Journal of Intelligent Algorithms and Scientific Computing
J. Intell. Algor. Sci. Comput., Vol. **1**, No. 1, pp. 1-23

DOI: 10.4208/jiasc.2026-72-1
xxx 2026

Convergence Analysis of Block Newton Methods for 1D Shallow Neural Network Approximation

Zhiqiang Cai[1,*], Anastassia Doktorova[1], Robert D. Falgout[2] and César Herrera[1]

[1] *Department of Mathematics, Purdue University, West Lafayette, IN 47907-2067, USA*
[2] *Livermore National Laboratory, Livermore, CA 94550-9698, USA*

Received 13 February 2026; Accepted (in revised version) 18 March 2026

Abstract. This paper analyzes local convergence of the block Newton (BN) method introduced in [5,6] for one-dimensional shallow neural network approximation to functions and diffusion-reaction problems. The BN method consists of the 2×2 block nonlinear Gauss-Seidel, linear Gauss-Seidel, or Jacobi method for outer iteration and the Newton method for inner iteration. The blocks are corresponding to the linear and the nonlinear parameters. Under some reasonable assumptions, we establish local convergence of the BN methods as well as the reduced BN (rBN) method for one-dimensional diffusion-reaction problems and least-squares function approximation. Unlike common optimization methods, the rBN allows for the reduction of the number of parameters during the optimization process when some neurons contribute little to the approximation or are at nearly optimal locations.

AMS subject classifications: to be provided by authors

Key words: Neural network, Elliptic problems, Least-Squares approximation, Newton's method, Local convergence analysis.

1 Introduction

One dimensional ReLU shallow neural network (NN) with n neurons generates a set of continuous piecewise linear functions. Specifically, the set with restriction of the biases in the interval $I=(0,1)$ is given by

$$\mathcal{M}_n(I)=\left\{\alpha+\sum_{i=0}^{n}c_i\sigma(x-b_i):\alpha\in\mathbb{R},c_i\in\mathbb{R},0=b_0<b_1<\cdots<b_n<b_{n+1}=1\right\},$$

*Corresponding author.
Email addresses: caiz@purdue.edu (Z. Cai), adoktoro@purdue.edu (A. Doktorova), rfalgout@llnl.gov (R. Falgout), herre125@purdue.edu (C. Herrera)

where $\sigma(t)=\max\{0,t\}$ is the ReLU activation function. Denote by $\mathbf{c}=(c_0,\cdots,c_n)^T\in\mathbb{R}^{n+1}$ and $\mathbf{b}=(b_1,\cdots,b_n)^T\in\mathbb{R}^n$ the respective linear parameters and nonlinear parameters, and denote by $\boldsymbol{\theta}=(c_0,c_1,\cdots,c_n,b_1,\cdots,b_n)^T\in\mathbb{R}^{2n+1}$ all parameters. Notice that the weights of all hidden layer is chosen to be one by normalization (see [11]). Each function in $\mathcal{M}_n(I)$

$$v(x;\mathbf{c},\mathbf{b})=\alpha+\sum_{i=0}^{n}c_i\sigma(x-b_i)$$

depends on the parameters $\mathbf{c}$ and $\mathbf{b}$ and is a piecewise linear function of $x\in[0,1]$ with respect to the partition by the breaking (mesh) points: $0=b_0<b_1<\cdots<b_n<b_{n+1}=1$.

It is known (see, e.g., [11]) that the set $\mathcal{M}_n(I)$ is equivalent to the free-knot splines (FKS) (see [16]), where FKS utilizes local hat basis functions. FKS can substantially enhance the approximation order and reduce the number of degrees of freedom for non-smooth functions (see [3] and the discussion in [5]). As an example of this, the order of the best approximation to $f(x)=x^\alpha$ $(0<\alpha<1)$ on I is merely $\alpha<1$ (i.e., $\mathcal{O}(n^{-\alpha})$) when using finite elements on a fixed uniform mesh, whereas for FKS, the order becomes one (i.e., $\mathcal{O}(n^{-1})$) no matter how small the exponent $\alpha>0$ is (see, e.g., [8,16]). This is a huge improvement in approximation.

Despite the remarkable approximation capability of FKS for non-smooth functions, there are two essential difficulties that have led numerical analysts moving away from FKS: (1) no successful extension of FKS to two or higher dimensions has been achieved, and (2) determining the optimal knot locations (the nonlinear parameters $\mathbf{b}$) results in a high-dimensional, non-convex optimization problem, that is computationally expensive and hence dismisses its benefit in approximation. To the best of our knowledge, there are no available efficient optimization schemes that enable the competitiveness of FKS. While the first issue may be addressed by employing neural networks due to their global supported basis functions, the second issue still remains a major challenge.

To address this challenge, a major advance on fast iterative solver was recently made in [5, 6] for solving non-convex optimization problems arising from shallow ReLU NN approximation to a given function or solutions of elliptic differential equations in one dimension. Specifically, a well-designed damped block Newton (dBN) method was developed. First, the dBN adopts a classical outer-inner iterative strategy (see, e.g., [1, 2, 7, 10, 12, 14, 15]), alternating between updates of the linear and nonlinear parameters. Second, to solve the resulting dense, ill-conditioned linear systems due to the global basis functions of NNs, the dBN uses the fact that the exact inversion of those matrices can be represented in terms of products of sparse matrices (see [6]). Third, the dBN deals directly with singularities of the Hessian for the nonlinear parameters by removing neurons whose linear parameters are small or whose nonlinear parameters have reached nearly optimal locations. As a result, the computational cost per iteration of the dBN is $\mathcal{O}(n)$, and numerical experiments show that this method is capable of moving mesh points effectively and efficiently. Beyond strong one dimensional results, the methodology of the dBN is conceptually promising for higher dimensions, as it demonstrates how to design

iterative solvers that exploit the problem structure and the approximation and geometric meanings of the NN parameters.

The purpose of this paper is to provide a theoretical guarantee on why this sophisticated dBN moves the mesh points efficiently. This will be done by analyzing local convergence of block Newton (BN) methods. The BN method consists of the 2×2 block nonlinear Gauss-Seidel, linear Gauss-Seidel, or Jacobi method for outer iteration and the Newton method for inner iteration. The blocks are corresponding to the linear and the nonlinear parameters. By following the machinery in [13] on local convergence of the componentwise Gauss-Seidel method, we first develop a local convergence theory for the block Newton methods applicable to shallow ReLU NNs in both one and multiple dimensions, provided that the Hessian matrix at a critical point is symmetric positive definite (SPD) and that the 2×2 block nonlinear Gauss-Seidel, linear Gauss-Seidel, or Jacobi matrix is invertible.

By expressing the BN method as a fixed-point iteration, local convergence of the BN method is established by showing that a norm of the corresponding Jacobian at the critical point is strictly less than one. Note that derivation of the Jacobian matrix is non-trivial (see Lemma 3.1). To guarantee feasibility of each Newton step, the BN method is modified to the reduced BN (rBN) method that allows a reduction in the number of parameters during the optimization process. Local convergence of the rBN is justified by showing that some nonlinear parameters are at nearly optimal locations.

The paper is structured as follows. Section 2 introduces BN methods with three different outer iteration methods. Local convergence analysis of these methods are presented in Section 3. Sufficient conditions on SPD of the Hessian for one-dimensional problems are derived in Section 4. Section 4.4 discusses the reduced Block Newton method and some conclusions and remarks are presented in Section 5.

