{\frac{\kappa(\mathsf{A})}{2^{k+1}}}\ ; \qquad (8.35)
\end{aligned}
$$

the last inequality is valid as a consequence of (8.30).

We can further estimate the measures

$$
\begin{aligned}
\mu(\Omega_0)&\le\Lambda-\lambda\ ,\\
\mu(\Omega_k)&\le\frac{4\Lambda}{2^k(\kappa(\mathsf{A})-1)}\quad\text{for } k\in\mathcal{N}\ , \qquad (8.36)
\end{aligned}
$$

$$
\mu\left(\langle\lambda,\Lambda\rangle\times\langle\lambda,\Lambda\rangle\setminus\bigcup_{i=0}^{k}\Omega_k\right)\le\frac{4\Lambda}{2^k(\kappa(\mathsf{A})-1)} \qquad (8.37)
$$

—all these results follow easily from Fig. 13. The estimate (8.37) together with (8.31) implies

$$
\mu\left(\langle\lambda,\Lambda\rangle\times\langle\lambda,\Lambda\rangle\right)=\mu(\Omega)
$$

A sufficient condition for the existence of the integral $J$ defined by (3.4) is the existence of the integral

$$
\begin{aligned}
\widetilde{J}&=\int\limits_{\Omega}|f(x,l)|\,\mathrm{d}\mu=\sum_{k\ge0}\int\limits_{\Omega_k}|f(x,l)|\,\mathrm{d}\mu\\
&=\int\limits_{\Omega_0}|f(x,l)|\,\mathrm{d}\mu+\sum_{k\ge\min(\log_2\frac{2\kappa(\mathsf{A})}{(\kappa(\mathsf{A})-1)^2},1)}\ \int\limits_{\Omega_k}|f(x,l)|\,\mathrm{d}\mu\ . \qquad (8.38)
\end{aligned}
$$

The condition $k\ge\log_2\frac{2\kappa(\mathsf{A})}{(\kappa(\mathsf{A})-1)^2}$ follows from the relation (8.29). It holds

$$
\int\limits_{\Omega_k}|f(x,l)|\,\mathrm{d}\mu\le\sup_{[x,l]\in\Omega_k}|f(x,l)|\mu(\Omega_k) \qquad (8.39)
$$

for any $k\ge0$. The first integral on the right-hand side of (8.38) is zero at $\Omega_0=\emptyset$, or—according to (8.39)

$$
\int\limits_{\Omega_0}|f(x,l)|\,\mathrm{d}\mu\le\frac{\ln(\kappa(\mathsf{A})-1)}{\lambda^2\sqrt{\frac{\kappa(\mathsf{A})}{2}}}(\Lambda-\lambda)\ ;
$$

in any case, its value is finite. For $k \geq k_0 = \min(\log_2 \frac{2\kappa(\mathsf{A})}{(\kappa(\mathsf{A})-1)^2}, 1)$, the terms in the sum on the right-hand side of (8.38) may be estimated by (8.39) using (8.34), (8.35), and (8.36)

$$\begin{aligned}\int\limits_{\Omega_k} |f(x,l)| \,\mathrm{d}\mu &\leq \frac{k \ln 2 + \ln(\kappa(\mathsf{A})-1)}{\lambda^2 \sqrt{\frac{\kappa(\mathsf{A})}{2^{k+1}}}} \cdot \frac{4\Lambda}{2^k(\kappa(\mathsf{A})-1)} \\ &= \frac{4\sqrt{2\kappa(\mathsf{A})}}{\lambda(\kappa(\mathsf{A})-1)} \cdot \frac{k \ln 2 + \ln(\kappa(\mathsf{A})-1)}{\left(\sqrt{2}\right)^k} \,.\end{aligned}$$

The sum

$$\begin{aligned}\sum_{k \geq k_0} \int\limits_{\Omega_k} |f(x,l)| \,\mathrm{d}\mu &\leq \frac{4\sqrt{2\kappa(\mathsf{A})}}{\lambda(\kappa(\mathsf{A})-1)} \sum_{k \geq k_0} \frac{k \ln 2 + \ln(\kappa(\mathsf{A})-1)}{\left(\sqrt{2}\right)^k} = \\ &= \frac{4\sqrt{2\kappa(\mathsf{A})}}{\lambda(\kappa(\mathsf{A})-1)} \cdot \left[\frac{k_0 \ln 2 + \ln(\kappa(\mathsf{A})-1)}{\left(\sqrt{2}\right)^{k_0-1}\left(\sqrt{2}-1\right)} + \frac{\ln 2}{\left(\sqrt{2}\right)^{k_0-1}\left(\sqrt{2}-1\right)^2}\right]\end{aligned}$$

So, the series on the right-hand side of (8.38) is absolutely convergent, and the integral $J$ exists.

## Appendix 4: Construction of the Optimal Step Length Distribution

First, we prove

**Lemma 4** *If a distribution function $F_L(l)$ generates the function $\widehat{E}_R(x)$ satisfying the conditions of Theorem 3, then the function $\widehat{E}_R(x)$ is defined for all $x \in \langle \lambda, \Lambda \rangle$.*

*Proof* We show that for any point $\hat{x}$ of the second or third type relative to the measure $\phi$, there exists such a $\delta > 0$ that for any point $x \in \mathcal{D}_\delta(\hat{x}) = (\hat{x}-\delta, \hat{x}+\delta) \cap \langle \lambda, \Lambda \rangle$ of the first type relative to the measure $\phi$ is $\widehat{E}_R(x) < \overline{E}_\phi - 1$. Since almost all the points $x \in \mathcal{D}_\delta(\hat{x})$ are the points of the first type relative to the measure $\phi$ (it is a consequence of Theorem 2) and $\nu\big(x \in \mathcal{D}_\delta(\hat{x})\big) \geq \frac{2\delta}{\Lambda(\Lambda-\lambda)} > 0$, the conditions of Theorem 3 are not satisfied in the considered case.

Suppose $\hat{x} \in \langle \lambda, \Lambda \rangle$ be a point of the third type relative to the measure $\phi$. According to the definition, $\phi(\hat{x}) = m > 0$. If $x \in \langle \lambda, \Lambda \rangle$ is a point of the first type relative to the measure $\phi$, then the mean value $\widehat{E}_R(x)$ exists, and

$$\begin{aligned}\widehat{E}_R(x) &= \int\limits_{\langle \lambda, \hat{x})} \ln\left|1 - \frac{x}{l}\right| \,\mathrm{d}\phi(l) + m \cdot \ln\left|1 - \frac{x}{\hat{x}}\right| + \int\limits_{(\hat{x}, \Lambda\rangle} \ln\left|1 - \frac{x}{l}\right| \,\mathrm{d}\phi(l) \\ &\leq (1-m)\ln(\kappa(\mathsf{A})-1) + m \cdot \ln\left|1 - \frac{x}{\hat{x}}\right| \qquad (8.40)\end{aligned}$$

due to (8.33). Let us denote, using the symbols applied in the proof of Theorem 3

$$, \delta = \hat{x} \cdot \frac{e^{\frac{I_d}{m\nu(\langle\lambda,\Lambda\rangle)}}}{e^{\frac{1}{m}}\,(\kappa(\mathsf{A})-1)^{\frac{1}{m}-1}} > 0\,.$$

For any $x \in \mathcal{D}_\delta(\hat{x})$ being the point of the first type relative to the measure $\phi$, the right-hand side of (8.40) may be estimated

$$\widehat{E}_R(x) \le (1-m)\ln(\kappa(\mathsf{A})-1) + m \cdot \ln\frac{\delta}{\hat{x}} = \frac{I_d}{\nu(\langle\lambda,\Lambda\rangle)} - 1 \le \overline{E}_\phi - 1$$

according to (3.10).