2 Block Newton methods

Let $\boldsymbol{\theta}=\begin{pmatrix}\mathbf{c}\\ \mathbf{b}\end{pmatrix}$, where $\mathbf{c}\in\mathbb{R}^{n+1}$ and $\mathbf{b}\in\mathbb{R}^n$. Given an open set $\mathcal{D}\subseteq\mathbb{R}^{2n+1}$ and a twice continuously differentiable function $F=F(\boldsymbol{\theta})=F(\mathbf{c},\mathbf{b}):\mathcal{D}\to\mathbb{R}$, we aim to find a minimizer $\boldsymbol{\theta}^*=\begin{pmatrix}\mathbf{c}^*\\ \mathbf{b}^*\end{pmatrix}\in\mathcal{D}$ such that

$$F(\boldsymbol{\theta}^*)=\min_{\boldsymbol{\theta}\in\mathcal{D}}F(\boldsymbol{\theta}).$$

Optimality conditions imply that $\boldsymbol{\theta}^*$ satisfies the system of nonlinear algebraic equations

$$\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\theta})=\begin{pmatrix}\nabla_{\mathbf{c}}F(\boldsymbol{\theta})\\ \nabla_{\mathbf{b}}F(\boldsymbol{\theta})\end{pmatrix}=\begin{pmatrix}\mathbf{0}\\ \mathbf{0}\end{pmatrix}, \tag{2.1}$$

where $\nabla_{\mathbf{c}}$ and $\nabla_{\mathbf{b}}$ denote the gradients of $F(\boldsymbol{\theta})$ with respect to $\mathbf{c}$ and $\mathbf{b}$, respectively. The Hessian matrix is given by

$$\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \begin{pmatrix} \nabla^2_{\mathbf{cc}} F(\boldsymbol{\theta}) & \nabla^2_{\mathbf{cb}} F(\boldsymbol{\theta}) \\ \nabla^2_{\mathbf{bc}} F(\boldsymbol{\theta}) & \nabla^2_{\mathbf{bb}} F(\boldsymbol{\theta}) \end{pmatrix} = \begin{pmatrix} \mathcal{H}_{11}(\boldsymbol{\theta}) & \mathcal{H}_{12}(\boldsymbol{\theta}) \\ \mathcal{H}_{21}(\boldsymbol{\theta}) & \mathcal{H}_{22}(\boldsymbol{\theta}) \end{pmatrix} \in \mathbb{R}^{(2n+1)\times(2n+1)},$$

where $\mathcal{H}_{ij}(\boldsymbol{\theta})$ for $i,j=1,2$ are given by

$$\mathcal{H}_{11}(\boldsymbol{\theta}) = \nabla^2_{\mathbf{cc}} F(\boldsymbol{\theta}), \quad \mathcal{H}_{12}(\boldsymbol{\theta}) = \nabla^2_{\mathbf{cb}} F(\boldsymbol{\theta}), \quad \mathcal{H}_{21}(\boldsymbol{\theta}) = \nabla^2_{\mathbf{bc}} F(\boldsymbol{\theta}), \quad \mathcal{H}_{22}(\boldsymbol{\theta}) = \nabla^2_{\mathbf{bb}} F(\boldsymbol{\theta}).$$

The nonlinear system in Eq. (2.1) can be solved using Newton's method, though this may be computationally expensive. To reduce cost, we can use a block Newton (BN) method by performing an outer-inner iteration that alternates between the variables $\mathbf{c}$ and $\mathbf{b}$. For the outer iteration, one may use a Gauss–Seidel or Jacobi scheme, and apply a Newton iteration to each block during the inner solve.

More specifically, let $(\mathbf{c}^{(k)},\mathbf{b}^{(k)})$ denote the current iterate. Then the block nonlinear Gauss-Seidel (see, e.g., [13]) method, as the outer iteration, computes the new iterate $(\mathbf{c}^{(k+1)},\mathbf{b}^{(k+1)})$ as follows:

1. Update the variable $\mathbf{c}^{(k+1)}$:

$$\mathbf{c}^{(k+1)} = \mathbf{c}^{(k)} - \left[\mathcal{H}_{11}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right)\right]^{-1} \nabla_{\mathbf{c}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right).$$

2. Update the variable $\mathbf{b}^{(k+1)}$:

$$\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)} - \left[\mathcal{H}_{22}\left(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)}\right)\right]^{-1} \nabla_{\mathbf{b}} F\left(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)}\right).$$

In other words, the new iterate $(\mathbf{c}^{(k+1)},\mathbf{b}^{(k+1)})$ is the solution of the block diagonal system of nonlinear algebraic equations

$$\begin{pmatrix} \mathcal{H}_{11}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_{22}\left(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)}\right) \end{pmatrix} \begin{pmatrix} \mathbf{c}^{(k+1)} - \mathbf{c}^{(k)} \\ \mathbf{b}^{(k+1)} - \mathbf{b}^{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla_{\mathbf{c}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \\ \nabla_{\mathbf{b}} F\left(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)}\right) \end{pmatrix}, \quad (2.2)$$

which can be solved sequentially by computing two systems of linear algebraic equations.

This nonlinear Gauss-Seidel method differs from the classical linear Gauss–Seidel method, where the new iterate is obtained by solving the block lower-triangular system of linear equations

$$\begin{pmatrix} \mathcal{H}_{11}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) & \mathbf{0} \\ \mathcal{H}_{21}(\mathbf{c}^{(k)},\mathbf{b}^{(k)}) & \mathcal{H}_{22}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \end{pmatrix} \begin{pmatrix} \mathbf{c}^{(k+1)} - \mathbf{c}^{(k)} \\ \mathbf{b}^{(k+1)} - \mathbf{b}^{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla_{\mathbf{c}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \\ \nabla_{\mathbf{b}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \end{pmatrix}. \quad (2.3)$$

On the other hand, in the Jacobi method, the new iterate $\left(\mathbf{c}^{(k+1)},\mathbf{b}^{(k+1)}\right)$ solves the block diagonal system of linear algebraic equations

$$\begin{pmatrix} \mathcal{H}_{11}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_{22}\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \end{pmatrix} \begin{pmatrix} \mathbf{c}^{(k+1)}-\mathbf{c}^{(k)} \\ \mathbf{b}^{(k+1)}-\mathbf{b}^{(k)} \end{pmatrix} = - \begin{pmatrix} \nabla_{\mathbf{c}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \\ \nabla_{\mathbf{b}} F\left(\mathbf{c}^{(k)},\mathbf{b}^{(k)}\right) \end{pmatrix}. \tag{2.4}$$

3 Convergence analysis

For brevity, we refer to the schemes defined in (2.2), (2.3), and (2.4) as NL-GS, L-GS, and JB, respectively. This section presents analytic tools for deriving local convergence conditions for the BN methods introduced in the previous section. To this end, we introduce the following assumption:

Invertibility Assumption: there exists an open set $\mathcal{O}\subseteq\mathcal{D}$ such that $\mathcal{H}_{11}(\boldsymbol{\theta})$ is invertible; moreover, $\mathcal{H}_{22}(G_1(\boldsymbol{\theta}),\mathbf{b})$ is invertible for NL-GS and $\mathcal{H}_{22}(\boldsymbol{\theta})$ is invertible for L-GS and JB.