Assuming $\hat{x}$ be a point of the second type relative to the measure $\phi$, there exists, consequently of Lemma 3, such a $\delta \in (0, \frac{\hat{x}}{2})$ that

$$\int\limits_{\langle\lambda,\Lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1-\frac{\hat{x}}{l}\right| \mathrm{d}\phi(l) < \overline{E}_\phi - 1 - 2\ln 2\,.$$

For any $x \in \mathcal{D}_\delta(\hat{x})$ being the point of first type relative to the measure $\phi$, the mean value

$$\begin{aligned}
\widehat{E}_R(x) &= \int\limits_{\langle\Lambda,\lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1-\frac{x}{l}\right| \mathrm{d}\phi(l) + \int\limits_{\mathcal{D}_\delta(\hat{x})} \ln\left|1-\frac{x}{l}\right| \mathrm{d}\phi(l) \\
&\le \int\limits_{\langle\Lambda,\lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \left(\ln\left|1-\frac{\hat{x}}{l}\right| + \ln\left|\frac{l-x}{l-\hat{x}}\right|\right) \mathrm{d}\phi(l) + \int\limits_{\mathcal{D}_\delta(\hat{x})} \ln 2\,\mathrm{d}\phi(l) \\
&< \overline{E}_\phi - 1 - 2\ln 2 + \int\limits_{\langle\Lambda,\lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1+\frac{\hat{x}-x}{l-\hat{x}}\right| \mathrm{d}\phi(l) + \ln 2\,. \qquad (8.41)
\end{aligned}$$

We can express the value $x$ in the form $x = \hat{x} + \alpha\delta$ for some $\alpha \in (-1, 1)$. Suppose first $\alpha > 0$. Then,

$$\begin{aligned}
\int\limits_{\langle\Lambda,\lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1+\frac{\hat{x}-x}{l-\hat{x}}\right| \mathrm{d}\phi(l) &= \int\limits_{(0,\hat{x}-\delta)\cap\langle\lambda,\hat{x}\rangle} \ln\left|1+\frac{\alpha\delta}{\hat{x}-l}\right| \mathrm{d}\phi(l) \\
&\quad + \int\limits_{\langle\hat{x}+\delta,\infty)\cap(\hat{x},\Lambda\rangle} \ln\left|1-\frac{\alpha\delta}{l-\hat{x}}\right| \mathrm{d}\phi(l)\,.
\end{aligned}$$

The integrand in the second integral on the right-hand side is always non-positive, and the denominator of the fraction in the first integral is always greater or equal to $\delta$; therefore, $1 + \frac{\alpha\delta}{\hat{x}-l} \le 1 + \alpha < 2$. It implies

$$\int\limits_{\langle\Lambda,\lambda\rangle\setminus(\hat{x}-\delta,\hat{x}+\delta)} \ln\left|1+\frac{\hat{x}-x}{l-\hat{x}}\right| \mathrm{d}\phi(l) < \int\limits_{(0,\hat{x}-\delta)\cap\langle\lambda,\hat{x}\rangle} \ln 2 \,\mathrm{d}\phi(l) \leq \ln 2$$

and in (8.41)

$$\widehat{E}_R(x) < \overline{E}_\phi - 1\,.$$

Analogously we get the same result at $\alpha < 0$ □

The wanted distribution function $F_L(l)$ is, as a consequence of Lemma 4, continuous, especially $F_L(\lambda) = 0$. This fact will be used later.

The function $\ln\left|1-\frac{x}{l}\right|$ may be expressed in the form of Fourier series of Chebyshev polynomials of the first kind of the variable $l$; the Fourier coefficients $c_n(x)$ are calculated in “Appendix 5.” We show that the partial sums of the created series are bounded by an integrable function for all $x \in \langle\lambda, \Lambda\rangle$, if the measure $\phi$ generates the function $\widehat{E}_R(x)$ satisfying the condition (3.6). Indeed, using the notation

$$\alpha = \arccos\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\,, \tag{8.42}$$

and applying the results (8.61), (8.62), we get (the parameters $\xi$, $\zeta$ defined by (8.63) and (8.64), respectively, are used)

$$\begin{aligned} Z_k(l,x) = c_0 + \sum_{n=1}^{k}\sqrt{2}c_n(x)T_n(\cos\alpha) \; &= \; -\zeta + \sum_{n=1}^{k}\frac{2}{n}(e^{-n\zeta} - \cos n\xi)\cos n\alpha \\ &= -\zeta + \sum_{n=1}^{k}\frac{2}{n}\left(e^{-n\zeta}-1\right)\cos n\alpha - \sum_{n=1}^{k}\frac{2}{n}\left(\cos n\xi - 1\right)\cos n\alpha \\ &= -\zeta - \sum_{n=1}^{k}\int_0^{\zeta} 2e^{-nt}\cos n\alpha\,\mathrm{d}t + \sum_{n=1}^{k}\int_0^{\xi} 2\sin nt\cdot\cos n\alpha\,\mathrm{d}t \\ &= -\zeta + \int_0^{\zeta}\Bigg(1 - e^{-kt}\cos k\alpha - \frac{\sinh t\left(1-e^{-kt}\cos k\alpha\right)}{\cosh t - \cos\alpha} \\ &\quad - \frac{e^{-kt}\sin k\alpha\cos\frac{\alpha}{2}\sin\frac{\alpha}{2}}{\sinh^2\frac{t}{2}+\sin^2\frac{\alpha}{2}}\Bigg)\mathrm{d}t \\ &\quad + \int_{\alpha}^{\xi+\alpha}\frac{\cos\frac{u}{2}-\cos(k+\frac{1}{2})u}{2\sin\frac{u}{2}}\,\mathrm{d}u + \int_{-\alpha}^{\xi-\alpha}\frac{\cos\frac{u}{2}-\cos(k+\frac{1}{2})u}{2\sin\frac{u}{2}}\,\mathrm{d}u\,. \end{aligned} \tag{8.43}$$

The integrated fraction in the last integral on the right-hand side of (8.43) is an odd and bounded function (it follows from the fact that it is a sum of $k$ sines); therefore, it is integrable on any bounded interval. Due to the oddness of the integrand, the last integral is equal to the integral of the same function over the interval $(\alpha, |\alpha - \xi|)$. Now, we can estimate the sum $Z_k(l, x)$

$$|Z_k(l,x)| \leq 3\zeta + \int\limits_0^{\zeta} \frac{2\sinh t}{\cosh t - \cos\alpha}\,\mathrm{d}t + \left|\sin k\alpha \cos\frac{\alpha}{2}\right| \int\limits_0^{\zeta} \frac{\cosh\frac{t}{2}\sin\frac{\alpha}{2}}{\sinh^2\frac{t}{2} + \sin^2\frac{\alpha}{2}}\,\mathrm{d}t$$
$$+ \int\limits_{\alpha}^{\xi+\alpha} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u + \left| \int\limits_{\alpha}^{|\alpha-\xi|} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u \right| . \tag{8.44}$$

The absolute value of the last integral in (8.44) is necessary because the ordering of its limits depends on the relation between the values $\alpha, \xi$. Nevertheless, the integrand is always a positive function, and since in the case $\alpha < |\alpha - \xi|$ it is $(\alpha, |\alpha - \xi|) \subset (\alpha, \alpha + \xi)$ and for $\alpha \geq |\alpha - \xi|$ it holds $(|\alpha - \xi|, \alpha) \subset (|\alpha - \xi|, \alpha + \xi)$, we can estimate

$$\left| \int\limits_{\alpha}^{|\alpha-\xi|} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u \right| \leq \int\limits_{\alpha}^{\alpha+\xi} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u + \int\limits_{|\alpha-\xi|}^{\alpha+\xi} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u . \tag{8.45}$$

The integrals on the right-hand side of (8.45) have their limits in ascending order, because $\alpha \in \langle 0, \pi\rangle, \xi \in \langle 0, \pi\rangle$. Furthermore, the last mentioned fact implies $(\alpha, \alpha+\xi) \subset (\alpha, \alpha + \pi)$, and we can continue estimating as follows:

$$\begin{aligned}
|Z_k(l,x)| &\leq 3\zeta + 2\ln\frac{\cosh\zeta - \cos\alpha}{1 - \cos\alpha} + 2\left|\sin k\alpha\cos\frac{\alpha}{2}\right| \arctan\frac{\sinh\frac{\zeta}{2}}{\sin\frac{\alpha}{2}} \\
&\quad +2\int\limits_{\alpha}^{\alpha+\pi} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u + \int\limits_{|\alpha-\xi|}^{\alpha+\xi} \frac{1}{\sin\frac{u}{2}}\,\mathrm{d}u \\
&\leq 3\zeta + 2\ln\frac{l}{l-\lambda} + \pi + 4\ln\frac{\tan\frac{\alpha+\pi}{4}}{\tan\frac{\alpha}{4}} + 2\ln\frac{\tan\frac{\alpha+\xi}{4}}{\tan\frac{|\alpha-\xi|}{4}} \\
&= 3\zeta + \pi - 2\ln\left(1 - \frac{\lambda}{l}\right) \\
&\quad + 2\ln\frac{4(1+\sin\frac{\alpha}{2})^2(1+\cos\frac{\alpha}{2})^2}{(1+\cos\alpha)(1-\cos\alpha)} + 2\ln\frac{2(\sin\frac{\alpha}{2}+\sin\frac{\xi}{2})^2}{|\cos\xi - \cos\alpha|} \\
&\leq 3\zeta + \pi - 2\ln\left(1 - \frac{\lambda}{l}\right) + 2\ln\frac{16(\Lambda-\lambda)^2}{(\Lambda-l)(l-\lambda)} + 2\ln\frac{4(\Lambda-\lambda)}{|l-x|} \\
&= 3\zeta + \pi - 2\ln\left(1 - \frac{\lambda}{l}\right) + 12\ln 2 + 6\ln\frac{\Lambda-\lambda}{l} \\
&\quad - 2\ln\frac{\Lambda-l}{l} - 2\ln\frac{l-\lambda}{l} - 2\ln\frac{|l-x|}{l}
\end{aligned}$$

$$\leq 3\zeta + \pi + 12\ln 2 + 6\ln(\kappa(\mathsf{A})-1) - 4\ln\left(1-\frac{\lambda}{l}\right)$$
$$-2\ln\left|1-\frac{\Lambda}{l}\right| - 2\ln\left|1-\frac{x}{l}\right| \;=\; z_x(l). \tag{8.46}$$

The procedure resulting in (8.46) is fully correct only when $\alpha > 0$; in the case $\alpha = 0$, the integration of the term $\frac{\cosh\frac{t}{2}\sin\frac{\alpha}{2}}{\sinh^2\frac{t}{2}+\sin^2\frac{\alpha}{2}}$ in (8.44) does not give the antiderivative $2\arctan\frac{\sinh\frac{t}{2}}{\sin\frac{\alpha}{2}}$. The integrand is, however, zero except for the point $t = 0$ (where it is undefined) in this case; therefore, the corresponding integral is zero, and the estimate (8.46) holds too.

As follows from Lemma 4, if the measure $\phi$ satisfies our requirements, then the integral

$$\int\limits_{\langle\lambda,\Lambda\rangle} \ln\left|1-\frac{x}{l}\right| \,\mathrm{d}\phi(l)$$

exists for any $x \in \langle\lambda, \Lambda\rangle$, especially for $x = \lambda$ and $x = \Lambda$ too. It implies the partial sums $\{Z_k(l,x)\}_{k\geq 0}$ are for any $x \in \langle\lambda, \Lambda\rangle$ dominated by a common integrable function $z_x(l)$, and due to the Lebesgue dominated convergence theorem,

$$\widehat{E}_R(x) = \int\limits_{\lambda}^{\Lambda} \left[c_0 + \sum_{n\geq 1}\sqrt{2}c_n(x)T_n\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right)\right]\mathrm{d}\phi(l)$$
$$= c_0 + \sqrt{2}\sum_{n\geq 1}c_n(x)\int\limits_{\lambda}^{\Lambda} T_n\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right)\mathrm{d}\phi(l) \tag{8.47}$$

for all $x \in \langle\lambda, \Lambda\rangle$.

The integrals on the right-hand side of (8.47) may be modified using the integration by parts

$$\int\limits_{\lambda}^{\Lambda} T_n\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right)\mathrm{d}\phi(l) = (-1)^n + \frac{2}{\Lambda-\lambda}\int\limits_{\lambda}^{\Lambda} nU_{n-1}\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right)F_L(l)\,\mathrm{d}l$$

(the basic properties of the distribution function $F_L$ are used: $F_L(\Lambda) = 1$, $F_L(\lambda) = 0$ due to its necessary continuity). We use again the substitution (8.42); the distribution function $F_L(l)$ changes over to the function $\widetilde{F}_L(\alpha) = F_L(\frac{\Lambda+\lambda}{2} - \frac{\Lambda-\lambda}{2}\cos\alpha)$ and

$$\int\limits_{\lambda}^{\Lambda} T_n\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right)\mathrm{d}\phi(l) = (-1)^n + n\int\limits_{0}^{\pi} \widetilde{F}_L(\alpha)\,\sin n\alpha\,\mathrm{d}\alpha\ .$$

The function $\widetilde{F}_L(\alpha)$ may be symmetrically extended onto the interval $\langle -\pi, \pi\rangle$ defining

$$\breve{F}_L(\alpha) = \begin{cases} \widetilde{F}_L(\alpha) & \text{for } \alpha \geq 0, \\ -\widetilde{F}_L(-\alpha) & \text{for } \alpha \leq 0 \end{cases}$$

($\breve{F}_L(0) = \widetilde{F}_L(0) = 0$ according to its definition). The function $\breve{F}_L$ is continuous, bounded on $\langle -\pi, \pi\rangle$ and odd; therefore, $\breve{F}_L \in L_2(-\pi, \pi)$, and it may be expressed in the form of a trigonometric series

$$\breve{F}_L(\alpha) = \sum_{j\geq 1} d_j \sin\, j\alpha\ . \tag{8.48}$$

The expansion (8.48) may be used also on the interval $\langle 0, \pi\rangle$, where $\widetilde{F}_L(l) = \breve{F}_L(l)$; then

$$\int\limits_{\lambda}^{\Lambda} T_n\left(\frac{\Lambda+\lambda-2l}{\Lambda-\lambda}\right) \mathrm{d}\phi(l) = (-1)^n + n\int\limits_{0}^{\pi} \sin\, n\alpha \sum_{j\geq 1} d_j \sin\, j\alpha\, \mathrm{d}\alpha = (-1)^n + n\frac{\pi}{2} d_n\ . \tag{8.49}$$

Substituting the results (8.49) and (8.61) into (8.47), we get

$$\begin{aligned}\widehat{E}_R(x) &= c_0 + 2\sum_{n\geq 1} \frac{1}{n}\left((-1)^n + n\frac{\pi}{2}d_n\right)\left[e^{-n\zeta} - \cos\left(n \arccos \frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\right)\right] \\ &= c_0 + 2\sum_{n\geq 1}\left(\frac{(-1)^n}{n} + \frac{\pi}{2}d_n\right)\left[e^{-n\zeta} - T_n\left(\frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\right)\right]\ . \end{aligned} \tag{8.50}$$

The expression on the right-hand side of (8.50) may be written in the form

$$\widehat{E}_R(x) = \sum_{n\geq 0} b_n\, T_n\left(\frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\right)\ . \tag{8.51}$$

The coefficient $b_0$ is finite, because $|d_n| \leq \frac{\|\breve{F}_L\|}{\sqrt{\pi}} \leq \sqrt{2}$ for any $n \in \mathcal{N}$ due to (8.48) and

$$\begin{aligned}|b_0| &= \left| c_0 + 2\sum_{n\geq 1}\left(\frac{(-1)^n}{n} + \frac{\pi}{2}d_n\right)e^{-n\zeta}\right| \leq |c_0| + 2\sum_{n\geq 1}\left(1 + \frac{\pi}{\sqrt{2}}\right)e^{-n\zeta} \\ &\leq |c_0| + \left(2 + \pi\sqrt{2}\right)\frac{e^{-\zeta}}{1-e^{-\zeta}}\ .\end{aligned}$$

Moreover, since $b_n = -\frac{(-1)^n}{n} - \frac{\pi}{2}d_n$ and both the sequences $\{\frac{(-1)^n}{n}\}_{n\geq 1}$, $\{d_n\}_{n\geq 1}$ belong to the space $l_2$, also $\{b_n\}_{n\geq 0} \in l_2$ and the series (8.51) is a Fourier expansion of a function $\widehat{E}_R(x) \in L_{2,f}(\lambda, \Lambda)$ with the weight function $f = \frac{1}{\sqrt{(\Lambda-x)(x-\lambda)}}$.

The condition necessary to satisfy (3.6) is $\widehat{E}_R(x) = const$ almost everywhere on $\langle\lambda, \Lambda\rangle$. The Fourier series representation (8.51) of such a function must have all coefficients $b_n = 0$ for $n \geq 1$, so

$$d_n = \frac{2}{\pi} \cdot \frac{(-1)^{n-1}}{n} \tag{8.52}$$

for all $n \in \mathcal{N}$. Therefore, the Fourier series of the distribution function $\widetilde{F}_L(\alpha)$ is unequivocally defined.

Let us suppose the existence of two distribution functions $\widetilde{F}_1(\alpha)$, $\widetilde{F}_2(\alpha)$ satisfying our requirements and their extensions $\breve{F}_1(\alpha)$, $\breve{F}_2(\alpha)$. The functions $\breve{F}_i(\alpha)$ are continuous, as proved above, and they have bounded variation (equal to 2, as follows from the fact that $\widetilde{F}_i(\alpha)$ are distribution functions). The trigonometric series (8.48) of both the functions has the coefficients (8.52). According to Dirichlet-Jordan theorem, this series converges uniformly to the functions $\breve{F}_1(\alpha)$, $\breve{F}_2(\alpha)$ on any closed interval $\langle a, b\rangle \subset (-\pi, \pi)$; therefore, $\breve{F}_1(\alpha) = \breve{F}_2(\alpha)$ for any $\alpha \in (-\pi, \pi)$. The equality $\breve{F}_1(\pi) = \breve{F}_2(\pi)$ follows from the fact that $\widetilde{F}_L(\pi) = F_L(\Lambda) = 1$ for any distribution function $F_L$. This implies that there exists at most one distribution function $F_L(l)$ which generates the function $\widehat{E}_R(x)$ satisfying the condition (3.6).