Define the mapping

$$G(\boldsymbol{\theta}) = \begin{pmatrix} G_1(\boldsymbol{\theta}) \\ G_2(\boldsymbol{\theta}) \end{pmatrix},$$

where $G_1:\mathcal{D}\to\mathbb{R}^{n+1}$ and $G_2:\mathcal{D}\to\mathbb{R}^n$ are respectively given by

$$G_1(\boldsymbol{\theta}) = \mathbf{c} - \mathcal{H}_{11}^{-1}(\boldsymbol{\theta})\nabla_{\mathbf{c}} F(\boldsymbol{\theta}), \tag{3.1}$$

and by

$$G_2(\boldsymbol{\theta}) = \begin{cases} \mathbf{b} - \mathcal{H}_{22}^{-1}(G_1(\boldsymbol{\theta}),\mathbf{b})\nabla_{\mathbf{b}} F(G_1(\boldsymbol{\theta}),\mathbf{b}) & \text{for } NL-GS, \\ \mathbf{b} - \mathcal{H}_{22}^{-1}(\boldsymbol{\theta})\left(\nabla_{\mathbf{b}} F(\boldsymbol{\theta}) + \mathcal{H}_{21}(\boldsymbol{\theta})\left(G_1(\boldsymbol{\theta})-\mathbf{c}\right)\right) & \text{for } L-GS, \\ \mathbf{b} - \mathcal{H}_{22}^{-1}(\boldsymbol{\theta})\nabla_{\mathbf{b}} F(\boldsymbol{\theta}), & \text{for } JB. \end{cases} \tag{3.2}$$

Then the BN methods can be expressed as the fixed-point iteration of $G(\boldsymbol{\theta})$, i.e.,

$$\boldsymbol{\theta}^{k+1} = G(\boldsymbol{\theta}^k) \quad \text{for } k\in\mathbb{N}=\{0,1,2,\cdots\}. \tag{3.3}$$

Denote by $\mathbf{J}_G(\boldsymbol{\theta})\in\mathbb{R}^{(2n+1)\times(2n+1)}$ the Jacobian matrix of G at $\boldsymbol{\theta}$. A sufficient condition for local convergence of the fixed-point iteration in (3.3) was given in Theorem 10.1.3 of [13]. For the convenience of readers, we state it below and provide its proof.

Theorem 3.1 (Ostroswki). *Suppose that $G:\mathcal{O}\to\mathbb{R}^{2n+1}$ has a fixed point $\boldsymbol{\theta}^*\in\mathcal{O}$ and that the mapping G is differentiable at $\boldsymbol{\theta}^*$. Denote by $\|\cdot\|$ a norm in $\mathbb{R}^{2n+1}$, if $\|\mathbf{J}_G(\boldsymbol{\theta}^*)\|=\sigma<1$, then the fixed-point iteration in* (3.3) *converges locally to $\boldsymbol{\theta}^*$.*

Proof. For any $0<\epsilon<1-\sigma$, the assumption on the differentiability of G at $\boldsymbol{\theta}^*$ implies that there exists a $\delta>0$ neighbourhood centered at $\boldsymbol{\theta}^*$, $\mathcal{B}(\boldsymbol{\theta}^*;\delta)=\{\boldsymbol{\theta}\in\mathbb{R}^{2n+1}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta\}\subset\mathcal{O}$, such that

$$\|G(\boldsymbol{\theta})-[G(\boldsymbol{\theta}^*)+\mathbf{J}_G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)]\|\leq\epsilon\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \quad \text{for all} \quad \boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*;\delta). \tag{3.4}$$

For any $k\in\mathbb{N}$, the assumption that $\boldsymbol{\theta}^*=G(\boldsymbol{\theta}^*)$ and (3.3) give

$$\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^*=G(\boldsymbol{\theta}^k)-G(\boldsymbol{\theta}^*)=\left(G(\boldsymbol{\theta}^k)-\left[G(\boldsymbol{\theta}^*)+\mathbf{J}_G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^k-\boldsymbol{\theta}^*)\right]\right)+\mathbf{J}_G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}^k-\boldsymbol{\theta}^*),$$

which, together with the triangle inequality, the assumption that $\boldsymbol{\theta}^k\in\mathcal{B}(\boldsymbol{\theta}^*;\delta)$, and (3.4), implies

$$\|\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^*\|\leq(\epsilon+\sigma)\|\boldsymbol{\theta}^k-\boldsymbol{\theta}^*\|<\|\boldsymbol{\theta}^k-\boldsymbol{\theta}^*\|.$$

If the initial $\boldsymbol{\theta}^0$ belongs to $\mathcal{B}(\boldsymbol{\theta}^*;\delta)$, then the second inequality implies that $\boldsymbol{\theta}^k\in\mathcal{B}(\boldsymbol{\theta}^*;\delta)$ by induction. Hence, we have

$$\|\boldsymbol{\theta}^{k+1}-\boldsymbol{\theta}^*\|\leq(\epsilon+\sigma)^k\|\boldsymbol{\theta}^0-\boldsymbol{\theta}^*\|,$$

which proves the theorem. □

Next, let

$$B(\boldsymbol{\theta})=\begin{pmatrix}\mathcal{H}_{11}(\boldsymbol{\theta}) & \mathbf{0}\\ \mathcal{H}_{21}(\boldsymbol{\theta}) & \mathcal{H}_{22}(\boldsymbol{\theta})\end{pmatrix} \quad \text{and} \quad B_1(\boldsymbol{\theta})=\begin{pmatrix}\mathcal{H}_{11}(\boldsymbol{\theta}) & \mathbf{0}\\ \mathbf{0} & \mathcal{H}_{22}(G_1(\boldsymbol{\theta}),\mathbf{b})\end{pmatrix}.$$

Assume that $\boldsymbol{\theta}^*$ is a minimizer of $F(\boldsymbol{\theta})$, then Taylor's expansion of $\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\theta})$ at $\boldsymbol{\theta}^*$ gives

$$\nabla_{\boldsymbol{\theta}}F(\boldsymbol{\theta})=\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)+R(\boldsymbol{\theta};\boldsymbol{\theta}^*), \tag{3.5}$$

where $R(\boldsymbol{\theta};\boldsymbol{\theta}^*)$ is the remainder satisfying

$$\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\|R(\boldsymbol{\theta};\boldsymbol{\theta}^*)\|/\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|=0.$$

Similarly, expanding $\nabla_{\mathbf{b}}F(G_1(\boldsymbol{\theta}),\mathbf{b})$ about $\boldsymbol{\theta}$ and using (3.1), we have

$$\nabla_{\mathbf{b}}F(G_1(\boldsymbol{\theta}),\mathbf{b})=\nabla_{\mathbf{b}}F(\boldsymbol{\theta})-\mathcal{H}_{21}(\boldsymbol{\theta})\mathcal{H}_{11}^{-1}(\boldsymbol{\theta})\nabla_{\mathbf{c}}F(\boldsymbol{\theta})+\tilde{R}(\boldsymbol{\theta}),$$

where the remainder $\tilde{R}(\boldsymbol{\theta})$ satisfies

$$\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\frac{\|\tilde{R}(\boldsymbol{\theta})\|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|}=0.$$

Let

$$B_2(\boldsymbol{\theta}) = \begin{pmatrix} I & \mathbf{0} \\ -\mathcal{H}_{21}(\boldsymbol{\theta})\mathcal{H}_{11}^{-1}(\boldsymbol{\theta}) & I \end{pmatrix},$$

then combining with (3.5) yields

$$\begin{pmatrix} \nabla_{\mathbf{c}} F(\boldsymbol{\theta}) \\ \nabla_{\mathbf{b}} F(G_1(\boldsymbol{\theta}),\mathbf{b}) \end{pmatrix} = B_2(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) + \begin{pmatrix} \mathbf{0} \\ \tilde{R}(\boldsymbol{\theta}) \end{pmatrix} = B_2(\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) + \hat{R}(\boldsymbol{\theta};\boldsymbol{\theta}^*), \quad (3.6)$$

where

$$\hat{R}(\boldsymbol{\theta};\boldsymbol{\theta}^*) = B_2(\boldsymbol{\theta})R(\boldsymbol{\theta};\boldsymbol{\theta}^*) + \begin{pmatrix} \mathbf{0} \\ \tilde{R}(\boldsymbol{\theta}) \end{pmatrix}.$$

The next lemma provides a formula for $\mathbf{J}_G(\boldsymbol{\theta}^*)$ when $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$ is symmetric and positive definite (SPD) for Gauss–Seidel schemes. For the componentwise Gauss–Seidel method, this result was proved in Theorem 10.3.3 of [13]. Here, we use block Gauss–Seidel methods.

Lemma 3.1. *Let $G:\mathcal{O}\to\mathbb{R}^{2n+1}$ be the mapping defined in* (3.2) *for the L-GS or NL-GS, and let $\boldsymbol{\theta}^*\in\mathcal{O}$ be a fixed point of G. Assume that $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$ is SPD. Then*

$$\mathbf{J}_G(\boldsymbol{\theta}^*) = I_{2n+1} - B^{-1}(\boldsymbol{\theta}^*)\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*), \quad (3.7)$$

where I_{2n+1} is the order-$(2n+1)$ identity matrix.

Proof. By definition of differentiability, to show the validity of (3.7), it suffices to show that if $\mathbf{J}_G(\boldsymbol{\theta}^*)$ is given in (3.7), then

$$\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*} \frac{\|G(\boldsymbol{\theta}) - [G(\boldsymbol{\theta}^*) + \mathbf{J}_G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)]\|}{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|} = 0. \quad (3.8)$$

To this end, for $\mathbf{J}_G(\boldsymbol{\theta}^*)$ given in (3.7), we have

$$B(\boldsymbol{\theta}^*)\mathbf{J}_G(\boldsymbol{\theta}^*) = \begin{pmatrix} \mathbf{0} & -\mathcal{H}_{12}(\boldsymbol{\theta}^*) \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \quad \text{and} \quad B(\boldsymbol{\theta}^*) - B(\boldsymbol{\theta}^*)\mathbf{J}_G(\boldsymbol{\theta}^*) = \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*),$$

which, together with the assumption that $\boldsymbol{\theta}^*$ is a fixed-point of $G(\boldsymbol{\theta})$, implies

$$\begin{aligned} \mathbf{a}(\boldsymbol{\theta},\boldsymbol{\theta}^*) &\equiv B(\boldsymbol{\theta}^*)\Big(G(\boldsymbol{\theta}) - \big[G(\boldsymbol{\theta}^*) + \mathbf{J}_G(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\big]\Big) \\ &= B(\boldsymbol{\theta}^*)\Big(G(\boldsymbol{\theta}) - \boldsymbol{\theta}\Big) + \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*). \end{aligned}$$

For the L-GS, (3.5) gives

$$G(\boldsymbol{\theta}) - \boldsymbol{\theta} = -B^{-1}(\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = -B^{-1}(\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*) - B^{-1}(\boldsymbol{\theta})R(\boldsymbol{\theta};\boldsymbol{\theta}^*),$$

which implies

$$\mathbf{a}(\boldsymbol{\theta},\boldsymbol{\theta}^*)=\Big(B(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)\Big)B^{-1}(\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)-B(\boldsymbol{\theta}^*)B^{-1}(\boldsymbol{\theta})R(\boldsymbol{\theta};\boldsymbol{\theta}^*). \tag{3.9}$$

For the NL-GS, (3.6) leads to

$$G(\boldsymbol{\theta})-\boldsymbol{\theta}=-B_1^{-1}(\boldsymbol{\theta})B_2(\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)-B_1^{-1}(\boldsymbol{\theta})\hat{R}(\boldsymbol{\theta};\boldsymbol{\theta}^*),$$

which, together with

$$B_3(\boldsymbol{\theta})\equiv B_2^{-1}(\boldsymbol{\theta})B_1(\boldsymbol{\theta})=\begin{pmatrix}\mathcal{H}_{11}(\boldsymbol{\theta}) & \mathbf{0}\\ \mathcal{H}_{21}(\boldsymbol{\theta}) & \mathcal{H}(G_1(\boldsymbol{\theta}),\mathbf{b})\end{pmatrix},$$

yields

$$\mathbf{a}(\boldsymbol{\theta},\boldsymbol{\theta}^*)=\Big(B_3(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)\Big)B_1^{-1}(\boldsymbol{\theta})B_2(\boldsymbol{\theta})\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)(\boldsymbol{\theta}-\boldsymbol{\theta}^*)-B(\boldsymbol{\theta}^*)B_1^{-1}(\boldsymbol{\theta})\hat{R}(\boldsymbol{\theta};\boldsymbol{\theta}^*). \tag{3.10}$$

By the assumptions that $F(\boldsymbol{\theta})$ is twice continuously differentiable and that $\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)$ is SPD, there exists a $\delta>0$ neighbourhood centered at $\boldsymbol{\theta}^*$, $\mathcal{B}(\boldsymbol{\theta}^*;\delta)=\{\boldsymbol{\theta}\in\mathbb{R}^{2n+1}:\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|<\delta\}\subset\mathcal{O}$, such that $\|B^{-1}(\boldsymbol{\theta})\|$ and $\|B_1^{-1}(\boldsymbol{\theta})\|$ are bounded for all $\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}^*;\delta)$ and that

$$\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\Big(B(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)\Big)=\mathbf{0}\quad\text{and}\quad\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\Big(B_3(\boldsymbol{\theta})-B(\boldsymbol{\theta}^*)\Big)=\mathbf{0}.$$

Now, (3.8) is a direct consequence of (3.9), (3.10), the triangle inequality, and the facts that

$$\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\|R(\boldsymbol{\theta};\boldsymbol{\theta}^*)\|/\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|=0\quad\text{and}\quad\lim_{\boldsymbol{\theta}\to\boldsymbol{\theta}^*}\|\hat{R}(\boldsymbol{\theta};\boldsymbol{\theta}^*)\|/\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\|=0.$$

This completes the proof of the lemma. □

The following lemma was stated and proved in [9]. For convenience of readers, a brief proof is provided.

Lemma 3.2. *Assume that $A\in\mathbb{R}^{m\times m}$ is SPD and that $M\in\mathbb{R}^{m\times m}$ is invertible. Then the matrix $M+M^T-A$ is SPD if and only if*

$$\|I_{2n+1}-M^{-1}A\|_A<1,$$

where $\|\mathbf{v}\|_A=\sqrt{\langle A\mathbf{v},\mathbf{v}\rangle}$ and $\langle\cdot,\cdot\rangle$ is the standard inner product in $\mathbb{R}^m$.

Proof. For any $\mathbf{v}\in\mathbb{R}^m$, it is easy to check that

$$\|(I-M^{-1}A)\mathbf{v}\|_A^2=\langle A\mathbf{v},\mathbf{v}\rangle-\Big\langle(M+M^T-A)(M^{-1}A)\mathbf{v},(M^{-1}A)\mathbf{v}\Big\rangle,$$

which implies the validity of the lemma. □

Now, we are ready to present a sufficient condition for local convergence of the fixed-point iteration in (3.3) for the NL-GS and L-GS.