The last task is the finding of the distribution function $F_l(l)$ corresponding to the series (8.48) with the coefficients (8.52). It is known that the series

$$\sum_{n\geq 1} \frac{2}{n}(-1)^{n-1} \sin nx$$

is trigonometric expansion of the function $x$ on the interval $(-\pi, \pi)$. Using this fact and the requirement $\widetilde{F}_L(\pi) = 1$, we get

$$\begin{aligned} \widetilde{F}_L(\alpha) &= \frac{\alpha}{\pi}\ , \\ F_L(l) &= \frac{1}{\pi} \arccos \frac{\Lambda + \lambda - 2l}{\Lambda - \lambda}\ . \end{aligned} \tag{8.53}$$

## Appendix 5: Variance of the Process with Optimal Step Length Distribution

Calculations in this appendix are technically complicated, and the details would require a lot of space. Therefore, we usually present only their basic principles and the final results.

Considering the distribution (3.13) of the random variable $L$, we express the variance $\widehat{\sigma}_R^2(x)$ using the second moment

$$\widehat{\sigma}_R^2(x) = \int_\lambda^\Lambda \frac{\left(\ln\left|1 - \frac{x}{l}\right|\right)^2}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(l - \frac{\Lambda+\lambda}{2}\right)^2}}\, \mathrm{d}l - \left(\widehat{E}_R(x)\right)^2\ . \tag{8.54}$$

The integral in (8.54) exists, because all the present singularities are integrable. The probability density $f$ defined by (3.13) may be considered a weight function on $(\lambda, \Lambda)$; the existence of the integral in (8.54) implies

$$\ln\left|1-\frac{x}{l}\right| \in L_{2,f}(\lambda,\Lambda),$$

and we can write

$$\widehat{\sigma}_R^2(x) = \left\|\ln\left|1-\frac{x}{l}\right|\right\|^2_{L_{2,f}(\lambda,\Lambda)} - \left(\widehat{E}_R(x)\right)^2 \tag{8.55}$$

If $\mathcal{B} = \{\mathsf{b}_i\}_{i\geq 0}$ is an orthonormal basis of a real Hilbert space $\mathcal{H}$, $\mathsf{u} \in \mathcal{H}$, then we can express the vector $\mathsf{u}$ using the Fourier series

$$\mathsf{u} = \sum_{i\geq 0} c_i \cdot \mathsf{b}_i = \sum_{i\geq 0} \langle \mathsf{b}_i | \mathsf{u} \rangle_{\mathcal{H}} \cdot \mathsf{b}_i ,$$

and its norm satisfies the Parseval's identity

$$\|\mathsf{u}\|^2 = \sum_{i\geq 0} c_i^2 = \sum_{i\geq 0} \langle \mathsf{b}_i | \mathsf{u} \rangle^2_{\mathcal{H}} . \tag{8.56}$$

The most common orthonormal basis in the Hilbert space $L_{2,f}(\lambda,\Lambda)$ with the weight (3.13) is constituted by the Chebyshev polynomials of the first kind

$$\begin{aligned}
\mathsf{b}_0 = T_0(l) &= 1\,, \\
\mathsf{b}_n = \sqrt{2}\cdot T_n\left(\frac{\lambda+\Lambda-2l}{\Lambda-\lambda}\right) &= \sqrt{2}\cdot\cos\left(n\cdot\arccos\frac{\lambda+\Lambda-2l}{\Lambda-\lambda}\right) \quad \text{for } n \in \mathcal{N}\,.
\end{aligned}$$

Thus, the problem of the variance expression is transformed to the calculation of the Fourier coefficients

$$\begin{aligned}
c_0(x) &= \int_\lambda^\Lambda \frac{\ln\left|1-\frac{x}{l}\right|}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2-\left(l-\frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}l \;=\; \widehat{E}_R(x)\,, \\
c_n(x) &= \sqrt{2}\int_\lambda^\Lambda \frac{\ln\left|1-\frac{x}{l}\right|\cdot T_n\left(\frac{2l-(\lambda+\Lambda)}{\Lambda-\lambda}\right)}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2-\left(l-\frac{\Lambda+\lambda}{2}\right)^2}}\,\mathrm{d}l \quad \text{for } n \in \mathcal{N}\,.
\end{aligned} \tag{8.57}$$

The relation (8.54) may be then written, using (8.56) and (8.57), in the form

$$\widehat{\sigma}_R^2(x) = \sum_{n\geq 1} c_n^2\,. \tag{8.58}$$

Let $\Omega_\varepsilon(x) = (\lambda, \Lambda) \times \langle 0, x\rangle \setminus \{[l, t] : |l - t| < \varepsilon\}$ for $x \in \langle \lambda, \Lambda\rangle$, $\varepsilon \in (0, \min(\lambda, \frac{\Lambda-\lambda}{3}))$, and

$$g_n(l, t) = \frac{T_n\left(\frac{\lambda+\Lambda-2l}{\Lambda-\lambda}\right)}{\pi(t-l)\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2 - \left(l - \frac{\Lambda+\lambda}{2}\right)^2}}$$

for integer $n \geq 0$. We express the integral

$$C_n(\varepsilon, x) = \int\limits_{\Omega_\varepsilon(x)} g_n(l, t)\, \mathrm{d}l\, \mathrm{d}t$$

The integral $C_n(\varepsilon, x)$ exists for arbitrary $x \in \langle \lambda, \Lambda\rangle$ and any $\varepsilon \in (0, \min\left(\lambda, \frac{\Lambda-\lambda}{3}\right))$, because the singularity $\frac{1}{\sqrt{(\Lambda-\lambda)^2-(2l-\Lambda-\lambda)^2}}$ is integrable. So, according to the Fubini's theorem,

$$C_n(\varepsilon, x) = \int\limits_{\lambda}^{\Lambda} k_n(l)\, \mathrm{d}l \tag{8.59}$$

where

$$k_n(l) = \int\limits_{\langle 0,x\rangle\setminus(l-\varepsilon,l-\varepsilon)} g_n(l, t)\, \mathrm{d}t\ .$$

After the integration (8.59), we get

$$c_n(x) = \sqrt{2} \lim_{\varepsilon\to 0^+} C_n(\varepsilon, x) \tag{8.60}$$

for $n \in \mathcal{N}$; analogous result without the factor $\sqrt{2}$ holds for the value $c_0$.

The change in the integration order when computing the integral $C_n(\varepsilon, x)$ gives

$$C_n(\varepsilon, x) = \int\limits_{0}^{x} h_n(t, \varepsilon)\, \mathrm{d}t$$

where

$$h_n(t, \varepsilon) = \int\limits_{\langle\lambda,\Lambda\rangle\setminus(t-\varepsilon,t+\varepsilon)} g_n(l, t)\, \mathrm{d}l\ .$$

The recurrence formula for the Chebyshev polynomials implies

$$g_n(l, t) = T_n\left(\frac{\lambda+\Lambda-2t}{\Lambda-\lambda}\right) g_0(l, t) + \sum_{i=1}^{n} U_{i-1}\left(\frac{\lambda+\Lambda-2t}{\Lambda-\lambda}\right) r_{n-i}(l)\ ,$$

where $U_n$ denote the Chebyshev polynomials of second kind and

$$r_n(l) = \frac{\frac{2(1+\mathrm{sign} n)}{\Lambda-\lambda}\cdot T_n\left(\frac{\lambda+\Lambda-2l}{\Lambda-\lambda}\right)}{\pi\sqrt{\left(\frac{\Lambda-\lambda}{2}\right)^2-\left(l-\frac{\Lambda+\lambda}{2}\right)^2}}\,.$$

After all the needed calculations (during the process, we find that many calculated integrals converge to zero at $\varepsilon \to 0^+$), we get the Fourier coefficients $c_n(x)$, according to (8.60)

$$\begin{aligned} c_n(x) &= \sqrt{2}\lim_{\varepsilon\to 0^+}\left[-\int_0^{\lambda-\varepsilon}\frac{T_n\left(\frac{\lambda+\Lambda-2t}{\Lambda-\lambda}\right)}{\sqrt{(\Lambda-t)(\lambda-t)}}\,\mathrm{d}t+\int_0^{x}\frac{2U_{n-1}\left(\frac{\lambda+\Lambda-2t}{\Lambda-\lambda}\right)}{\Lambda-\lambda}\,\mathrm{d}t\right] \\ &= \frac{\sqrt{2}}{n}\left[e^{-n\,\mathrm{argcosh}\,\frac{\lambda+\Lambda}{\Lambda-\lambda}}-\cos\left(n\arccos\frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\right)\right]\quad \text{for } n\in\mathcal{N}\,, \end{aligned} \tag{8.61}$$

$$c_0(x) = -\lim_{\varepsilon\to 0^+}\int_0^{\lambda-\varepsilon}\frac{1}{\sqrt{(\Lambda-t)(\lambda-t)}}\,\mathrm{d}t \;=\; -\,\mathrm{argcosh}\,\frac{\lambda+\Lambda}{\Lambda-\lambda}\,. \tag{8.62}$$

The Fourier coefficients $c_n(x)$ were calculated with the aim to express the variance $\widehat{\sigma}_R^2(x)$ according to (8.58). Before the attempt to calculate this sum, we define the parameters

$$\xi = \arccos\frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\,, \tag{8.63}$$

$$\zeta = \mathrm{argcosh}\,\frac{\lambda+\Lambda}{\Lambda-\lambda} = \ln\frac{\sqrt{\kappa(\mathsf{A})}+1}{\sqrt{\kappa(\mathsf{A})}-1}\,. \tag{8.64}$$

The sum (8.58) may be written using (8.61) in the form

$$\widehat{\sigma}_R^2(x) = 2\sum_{n\ge 1}\frac{1}{n^2}\left[\left(1-e^{-n\zeta}\right)^2-2\left(1-e^{-n\zeta}\right)(1-\cos n\xi)+(1-\cos n\xi)^2\right] \tag{8.65}$$

We did not find the analytic expression of the sum (8.65); nevertheless, it may be approximated rather exactly, namely for small values $\zeta$ corresponding to large condition numbers $\kappa(\mathsf{A})$.

The contribution of the first term in the square bracket in (8.65) to the total sum is

$$\sum_{n\ge 1}\frac{1}{n^2}\left(1-e^{-n\zeta}\right)^2 = \sum_{n\ge 1}\left[\int_0^{\zeta}e^{-nt}\,\mathrm{d}t\right]^2 = \sum_{n\ge 1}\int_{\langle 0,\zeta\rangle\times\langle 0,\zeta\rangle}e^{-n(t+u)}\,\mathrm{d}S\,. \tag{8.66}$$

**Table 2** Inaccuracy inserted into the sum (8.65) applying the estimate (8.68) in (8.67)

| $\kappa(\mathsf{A})$ | 1.2 | 3 | 10 | 30 | 100 |
|---|---|---|---|---|---|
| Inaccuracy | $9.802 \times 10^{-3}$ | $2.184 \times 10^{-4}$ | $4.955 \times 10^{-6}$ | $1.853 \times 10^{-7}$ | $5.180 \times 10^{-9}$ |

We describe the integration area $[t, u] \in \langle 0, \zeta\rangle \times \langle 0, \zeta\rangle$ using the variables

$$\begin{aligned} v &= t + u \,, \\ w &= t - u \,. \end{aligned}$$

Then, the integral (8.66) may be transformed into the form

$$\sum_{n\geq 1} \frac{1}{n^2}\left(1 - e^{-n\zeta}\right)^2 = 2\zeta \ln\left(1 + e^{-\zeta}\right) + 2\int_0^{\zeta} \frac{v e^{-v}}{1 + e^{-v}}\, \mathrm{d}v \,. \tag{8.67}$$

We can estimate the remaining integral on the right-hand side of (8.67) using the approximation

$$v \approx \frac{2}{3}\sinh\frac{v}{2} + \frac{8}{3}\tanh\frac{v}{4}\,. \tag{8.68}$$

This (upper) estimate is very accurate for common condition numbers $\kappa(\mathsf{A})$; values of the inaccuracy of the result are demonstrated in Table 2.

The inaccuracy behaves proportionally to $[\kappa(\mathsf{A})]^{-3}$ for sufficiently large $\kappa(\mathsf{A})$, and its values are negligible in comparison with the total sum in (8.65).

After the integration in (8.67) using the approximation (8.68), we get

$$\begin{aligned} 2\sum_{n\geq 1}\frac{1}{n^2}\left(1 - e^{-n\zeta}\right)^2 \approx{}& 4\ln\frac{\sqrt{\kappa(\mathsf{A})}+1}{\sqrt{\kappa(\mathsf{A})}-1}\ln\frac{2\sqrt{\kappa(\mathsf{A})}}{\sqrt{\kappa(\mathsf{A})}+1} + \frac{8}{3}\left[\frac{\sqrt{\kappa(\mathsf{A})-1}}{\sqrt{\kappa(\mathsf{A})-1}} - 1\right] \\ &+ \frac{32}{3}\ln\frac{\sqrt{\kappa(\mathsf{A})-1}+\sqrt{\kappa(\mathsf{A})}}{2(\sqrt{\kappa(\mathsf{A})}+1)} + \frac{40}{3}\arctan\frac{1}{\sqrt{\kappa(\mathsf{A})-1}}\,. \end{aligned} \tag{8.69}$$

The second term in the square bracket in (8.65) may be expressed in a similar way (without the use of variables $v$, $w$). The application of Euler's formula and the approximation (8.68) give

$$\begin{aligned} -4\sum_{n\geq 1}\frac{1}{n^2}\left(1 - e^{-n\zeta}\right)(1 - \cos n\xi) &= -4\sum_{n\geq 1}\int_0^{\zeta} e^{-nt}\,\mathrm{d}t\int_0^{\xi}\sin nu\,\mathrm{d}u \\ &= -4\int_0^{\zeta}\int_0^{\xi}\frac{e^{-t}\sin u}{1 - 2e^{-t}\cos u + e^{-2t}}\,\mathrm{d}u\,\mathrm{d}t \;= \end{aligned}$$

**Table 3** Inaccuracy inserted into the sum (8.65) applying the estimate (8.68) calculating (8.70)

| $\kappa(\mathsf{A})$ | 1.2 | 3 | 10 | 30 | 100 |
|---|---|---|---|---|---|
| $I(\kappa(\mathsf{A}))$ | $4.990 \times 10^{-2}$ | $1.607 \times 10^{-4}$ | $9.651 \times 10^{-5}$ | $9.710 \times 10^{-6}$ | $8.452 \times 10^{-7}$ |

$$
\begin{aligned}
&= -2\left[\zeta \ln \frac{\cosh\zeta - \cos\xi}{\cosh\zeta - 1} - \int\limits_0^{\zeta} \frac{t\sinh t(\cos\xi - 1)}{(\cosh t - \cos\xi)(\cosh t - 1)}\,\mathrm{d}t\right] \\
&\approx -2\ln\frac{\sqrt{\kappa(\mathsf{A})}+1}{\sqrt{\kappa(\mathsf{A})}-1}\ln\frac{x}{\lambda} + \frac{32}{3}\left(\sqrt{\kappa(\mathsf{A})} - \sqrt{\kappa(\mathsf{A})-1}\right) \\
&-\frac{8}{3}\left(5 + \sqrt{\frac{\Lambda - x}{\Lambda - \lambda}}\right)\frac{\sqrt{x-\lambda}}{\sqrt{\Lambda-\lambda}+\sqrt{\Lambda - x}}\,\mathrm{arccotan}\sqrt{\frac{x}{\lambda} - 1} \\
&-\frac{64}{3}\sqrt{\frac{\Lambda - x}{x-\lambda}}\arctan\frac{\left(\sqrt{\kappa(\mathsf{A})} - \sqrt{\kappa(\mathsf{A})-1}\right)\sqrt{x-\lambda}}{\sqrt{\Lambda-\lambda}+\sqrt{\Lambda-x}}\;; \qquad (8.70)
\end{aligned}
$$

the inserted inaccuracy may be estimated, using the Schwartz' inequality, as

$$
\begin{aligned}
|D| &= 2\int\limits_0^{\zeta}\frac{(\frac{2}{3}\sinh\frac{t}{2} + \frac{8}{3}\tanh\frac{t}{4} - t)\sinh t(1-\cos\xi)}{(\cosh t - \cos\xi)(\cosh t - 1)}\,\mathrm{d}t \\
&= \sin^2\frac{\xi}{2}\int\limits_0^{\zeta}\frac{(\frac{2}{3}\sinh\frac{t}{2} + \frac{8}{3}\tanh\frac{t}{4} - t)\sinh t}{(\sinh^2\frac{t}{2} + \sin^2\frac{\xi}{2})\sinh^2\frac{t}{2}}\,\mathrm{d}t \\
&\leq \sin^2\frac{\xi}{2}\int\limits_0^{\zeta}\frac{(\frac{2}{3}\sinh\frac{t}{2} + \frac{8}{3}\tanh\frac{t}{4} - t)\cosh\frac{t}{2}}{\sinh^2\frac{t}{2}\sin\frac{\xi}{2}}\,\mathrm{d}t \\
&= \sin\frac{\xi}{2}\cdot I(\kappa(\mathsf{A})) = \sqrt{\frac{x-\lambda}{\Lambda-\lambda}}\cdot I(\kappa(\mathsf{A}))\;;
\end{aligned}
$$

values of the integral $I(\kappa(\mathsf{A}))$ are demonstrated in Table 3.