Theorem 3.2. *If $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$ is SPD, then the fixed point iteration* (3.3) *for the L-GS or NL-GS converges locally to $\boldsymbol{\theta}^*$ in the norm induced by $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$.*

Proof. Symmetry and positive definiteness of the Hessian $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$ implies that

$$B(\boldsymbol{\theta}^*)^T + B(\boldsymbol{\theta}^*) - \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*) = \begin{pmatrix} \mathcal{H}_{11}(\boldsymbol{\theta}^*) & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_{22}(\boldsymbol{\theta}^*) \end{pmatrix}$$

is SPD. Hence, by Lemma 3.2, we have

$$\| I_{2n+1} - B(\boldsymbol{\theta}^*)^{-1} \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*) \|_{\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)} = \sigma < 1,$$

which, together with Theorem 3.1 and Lemma 3.1, implies the validity of the theorem. This completes the proof. □

Remark 3.1. For the Jacobi method described in (2.4), if $\nabla^2 F(\boldsymbol{\theta}^*)$ is SPD, then a similar argument as that of Lemma 3.1 gives the following Jacobian matrix $\mathbf{J}_G(\boldsymbol{\theta}^*)$ of $G(\boldsymbol{\theta})$ at the fixed point $\boldsymbol{\theta}^*$

$$\mathbf{J}_G(\boldsymbol{\theta}^*) = I_{2n+1} - \tilde{B}(\boldsymbol{\theta}^*)^{-1} \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*), \quad \text{where} \quad \tilde{B}(\boldsymbol{\theta}) = \begin{pmatrix} \mathcal{H}_{11}(\boldsymbol{\theta}) & \mathbf{0} \\ \mathbf{0} & \mathcal{H}_{22}(\boldsymbol{\theta}) \end{pmatrix}.$$

Moreover, the iteration is locally convergent to $\boldsymbol{\theta}^*$ if and only if

$$\tilde{B}(\boldsymbol{\theta}^*)^T + \tilde{B}(\boldsymbol{\theta}^*) - \nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*) = \begin{pmatrix} \mathcal{H}_{11}(\boldsymbol{\theta}^*) & -\mathcal{H}_{12}(\boldsymbol{\theta}^*) \\ -\mathcal{H}_{21}(\boldsymbol{\theta}^*) & \mathcal{H}_{22}(\boldsymbol{\theta}^*) \end{pmatrix}$$

is SPD.

4 Applications

This section applies the local convergence theory developed in the previous section to one-dimensional least-squares (LS) function approximation and diffusion-reaction (DR) problems studied in [6].

For the diffusion-reaction problem, let u be the exact solution of the differential equation

$$\begin{cases} -(a(x)u'(x))' + r(x)u(x) = f(x) \quad \text{in} \quad I = (0,1), \\ u(0) = \alpha, \quad u(1) = \beta, \end{cases} \tag{4.1}$$

where the diffusion coefficient $a(x)$, the reaction coefficient $r(x)$, and the right-hand side $f(x)$ are given real-valued functions defined on I. Assume that $a(x)$ and $r(x)$ are

bounded below by the respective positive constant $\mu>0$ and non-negative constant $r_0\geq 0$ almost everywhere on I. Assume that $f(x)\in C(I)$ and that $a(x)\in C^1(I)$. For the least-squares approximation problem, we assume that the weight function $r(x)\geq r_0>0$. For these problems, the NN approximation is to seek $u_n^*(x)=u_n(x;\boldsymbol{\theta}^*)\in\mathcal{M}_n(I)$ such that

$$u_n^*(x)=u_n(x;\boldsymbol{\theta}^*)=\underset{u_n\in\mathcal{M}_n(I)}{\operatorname{argmin}} J(u_n(\cdot;\boldsymbol{\theta})), \tag{4.2}$$

where the functional $J(v)$ is given by

$$J(v)=\begin{cases}\dfrac{1}{2}\displaystyle\int_0^1\Big[a(x)(v'(x))^2+r(x)(v(x))^2\Big]dx-\int_0^1 f(x)v(x)dx+\frac{\gamma}{2}(v(1)-\beta)^2, & \text{DR},\\ \dfrac{1}{2}\displaystyle\int_0^1 r(x)[(v(x)-u(x)]^2dx, & \text{LS}.\end{cases} \tag{4.3}$$

4.1 Hessian

First, we calculate the Hessian matrices. To this end, denote by

$$H(t)=\begin{cases}1, & t>0,\\ \dfrac{1}{2}, & t=0,\\ 0, & t<0,\end{cases}\quad\text{and}\quad \delta(t)=\begin{cases}+\infty, & t=0,\\ 0, & t\neq 0,\end{cases}$$

the Heaviside (unit) step function and the Dirac delta function, respectively. Clearly, $H(t)=\sigma'(t)$ and $\delta(t)=\sigma''(t)$ everywhere except at $t=0$. For each $i=0,1,\cdots,n$, let

$$\sigma_i(x)=\sigma(x-b_i),\quad H_i(x)=H(x-b_i),\quad \delta_i(x)=\delta(x-b_i),$$

and let

$$\boldsymbol{\Sigma}_{n+1}(x)=\begin{pmatrix}\sigma_0(x)\\ \sigma_1(x)\\ \vdots\\ \sigma_n(x)\end{pmatrix},\qquad \mathbf{H}_n(x)=\begin{pmatrix}H_1(x)\\ \vdots\\ H_n(x)\end{pmatrix},$$

$$\mathbf{H}_{n+1}(x)=\begin{pmatrix}H_0(x)\\ H_1(x)\\ \vdots\\ H_n(x)\end{pmatrix},\qquad \boldsymbol{\Lambda}_n(x)=\begin{pmatrix}\delta_1(x)\\ \vdots\\ \delta_n(x)\end{pmatrix}.$$

For $i=1,\cdots,n$, let

$$g_i=g_i(\boldsymbol{\theta})=\begin{cases}r(b_i)u_n(b_i)-f(b_i)-a'(b_i)u_n'(b_i), & \text{DR},\\ r(b_i)(u_n(b_i)-u(b_i)), & \text{LS},\end{cases} \tag{4.4}$$

where $u_n'(b_i) := \sum_{j=0}^{i-1} c_j + \frac{c_i}{2}$, and set

$$\mathbf{g}(\boldsymbol{\theta}) = (g_1, \cdots, g_n)^T, \quad \mathbf{d} = (1-b_1, \cdots, 1-b_n)^T, \quad \text{and} \quad \hat{\mathbf{c}} = (c_1, \cdots, c_n)^T.$$

For any $u_n(x;\boldsymbol{\theta}) \in \mathcal{M}_n(I)$, the value of the functional $J(u_n(\cdot;\boldsymbol{\theta}))$ at $u_n(x;\boldsymbol{\theta})$ is a function of parameters $\boldsymbol{\theta}$. For simplicity of notation, denote this function by F, i.e., $F(\boldsymbol{\theta}) = J(u_n(\cdot;\boldsymbol{\theta}))$. In [6], we derived the principle blocks of the Hessian matrix $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$ as follows

$$\mathcal{H}_{11}(\boldsymbol{\theta}) = \begin{cases} \displaystyle\int_0^1 a(x)\mathbf{H}_{n+1}\mathbf{H}_{n+1}^T dx + \int_0^1 r(x)\boldsymbol{\Sigma}_{n+1}\boldsymbol{\Sigma}_{n+1}^T dx + \gamma \mathbf{d}\mathbf{d}^T, & \text{DR}, \\ \displaystyle\int_0^1 r(x)\boldsymbol{\Sigma}_{n+1}\boldsymbol{\Sigma}_{n+1}^T dx, & \text{LS}, \end{cases} \tag{4.5}$$

$$\mathcal{H}_{22}(\boldsymbol{\theta}) = \begin{cases} \mathbf{D}(\hat{\mathbf{c}})\mathbf{D}(\mathbf{g}) + \mathbf{D}(\hat{\mathbf{c}}) \left(\displaystyle\int_0^1 r(x)\mathbf{H}_n\mathbf{H}_n^T dx \right) \mathbf{D}(\hat{\mathbf{c}}) + \gamma \hat{\mathbf{c}}\hat{\mathbf{c}}^T, & \text{DR}, \\ \mathbf{D}(\hat{\mathbf{c}})\mathbf{D}(\mathbf{g}) + \mathbf{D}(\hat{\mathbf{c}}) \left(\displaystyle\int_0^1 r(x)\mathbf{H}_n\mathbf{H}_n^T dx \right) \mathbf{D}(\hat{\mathbf{c}}), & \text{LS}, \end{cases} \tag{4.6}$$

where $\mathbf{D}(\mathbf{g}) = \text{diag}(g_1, \cdots, g_n)$ and $\mathbf{D}(\hat{\mathbf{c}}) = \text{diag}(c_1, \cdots, c_n)$ are diagonal matrices.