The value $I(\kappa(\mathsf{A}))$ behaves proportionally to $[\kappa(\mathsf{A})]^{-2}$ for sufficiently large $\kappa(\mathsf{A})$, which is still negligible.

The third term in the square bracket in (8.65) may be expressed exactly because of the known fact that the sum $\sum \frac{1}{n^2}\cos nx$ is the trigonometric series representation of the function $\frac{1}{4}(x-\pi)^2 - \frac{\pi^2}{12}$ on the interval $\langle 0, 2\pi\rangle$. Since $\xi \in \langle 0, \pi\rangle$, we can use this result also for the sum containing $\cos^2\xi = \frac{1}{2}(1+\cos 2\xi)$ and consequently

$$
\begin{aligned}
\sum_{n\geq 1}\frac{1}{n^2}(1-\cos(n\xi))^2 &= \frac{\pi^2}{6} - \left[\frac{1}{2}(\xi-\pi)^2 - \frac{\pi^2}{6}\right] + \frac{\pi^2}{12} + \frac{1}{8}(2\xi-\pi)^2 - \frac{\pi^2}{24} \\
&= \frac{\pi}{2}\xi = \frac{\pi}{2}\arccos\frac{\lambda + \Lambda - 2x}{\Lambda - \lambda}\,. \qquad (8.71)
\end{aligned}
$$

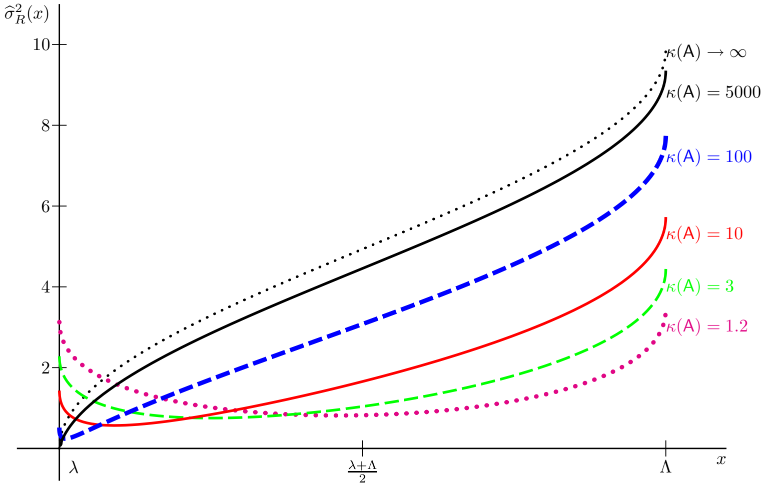

**Fig. 14** Dependence of the function $\widehat{\sigma}_R^2(x)$ on the condition number $\kappa(\mathsf{A})$

We can substitute the results (8.69), (8.70), (8.71) into (8.65); however, the obtained result is very complicated and chaotic. Therefore, we prefer the graphic illustration in Fig. 14. When the matrix $\mathsf{A}$ is sufficiently ill-conditioned ($\kappa(\mathsf{A}) > 10^4$), the contributions of the expressions (8.69), (8.70) to the sum (8.65) are negligible except for the values $x \sim \lambda$; therefore, we can write

$$\widehat{\sigma}_R^2(x) \approx \pi \arccos \frac{\lambda + \Lambda - 2x}{\Lambda - \lambda} .$$

The used method allows the calculation of the correlations of the random variables $R_x$, $R_y$, too. The covariance of these variables may be written in the form analogous to (8.55)

$$\operatorname{cov}(R_x, R_y) = \left\langle \ln \left|1 - \frac{x}{l}\right| \mid \ln \left|1 - \frac{y}{l}\right| \right\rangle_{L_{2,f}(\lambda,\Lambda)} - E_R(x)E_R(y) = \sum_{n\geq 1} c_n(x)c_n(y),$$

and the results (8.61) may be used. Analogously to (8.65)

$$\begin{aligned} \operatorname{cov}(R_x, R_y) = 2\sum_{n\geq 1} \frac{1}{n^2} \Big[ & \left(1 - e^{-n\zeta}\right)^2 - \left(1 - e^{-n\zeta}\right)(1 - \cos n\xi) \\ & - \left(1 - e^{-n\zeta}\right)(1 - \cos n\eta) + (1 - \cos n\xi)(1 - \cos n\eta) \Big] , \end{aligned} \tag{8.72}$$

where

$$\eta = \arccos \frac{\lambda + \Lambda - 2y}{\Lambda - \lambda} .$$

The summation of the first three terms in the square bracket in (8.72) is performed above, the last term may be expressed using the trigonometric sum applied in (8.71). When applying it, we must take into account the fact that the used formula is valid for arguments $x \geq 0$:

$$2\sum_{n\geq 1}\frac{1}{n^2}\left(1-\cos n\xi\right)\left(1-\cos n\eta\right) = \frac{1}{2}(\xi+\eta)\pi - \frac{1}{2}|\xi-\eta|\pi \;=\; \pi\,\min(\xi,\eta)$$
$$= \pi\,\arccos\frac{\lambda+\Lambda-2\min(x,y)}{\Lambda-\lambda}\,.$$

If $\kappa(\mathsf{A}) \gg 1$ and $\min(x,y) \gtrsim \sqrt{\lambda\Lambda}$, this term significantly exceeds the remaining ones in (8.72), and the correlation

$$\varrho(R_x,R_y) \approx \frac{\pi\,\arccos\frac{\lambda+\Lambda-2\min(x,y)}{\Lambda-\lambda}}{\pi\sqrt{\arccos\frac{\lambda+\Lambda-2x}{\Lambda-\lambda}\,\arccos\frac{\lambda+\Lambda-2y}{\Lambda-\lambda}}} = \sqrt{\frac{\arccos\frac{\lambda+\Lambda-2\min(x,y)}{\Lambda-\lambda}}{\arccos\frac{\lambda+\Lambda-2\max(x,y)}{\Lambda-\lambda}}}\,.$$

## Appendix 6: Estimates of the Hessian Matrix Eigenvalues

If $V(\mathsf{x})$ is a quadratic function, then the estimate $\hat{\lambda}^{(j+1)}$ given by (4.4) represents the second derivative of $V(\mathsf{x})$ in the direction $\mathsf{r}_j$ (independently of the point $\mathsf{x}$). In the case of non-quadratic $V(\mathsf{x})$, we can proceed analogously: We will propose a suitable value of the second derivative of $V(\mathsf{x})$ along the $j$th step of the iterative procedure for the estimate $\hat{\lambda}^{(j+1)}$.

Let $V(\mathsf{x}_{j-1})$, $V(\mathsf{x}_j)$ and $\mathsf{g}_{j-1} = \nabla V(\mathsf{x}_{j-1})$, $\mathsf{g}_j = \nabla V(\mathsf{x}_j)$ be given. We define a function of one variable

$$\widehat{V}(t) = V(\mathsf{x}_{j-1} + t\mathsf{p}_j)\,,$$

where $\mathsf{p}_j = \mathsf{x}_j - \mathsf{x}_{j-1} = -\gamma_j \mathsf{g}_{j-1}$. The second derivative of $V$ taken at the point $x_{j-1} + tp_j$ along the direction $\mathsf{p}_j$ is

$$\frac{\partial^2 V(\mathsf{x}_{j-1}+t\mathsf{p}_j)}{\partial\left(\mathsf{p}_j/\|\mathsf{p}_j\|\right)^2} = \frac{1}{\|p_j\|^2}\frac{d^2}{dt^2}V(x_{j-1}+tp_j) = \frac{\widehat{V}''(t)}{\|p_j\|^2}\,. \tag{8.73}$$

The values

$$\widehat{V}(0) = V(\mathsf{x}_{j-1})\,,\;\; \widehat{V}(1) = V(\mathsf{x}_j)\,,\;\; \widehat{V}'(0) = \langle \mathsf{g}_{j-1}|\mathsf{p}_j\rangle\,,\;\; \widehat{V}'(1) = \langle \mathsf{g}_j|\mathsf{p}_j\rangle \tag{8.74}$$

are known; therefore, we can interpolate the function $\widehat{V}(t)$ by a cubic polynomial

$$\widehat{P}(t) = at^3 + bt^2 + ct + d \tag{8.75}$$

where

$$a = \langle \mathsf{g}_j | \mathsf{p}_j \rangle + \langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle - 2\left[ V(\mathsf{x}_j) - V(\mathsf{x}_{j-1}) \right] , \tag{8.76}$$
$$b = 3\left[ V(\mathsf{x}_j) - V(\mathsf{x}_{j-1}) \right] - 2\langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle - \langle \mathsf{g}_j | \mathsf{p}_j \rangle , \tag{8.77}$$
$$c = \langle \mathsf{g}_{j-1} | \mathsf{p}_j \rangle , \tag{8.78}$$

satisfying the relations (8.74). Substituting the polynomial $\widehat{P}(t)$ for $\widehat{V}(t)$ in (8.73) gives

$$\frac{\partial^2 V(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\partial \left( \mathsf{p}_j / \|\mathsf{p}_j\| \right)^2} = \frac{6at + 2b}{\|\mathsf{p}_j\|^2} . \tag{8.79}$$

The considered step $\mathsf{p}_j$ is characterized by the values $t \in \langle 0, 1 \rangle$. In accordance with the considerations performed above, the maximum suitable value provided by (8.79) will be selected. So, depending on the values $a$, $b$, $c$:

1. If $a \leq 0$, then the corresponding maximum is achieved for $t = 0$, so $\hat{\lambda}^{(j+1)} = \frac{2b}{\|\mathsf{p}_j\|^2}$. Nevertheless, the substitutive cubic polynomial $\widehat{P}(t)$ may be decreasing on the whole set $\mathcal{R}$, and in this case, it does not give any information about the behavior of the function $V$ around its potential local minimum. This situation occurs, when the derivative of the polynomial $\widehat{P}(t)$ has at most one real root, i.e.,

$$b^2 - 3ac \leq 0 . \tag{8.80}$$

   In this case, the result is not accepted. If $b \leq 0$, then the function $\widehat{V}(t)$ is not convex in the observed region, and its behavior has no relation to its properties in the neighborhood of its local minimum; therefore, the result is disregarded too.
2. If $a > 0$, then the value (8.79) increases when increasing $t$. Since $\widehat{V}'(0) < 0$ (the vectors $\mathsf{g}_{j-1}$ and $\mathsf{p}_j$ are oppositely oriented), there exists such a $t_m > 0$ that the function $\widehat{V}(t)$ acquires its local minimum for $t = t_m$; evidently, $\widehat{V}''(t_m) > 0$. Now, two situations are possible:
   (a) If $t_m \leq 1$ (which is equivalent to $\widehat{V}'(1) = \langle \mathsf{g}_j | \mathsf{p}_j \rangle \geq 0$), then the value $t = 1$ provides the maximum relevant value $\widehat{V}''(t)$ associated with the $j$th step and $\hat{\lambda}^{(j+1)} = \frac{6a+2b}{\|\mathsf{p}_j\|^2}$.
   (b) If $t_m > 1$, then the step $\mathsf{p}_j$ is too short, and it does not achieve the local minimum of the function $\widehat{V}$. Since the points of minima are in the center of our interest, we prefer to work with the value $\widehat{V}''(t_m) > \widehat{V}''(1)$. The value $t_m$ is the positive root of the equation

$$3at^2 + 2bt + c = 0 ,$$

   thus

$$\hat{\lambda}^{(j+1)} = \frac{6at_m + 2b}{\|\mathsf{p}_j\|^2} = \frac{2\sqrt{b^2 - 3ac}}{\|\mathsf{p}_j\|^2} .$$

As for the estimate $\hat{\Lambda}^{(j+1)}$, the formula (4.4) based on (4.2) represents the second derivative of the function $V$ in the direction $\sqrt{\mathsf{H}}\mathsf{p}_j$ in the case of quadratic function $V(\mathsf{x})$ bounded from below—then, the Hessian matrix $\mathsf{H}$ is positive, and its square root exists. Therefore, we try to estimate the value $\hat{\Lambda}^{(j+1)}$ corresponding to the $j$th step as the minimum relevant value

$$\hat{\Lambda}^{(j+1)} = \frac{\partial^2 V(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\partial(\sqrt{\mathsf{H}}\mathsf{p}_j / \|\sqrt{\mathsf{H}}\mathsf{p}_j\|)^2} = \frac{\|\mathsf{H}(\mathsf{x}_{j-1} + t\mathsf{p}_j)\mathsf{p}_j\|^2}{\langle \mathsf{p}_j | \mathsf{H}(\mathsf{x}_{j-1} + t\mathsf{p}_j)\mathsf{p}_j \rangle} . \tag{8.81}$$

The vector $\mathsf{H}(\mathsf{x}_{j-1} + t\mathsf{p}_j)\mathsf{p}_j$ may be expressed in the form

$$\mathsf{H}(\mathsf{x}_{j-1} + t\mathsf{p}_j)\mathsf{p}_j = \frac{\mathrm{d}\,\mathsf{g}(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\mathrm{d}\,t}. \tag{8.82}$$

We decompose the gradients $\mathsf{g}(\mathsf{x})$ into two parts —the component $\mathsf{g}_{\|}$ parallel to $\mathsf{p}_j$ and the component $\mathsf{g}_{\perp}$ perpendicular to $\mathsf{p}_j$:

$$\mathsf{g}_{\|}(\mathsf{x}_{j-1} + t\mathsf{p}_j) = \frac{\langle \mathsf{g}(\mathsf{x}_{j-1} + t\mathsf{p}_j) | \mathsf{p}_j \rangle}{\|\mathsf{p}_j\|^2}\mathsf{p}_j \ ,$$
$$\mathsf{g}_{\perp}(\mathsf{x}_{j-1} + t\mathsf{p}_j) = \mathsf{g}(\mathsf{x}_{j-1} + t\mathsf{p}_j) - \mathsf{g}_{\|}(\mathsf{x}_{j-1} + t\mathsf{p}_j) \ ,$$

Since we have not enough information about the gradients $\mathsf{g}(\mathsf{x}_{j-1} + t\mathsf{p}_j)$, we must approximate them. For the component $\mathsf{g}_{\|}$, the cubic interpolation (8.75) of the function $\widehat{V}(t)$ is the most accessible one, since

$$\langle \mathsf{g}(\mathsf{x}_{j-1} + t\mathsf{p}_j) | \mathsf{p}_j \rangle = \frac{\mathrm{d}V(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\mathrm{d}\,t} \approx 3at^2 + 2bt + c \ .$$

For the perpendicular component, we have only two available entries —the values $\mathsf{g}_{j-1\perp} = \mathsf{0}$ and $\mathsf{g}_{j\perp}$ corresponding to $t = 1$. Therefore, the linear interpolation

$$\mathsf{g}_{\perp}(\mathsf{x}_{j-1} + t\mathsf{p}_j) \approx t\left[\mathsf{g}_j - \frac{\langle \mathsf{g}_j | \mathsf{p}_j \rangle}{\|\mathsf{p}_j\|^2}\mathsf{p}_j\right]$$

is applied. The derivatives of the particular components of gradients will be

$$\frac{\mathrm{d}\,\mathsf{g}_{\|}(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\mathrm{d}\,t} = \frac{6at + 2b}{\|\mathsf{p}_j\|^2} \cdot \mathsf{p}_j \ , \quad \frac{\mathrm{d}\,\mathsf{g}_{\perp}(\mathsf{x}_{j-1} + t\mathsf{p}_j)}{\mathrm{d}\,t} = \mathsf{g}_j - \frac{\langle \mathsf{g}_j | \mathsf{p}_j \rangle}{\|\mathsf{p}_j\|^2}\mathsf{p}_j \ ,$$

and according to (8.82), we get the interpolation

$$\mathsf{H}(\mathsf{x}_{j-1} + t\mathsf{p}_j)\mathsf{p}_j = \frac{6at + 2b}{\|\mathsf{p}_j\|^2} \cdot \mathsf{p}_j + \mathsf{g}_j - \frac{\langle \mathsf{g}_j | \mathsf{p}_j \rangle}{\|\mathsf{p}_j\|^2}\mathsf{p}_j \ .$$

Using this result in (8.81), we get

$$\begin{aligned}\hat{\Lambda}^{(j+1)} &= \frac{(6at+2b)^2 + \|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}{(6at+2b)\|\mathsf{p}_j\|^2} \\ &= \frac{1}{\|\mathsf{p}_j\|^2} \cdot \left[6at + 2b + \frac{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}{6at+2b}\right] \end{aligned} \tag{8.83}$$

for an appropriate $t$.