To compute the off-diagonal block $\mathcal{H}_{12}(\boldsymbol{\theta}) = \left(\frac{\partial^2 F(\boldsymbol{\theta})}{\partial c_i \partial b_j} \right)_{(n+1)\times n}$ of the Hessain matrix $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta})$, let

$$F_j(\boldsymbol{\theta}) = \begin{cases} \displaystyle\int_0^1 \Big(f(x)H_j(x) - a(x)u_n'(x)\delta_j(x) - r(x)u_n(x)H_j(x) \Big) dx - \gamma(u_n(1) - \beta), & \text{DR}, \\ \displaystyle\int_0^1 r(x)(u(x) - u_n(x))H_j(x) dx, & \text{LS}, \end{cases}$$

for $j = 1, \cdots, n$. From (3.5) and (6.3) in [5], we have

$$\frac{\partial}{\partial b_j} F(\boldsymbol{\theta}) = c_j F_j(\boldsymbol{\theta}), \tag{4.7}$$

then together with the fact that $\frac{\partial c_j}{\partial c_i} = \delta_{ij}$ (the Kronecker delta), we obtain

$$\frac{\partial^2 F}{\partial c_i \partial b_j}(\boldsymbol{\theta}) = \delta_{ij} F_j(\boldsymbol{\theta}) + c_j \frac{\partial F_j}{\partial c_i}(\boldsymbol{\theta}). \tag{4.8}$$

Lemma 4.1. *For the functional in* (4.3)*, the off-diagonal block* $\mathcal{H}_{12}(\boldsymbol{\theta})$ *has the form*

$$\mathcal{H}_{12}(\boldsymbol{\theta}) = \begin{cases} \mathcal{R}(\boldsymbol{\theta}) - \left(\displaystyle\int_0^1 a(x)\mathbf{H}_{n+1}\boldsymbol{\Lambda}_n^T dx + \int_0^1 r(x)\boldsymbol{\Sigma}_{n+1}\mathbf{H}_n^T dx \right) \mathbf{D}(\hat{\mathbf{c}}) - \gamma \mathbf{d}\hat{\mathbf{c}}^T, & \text{DR}, \\ \mathcal{R}(\boldsymbol{\theta}) - \left(\displaystyle\int_0^1 r(x)\boldsymbol{\Sigma}_{n+1}\mathbf{H}_n^T dx \right) \mathbf{D}(\hat{\mathbf{c}}), & \text{LS}, \end{cases}$$

where $F_0(\boldsymbol{\theta}) = 0$ *and* $\mathcal{R}(\boldsymbol{\theta}) = \big(\mathcal{R}_{ij}(\boldsymbol{\theta})\big) = \big(\delta_{ij} F_i(\boldsymbol{\theta})\big)_{(n+1)\times n}$.

Proof. The lemma is a direct consequence of (4.8) and the fact that

$$\frac{\partial F_j}{\partial c_i}(\boldsymbol{\theta}) = \begin{cases} -\int_0^1 a(x)H_i(x)\delta_j(x)dx - \int_0^1 r(x)\sigma_i(x)H_j(x)dx - \gamma(1-b_i), & \text{DR,} \\ -\int_0^1 r(x)\sigma_i(x)H_j(x)dx, & \text{LS,} \end{cases}$$

for all $i=0,1,\ldots,n$. □

For a given function $w: I \to \mathbb{R}$, define the matrices

$$\mathbb{H}_\Sigma(w;\boldsymbol{\theta}) = \begin{pmatrix} \int_0^1 w(x)\boldsymbol{\Sigma}_{n+1}\boldsymbol{\Sigma}_{n+1}^T dx & -\int_0^1 w(x)\boldsymbol{\Sigma}_{n+1}\mathbf{H}_n^T dx \\ -\int_0^1 w(x)\mathbf{H}_n\boldsymbol{\Sigma}_{n+1}^T dx & \int_0^1 w(x)\mathbf{H}_n\mathbf{H}_n^T dx \end{pmatrix}$$

and

$$\mathbb{H}_\Lambda(w;\boldsymbol{\theta}) = \begin{pmatrix} \int_0^1 w(x)\mathbf{H}_{n+1}\mathbf{H}_{n+1}^T dx & -\int_0^1 w(x)\mathbf{H}_{n+1}\boldsymbol{\Lambda}_n^T dx \\ -\int_0^1 w(x)\boldsymbol{\Lambda}_n\mathbf{H}_{n+1}^T dx & \mathbf{0} \end{pmatrix}.$$

Let $\mathbb{H}_{\Sigma+\Lambda}(\boldsymbol{\theta}) = \mathbb{H}_\Sigma(r;\boldsymbol{\theta}) + \mathbb{H}_\Lambda(a;\boldsymbol{\theta})$. Combining (4.5), (4.6), and Lemma 4.1, the Hessian matrix at $\boldsymbol{\theta}$ has the form

$$\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}) = \mathbb{R}(\boldsymbol{\theta}) + \begin{cases} \mathcal{D}(\hat{\mathbf{c}})\mathbb{H}_{\Sigma+\Lambda}(\boldsymbol{\theta})\mathcal{D}(\hat{\mathbf{c}}) + \mathcal{D}_0(\boldsymbol{\theta}) + \gamma \begin{pmatrix} \mathbf{d} \\ -\hat{\mathbf{c}} \end{pmatrix} \left(\mathbf{d}^T, -\hat{\mathbf{c}}^T\right), & \text{DR,} \\ \mathcal{D}(\hat{\mathbf{c}})\mathbb{H}_\Sigma(r;\boldsymbol{\theta})\mathcal{D}(\hat{\mathbf{c}}) + \mathcal{D}_0(\boldsymbol{\theta}), & \text{LS,} \end{cases} \tag{4.9}$$

where

$$\mathcal{D}(\hat{\mathbf{c}}) = \begin{pmatrix} I_{n+1} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(\hat{\mathbf{c}}) \end{pmatrix}, \quad \mathcal{D}_0(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{D}(\hat{\mathbf{c}})\mathbf{D}(\mathbf{g}) \end{pmatrix} \quad \text{and} \quad \mathbb{R}(\boldsymbol{\theta}) = \begin{pmatrix} \mathbf{0} & \mathcal{R}(\boldsymbol{\theta}) \\ \mathcal{R}^T(\boldsymbol{\theta}) & \mathbf{0} \end{pmatrix}.$$

4.2 Local convergence

In this section, we present sufficient conditions for the local convergence of the BN method for both the diffusion-reaction and the least-squares approximation problems.

First, we establish lower bounds of the smallest eigenvalues of the matrices $\mathbb{H}_\Sigma(w;\boldsymbol{\theta})$ and $\mathbb{H}_\Lambda(w;\boldsymbol{\theta})$.To this end, let

$$h_i = b_{i+1} - b_i, \quad i=0,\cdots,n; \quad h_{\min} = \min_{0\le i\le n}\{h_i\}, \quad \tilde{h}_i = \min\{h_{i-1}, h_i\}, \quad i=1,\cdots,n.$$

Lemma 4.2. *Suppose that* $0 \le w_0 \le w(x) \in C(I)$. *Then, for any* $\boldsymbol{\alpha} = (\alpha_0,\alpha_1,\cdots,\alpha_n)^T \in \mathbb{R}^{n+1}$, $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_n)^T \in \mathbb{R}^n$, *and* $\tau \in (0,1]$, *we have*

$$\begin{cases} (\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\mathbb{H}_\Sigma(w;\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} \ge \dfrac{w_0 h_{\min}^3}{96}|\boldsymbol{\alpha}|^2 + \dfrac{w_0}{24}\displaystyle\sum_{i=1}^n \tilde{h}_i\beta_i^2 \\ (\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\mathbb{H}_\Lambda(w;\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} \ge \dfrac{w_0(1-\tau)h_{\min}}{4}|\boldsymbol{\alpha}|^2 - \dfrac{1}{2\tau w_0}\displaystyle\sum_{i=1}^n w(b_i)^2\left(h_{i-1}^{-1}+h_i^{-1}\right)\beta_i^2, \end{cases} \tag{4.10}$$

where $|\boldsymbol{\alpha}|$ *is the magnitude of* $\boldsymbol{\alpha}$.

Proof. For any $\boldsymbol{\alpha} = (\alpha_0,\alpha_1,\cdots,\alpha_n)^T \in \mathbb{R}^{n+1}$ and any $\boldsymbol{\beta} = (\beta_1,\cdots,\beta_n)^T \in \mathbb{R}^n$, let $\beta_0 = 0$ and define the following piecewise linear and constant functions

$$\alpha_i(x) = \sum_{j=0}^i \alpha_j\sigma_j(x) = \sum_{j=0}^i \alpha_j\sigma(x-b_j) \quad \text{and} \quad \beta_i(x) = \sum_{j=0}^i \beta_j H_j(x) \quad \text{for} \quad i=0,1,\ldots,n,$$

respectively. For $x \in (b_i,b_{i+1})$, $\alpha_n(x) = \alpha_i(x)$ and $\beta_n(x) = \beta_i(x) = \sum_{j=0}^i \beta_j := a_i$. Clearly, we have

$$\left(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T\right)\mathbb{H}_\Sigma(w;\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix} = \int_0^1 w(\alpha_n(x)-\beta_n(x))^2dx \ge w_0\sum_{i=0}^n\int_{b_i}^{b_{i+1}}(\alpha_i(x)-a_i)^2dx. \tag{4.11}$$

By the Simpson rule,

$$\alpha_i(b_i) = \alpha_{i-1}(b_i) \quad \text{and} \quad \sum_{i=0}^n(\alpha_i(b_{i+1})-a_i)^2 \ge \sum_{i=1}^n(\alpha_{i-1}(b_i)-a_{i-1})^2,$$

we obtain

$$\begin{aligned} &12\sum_{i=0}^n\int_{b_i}^{b_{i+1}}(\alpha_i(x)-a_i)^2dx \ge 2\sum_{i=0}^n h_i\left[(\alpha_i(b_i)-a_i)^2+(\alpha_i(b_{i+1})-a_i)^2\right] \\ &\ge \sum_{i=0}^n h_i\left[(\alpha_i(b_i)-a_i)^2+(\alpha_i(b_{i+1})-a_i)^2\right] + \sum_{i=1}^n h_i(\alpha_{i-1}(b_i)-a_i)^2 + \sum_{i=1}^n h_{i-1}(\alpha_{i-1}(b_i)-a_{i-1})^2 \\ &\ge \frac{1}{2}\sum_{i=0}^n h_i(\alpha_i(b_i)-\alpha_i(b_{i+1}))^2 + \frac{1}{2}\sum_{i=0}^n\tilde{h}_i(a_i-a_{i-1})^2 = \frac{1}{2}\sum_{i=0}^n h_i^3\left(\sum_{j=0}^i\alpha_j\right)^2 + \frac{1}{2}\sum_{i=0}^n\tilde{h}_i\beta_i^2. \end{aligned}$$

Now, the first inequality in (4.10) is a direct consequence of (4.11) and the following inequality

$$|\boldsymbol{\alpha}|^2 = \sum_{i=0}^n\left(\sum_{j=0}^i\alpha_j - \sum_{j=0}^{i-1}\alpha_j\right)^2 \le 2\sum_{i=0}^n\left(\sum_{j=0}^i\alpha_j\right)^2 + 2\sum_{i=0}^n\left(\sum_{j=0}^{i-1}\alpha_j\right)^2 \le 4\sum_{i=0}^n\left(\sum_{j=0}^i\alpha_j\right)^2. \tag{4.12}$$

To prove the validity of the second inequality in (4.10), let $\delta^{\epsilon+}(t)$ and $\delta^{\epsilon-}(t)$ be approximations to the Dirac delta function such that

$$\delta^{\epsilon+}(t)=\begin{cases}\dfrac{1}{\epsilon+\rho}, & t\in[-\rho/2,\epsilon+\rho/2],\\ 0, & \text{otherwise,}\end{cases}\quad\text{and}\quad \delta^{\epsilon-}(t)=\begin{cases}\dfrac{1}{\epsilon+\rho}, & t\in[-\rho/2-\epsilon,\rho/2],\\ 0, & \text{otherwise,}\end{cases}$$

for any $\epsilon>0$ and $\rho\in(0,h_{\min})$.

For each $i=1,\cdots,n$, let $\epsilon_i=\tau h_i/2$, $\delta_i^+(x)=\delta^{\epsilon_i+}(x-b_i)$, and $\delta_i^-(x)=\delta^{\epsilon_{i-1}-}(x-b_i)$. Set

$$\hat{\alpha}_n(x)=\sum_{i=0}^{n}\alpha_i H_i(x),\qquad \hat{\beta}_n(x)=\sum_{i=1}^{n}\beta_i\delta_i(x),$$
$$\beta_+(x)=\sum_{i=1}^{n}w(b_i)\beta_i\delta_i^+(x),\qquad \beta_-(x)=\sum_{i=1}^{n}w(b_i)\beta_i\delta_i^-(x).$$

For all $i=0,\cdots,n$ and $j=1,\cdots,n$, it is easy to check that

$$\int_0^1 w(x)H_i(x)\delta_j(x)dx=\frac{1}{2}\left(\int_0^1 w(b_j)H_i(x)\delta_j^+(x)dx+\int_0^1 w(b_j)H_i(x)\delta_j^-(x)dx\right),$$

which, together with multiplying by $\alpha_i\beta_j$ and summing over i and j, implies

$$\int_0^1 w(x)\hat{\alpha}_n(x)\hat{\beta}_n(x)dx=\frac{1}{2}\left(\int_0^1\hat{\alpha}_n(x)\beta_+(x)dx+\int_0^1\hat{\alpha}_n(x)\beta_-(x)dx\right).$$

Then the quadratic form of $\mathbb{H}_\Lambda(w;\boldsymbol{\theta})$ is bounded below by

$$\begin{aligned}&\left(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T\right)\mathbb{H}_\Lambda(w;\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\\ =&\int_0^1 w(x)\hat{\alpha}_n^2(x)dx-2\int_0^1 w(x)\hat{\alpha}_n(x)\hat{\beta}_n(x)dx\\ \geq& w_0\int_0^1\hat{\alpha}_n^2(x)dx-\int_0^1\hat{\alpha}_n(x)\beta_+(x)dx-\int_0^1\hat{\alpha}_n(x)\beta_-(x)dx\\ =&\frac{1}{2}\int_0^1\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_+(x)\right)^2dx+\frac{1}{2}\int_0^1\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_-(x)\right)^2dx\\ &-\frac{1}{2w_0}\left(\int_0^1\beta_+^2(x)+\beta_-^2(x)dx\right).\end{aligned}\tag{4.13}$$