If $6at + 2b > 0$, the expression (8.83) acquires its minimal value at

$$6at + 2b = \sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}\,,$$

assuming $\mathsf{g}_j$ not to be parallel to $\mathsf{p}_j$. The choice of the parameter $t$ depends on the values $a, b, c$

1. If $a \leq 0$ and $b \leq 0$, then the value $\hat{\Lambda}^{(j+1)}$ is not proposed.
2. If $a \leq 0$ and $b \in (0, \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}\rangle$, or $a \geq 0$ and $b \geq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}$, then the expression (8.83) increases for $t \geq 0$, so the value $t = 0$ is used in (8.83).
3. If $a > 0$ and $b \leq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}$, or $a < 0$ and $b \geq \frac{1}{2}\sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2}$, then the value

$$t_0 = \frac{1}{6a}\left[\sqrt{\|\mathsf{g}_j\|^2\|\mathsf{p}_j\|^2 - \langle \mathsf{g}_j|\mathsf{p}_j\rangle^2} - 2b\right] \geq 0$$

   is expressed. Now:

   (a) If $t_0 \leq 1$, then the minimal possible value of the expression (8.83) is met in the inside point of the step $\mathsf{p}_j$, so the value $t_0$ is used in (8.83).

   (b) If $t_0 > 1$, then the expression (8.83) is a decreasing function of the parameter $t \in \langle 0, t_0\rangle$; therefore, the selection $t = 1$ is recommended. Nevertheless, if, in addition, $a > 0$, then we can consider also all the values $t \leq t_m$ ($t_m$ is the parameter minimizing the value $\widehat{V}(t)$, see above). The possibility $t_m > 1$ is characterized by the condition $\langle \mathsf{p}_j|\mathsf{g}_j\rangle < 0$, and in this case, the value $t = \min(t_0, t_m)$ is used in (8.83).
4. Analogously to the calculation of the value $\hat{\lambda}^{(j+1)}$, the results are disregarded, when the condition (8.80) is satisfied.

The expression of the values $\hat{\lambda}^{(j+1)}$, $\hat{\Lambda}^{(j+1)}$ is impossible, when $a \leq 0$ and $b \leq 0$, or in the case (8.80); in all the remaining situations, the estimates are available. The inaccessibility of the values $\hat{\lambda}^{(j+1)}$, $\hat{\Lambda}^{(j+1)}$ may be solved by two different ways:

1. In the first step ($j = 0$), we need initialize the estimates $\overline{\lambda}^{(1)}$, $\underline{\Lambda}^{(1)}$. First, we propose a step length parameter $\gamma_1$. If the unwanted situation mentioned above occurs, the easiest way to get some results is to increase the value $\gamma_1$. Really, if $\mathsf{p}_1 = -\gamma_1\mathsf{g}_0$, then adding of the Eqs. (8.76) and (8.77) results in

$$a + b = V(\mathsf{x}_1) - V(\mathsf{x}_0) + \gamma_1 \|\mathsf{g}_0\|^2 \geq V_0 - V(\mathsf{x}_0) + \gamma_1 \|\mathsf{g}_0\|^2 \,, \tag{8.84}$$

since the function $V(\mathsf{x})$ is assumed to be bounded from below by a value $V_0$. Therefore, if $\mathsf{g}_0 \neq \mathsf{0}$, then it exists a $\gamma_1 > 0$ for which the expression on the right-hand side of (8.84) is positive, which implies $a > 0$ or $b > 0$. However, the possibility (8.80) is not yet excluded. The requirement opposite to (8.80) may be transformed, using (8.76), (8.77), (8.78), into the form

$$\begin{aligned} \langle \mathsf{g}_1 | \mathsf{p}_1 \rangle^2 - \left[ 6\,(V(\mathsf{x}_1) - V(\mathsf{x}_0)) + \gamma_1 \|\mathsf{g}_0\|^2 \right] \langle \mathsf{g}_1 | \mathsf{p}_1 \rangle \\ + \left[ 3\,(V(\mathsf{x}_1) - V(\mathsf{x}_0)) + \gamma_1 \|\mathsf{g}_0\|^2 \right]^2 > 0 \,. \end{aligned}$$

This inequality is satisfied for any value $\langle \mathsf{g}_1 | \mathsf{p}_1 \rangle$,[14] when its discriminant is negative; thus, for $\gamma_1 > 0$,

$$4\,(V(\mathsf{x}_1) - V(\mathsf{x}_0)) + \gamma_1 \|\mathsf{g}_0\|^2 > 0 \,; \tag{8.85}$$

this holds, when

$$\gamma_1 > \frac{4\,(V(\mathsf{x}_0) - V_0)}{\|\mathsf{g}_0\|^2} \,. \tag{8.86}$$

The condition (8.86) is stronger than the condition following from (8.84); therefore, for any $\gamma_1$ satisfying (8.86), the initial estimates $\overline{\lambda}^{(1)} = \hat{\lambda}^{(1)} > 0$, $\underline{\Lambda}^{(1)} = \hat{\Lambda}^{(1)} > 0$ exist. The value $V_0$ is, however, unknown in general; nevertheless, all our requirements guaranteeing the existence of estimates $\hat{\lambda}^{(1)} > 0$, $\hat{\Lambda}^{(1)} > 0$ are met, as the condition (8.85) is satisfied.

2. For $j \geq 1$, the process described above may be used too, or simply the results for $\hat{\lambda}^{(j+1)}$, $\hat{\Lambda}^{(j+1)}$ are not employed.

Since the maximum possible estimate of the minimal eigenvalue of the Hessian matrix is selected, and the estimate of the maximal eigenvalue is minimized, the possibility $\hat{\lambda}^{(j+1)} > \hat{\Lambda}^{(j+1)}$ is not excluded. In general, the use of shorter steps in the steepest descent method causes less damage, so in the nonsense case $\overline{\lambda}^{(j)} > \underline{\Lambda}^{(j)}$, we prefer the application of the value $l_j = \overline{\lambda}^{(j)}$; this is supported also by the use of a finer interpolation during its calculation.

## References

1. H. Akaike: On a successive transformation of probability distribution and its application to the analysis of the optimum gradient method. Ann. Inst. Statist. Math. Tokyo 11 (1959), 1 – 16
2. J. Barzilai and J. M. Borwein: Two-point step size gradient methods. IMA Journal of Numerical Analysis 8 (1988), 141 – 148
3. A. Cauchy: Méthode génerale pour la résolution des systèmes d' équations simultanées. Comptes Rendus Hebd. Séances Acad. Sci. 25 (1847), 536–538,

[14] We refuse to waste time calculating the gradients $\mathsf{g}(\mathsf{x}_0 - \gamma_1 \mathsf{g}_0)$ until it is needed for the following step creation.

4. Y. H. Dai and R. Fletcher: Projected Barzilai-Borwein methods for large-scale box-constrained quadratic programming. Numerische Mathematik 100 (2005), No. 1, 21 – 47
5. Y. H. Dai and L. Z. Liao: $R$-linear convergence of the Barzilai-Borwein gradient method. IMA Journal of Numerical Analysis vol. 22 (2002), 1 – 10
6. R. Fletcher and C. M. Reeves: Function minimization by conjugate gradients, The Computer Journal 1964, 149 – 154
7. G. E. Forsythe: On the asymptotic directions of the $s$-dimensional optimum gradient method. Numerische Mathematik 11 (1968), 57 – 76
8. A. Friedlander, J. M. Martínez, N. Molina and M. Raydan: Gradient method with retards and generalizations. SIAM J. Numer. Anal. 36 (1999), 275 – 289
9. Kalousek, Z.: Appeal of inexact calculations in Proceedings of conference “Modern mathematical methods in engineering”, VŠB-TU Ostrava, 2013
10. F. Luengo and M. Raydan: Gradient method with dynamical retards for large-scale optimization problems. Electronic Transactions on Numerical Analysis, 16 (2003), 186 – 193
11. Y. Narushima, T. Wakamatsu and H. Yabe: Extended Barzilai-Borwein method for unconstrained minimization problems. Pacific journal of optimization 6 (2010), 591 - 613
12. L. Pronzato, P. Wynn and A. A. Zhigljavsky: A dynamical-system analysis of the optimum $s$-gradient algorithm in Optimal design and related areas in optimization and statistics (editors L. Pronzato and A. A. Zhigljavsky) , Springer, 2009, pp. 39 – 80
13. L. Pronzato and A. A. Zhigljavsky: Gradient algorithms for quadratic optimization with fast convergence rates. Computational Optimization and Applications 50 (2011), 597 – 617
14. M. Raydan: The Barzilai and Borwein gradient method for the large scale unconstrained minimization problem. SIAM Journal on Optimization 7 (1997), 26–33
15. M. Raydan and B. F. Svaiter: Relaxed steepest descent and Cauchy-Barzilai-Borwein method. Computational Optimization and Applications 21 (2002), 155 – 167