To bound the first term below, let $\beta_{n+1}=c_{n+1}=\beta_0=0$ and $\epsilon_0=\frac{h_0}{2}$. It follows from the facts

that $\hat{\alpha}_n(x)$ and $\beta_+(x)$ are piecewise constants, $h_i-\epsilon_i=\epsilon_i+(1-\tau)h_i$ and (4.12) that

$$\begin{aligned}
&\int_0^1\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_+(x)\right)^2dx\\
\geq&\sum_{i=0}^{n}\left(\int_{b_i}^{b_i+\epsilon_i+\rho/2}+\int_{b_i+\epsilon_i+\rho/2}^{b_{i+1}-\rho/2}\right)\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_+(x)\right)^2dx\\
=&\sum_{i=0}^{n}\left(\epsilon_i+\frac{\rho}{2}\right)\left(\sqrt{w_0}\sum_{j=0}^{i}\alpha_j-\frac{w(b_i)\beta_i}{(\epsilon_i+\rho)\sqrt{w_0}}\right)^2+\sum_{i=0}^{n}(h_i-\epsilon_i-\rho)\left(\sqrt{w_0}\sum_{j=0}^{i}\alpha_j\right)^2\\
\geq&\sum_{i=0}^{n}\frac{\epsilon_i}{2}\left(\frac{w(b_i)\beta_i}{(\epsilon_i+\rho)\sqrt{w_0}}\right)^2+\sum_{i=0}^{n}[(1-\tau)h_i-\rho]\left(\sqrt{w_0}\sum_{j=0}^{i}\alpha_j\right)^2\\
\geq&\frac{1}{2w_0}\sum_{i=1}^{n}\frac{\epsilon_i w^2(b_i)\beta_i^2}{(\epsilon_i+\rho)^2}+w_0\frac{(1-\tau)h_{\min}-\rho}{4}|\boldsymbol{\alpha}|^2.
\end{aligned}\tag{4.14}$$

Similarly, for the second term in (4.13), following an argument analogous to the one used for inequality (4.14), we obtain

$$\begin{aligned}
&\int_0^1\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_-(x)\right)^2dx\\
\geq&\sum_{i=1}^{n+1}\left(\int_{b_{i-1}+\rho/2}^{b_{i-1}-\epsilon_{i-1}-\rho/2}+\int_{b_i-\epsilon_{i-1}-\rho/2}^{b_i}\right)\left(\sqrt{w_0}\hat{\alpha}_n(x)-\frac{1}{\sqrt{w_0}}\beta_-(x)\right)^2dx\\
\geq&\frac{1}{2w_0}\sum_{i=1}^{n}\frac{\epsilon_{i-1}w^2(b_i)\beta_i^2}{(\epsilon_{i-1}+\rho)^2}+w_0\frac{(1-\tau)h_{\min}-\rho}{4}|\boldsymbol{\alpha}|^2.
\end{aligned}\tag{4.15}$$

Since

$$\int_0^1\beta_+^2(x)dx=\sum_{i=1}^{n}(\epsilon_i+\rho)^{-1}w^2(b_i)\beta_i^2\quad\text{and}\quad\int_0^1\beta_-^2(x)dx=\sum_{i=1}^{n}(\epsilon_{i-1}+\rho)^{-1}w^2(b_i)\beta_i^2,$$

after letting $\rho\to0$, it follows from (4.13), (4.14) and (4.15) that

$$\begin{pmatrix}\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T\end{pmatrix}\mathbb{H}_\Lambda(w;\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\\boldsymbol{\beta}\end{pmatrix}\geq\frac{w_0(1-\tau)h_{\min}}{4}|\boldsymbol{\alpha}|^2-\frac{1}{2\tau w_0}\sum_{i=1}^{n}w(b_i)^2\left(h_{i-1}^{-1}+h_i^{-1}\right)\beta_i^2,$$

and the lemma is proved. □

Lemma 4.3. *For all $i\in\{1,\cdots,n\}$, assume that $c_i\neq0$ and that*

$$F_i^2(\boldsymbol{\theta})<\begin{cases}c_i^2\left(\dfrac{r_0h_{\min}^3}{96}+\dfrac{\mu(1-\tau)h_{\min}}{4}\right)\left(\dfrac{g_i}{c_i}+\dfrac{r_0\tilde{h}_i}{24}-\dfrac{a^2(b_i)}{2\tau\mu}\left(h_{i-1}^{-1}+h_i^{-1}\right)\right), & \text{DR},\\ c_i^2\left(\dfrac{r_0h_{\min}^3}{96}\right)\left(\dfrac{g_i}{c_i}+\dfrac{r_0\tilde{h}_i}{24}\right), & \text{LS},\end{cases}\tag{4.16}$$

for some $0<\tau<1$, then $\nabla_{\boldsymbol{\theta}}^2F(\boldsymbol{\theta})=\nabla_{\boldsymbol{\theta}}^2J(u_n(\cdot;\boldsymbol{\theta}))$ is SPD.

Proof. For any $\boldsymbol{\alpha}=(\alpha_0,\alpha_1,\cdots,\alpha_n)^T\in\mathbb{R}^{n+1}$ and $\boldsymbol{\beta}=(\beta_1,\cdots,\beta_n)^T\in\mathbb{R}^n$, let $\hat{\boldsymbol{\beta}}=\mathbf{D}(\hat{\mathbf{c}})\boldsymbol{\beta}$. For the diffusion-reaction problem, it follows from (4.9), the fact that $\left(\boldsymbol{\alpha}^T\mathbf{d}-\boldsymbol{\beta}^T\hat{\mathbf{c}}\right)^2\geq 0$, (4.10), and the assumptions on the lower bounds of $a(x)$ and $r(x)$ that

$$
\begin{aligned}
&(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\\
=&(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\mathbb{R}(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}+(\boldsymbol{\alpha}^T,\hat{\boldsymbol{\beta}}^T)\mathbb{H}_{\Sigma+\Lambda}(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}+\hat{\boldsymbol{\beta}}^T\mathbf{D}(\mathbf{g})\mathbf{D}^{-1}(\hat{\mathbf{c}})\hat{\boldsymbol{\beta}}+\gamma\left(\boldsymbol{\alpha}^T\mathbf{d}-\boldsymbol{\beta}^T\hat{\mathbf{c}}\right)^2\\
\geq&(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\mathbb{R}(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}+(\boldsymbol{\alpha}^T,\hat{\boldsymbol{\beta}}^T)\mathbb{H}_{\Sigma+\Lambda}(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}+\hat{\boldsymbol{\beta}}^T\mathbf{D}(\mathbf{g})\mathbf{D}^{-1}(\hat{\mathbf{c}})\hat{\boldsymbol{\beta}}\\
\geq&2\sum_{i=1}^n\alpha_i\beta_iF_i+\frac{r_0(h_{\min})^3}{96}|\boldsymbol{\alpha}|^2+\frac{r_0}{24}\sum_{i=1}^n\tilde{h}_i(c_i\beta_i)^2+\frac{\mu(1-\tau)h_{\min}}{4}|\boldsymbol{\alpha}|^2\\
&+\sum_{i=1}^n\left(\frac{g_i}{c_i}-\frac{a^2(b_i)}{2\mu}(h_{i-1}^{-1}+h_i^{-1})\right)(c_i\beta_i)^2\\
=&\sum_{i=1}^n\left[\left(\frac{r_0h_{\min}^3}{96}+\frac{\mu(1-\tau)h_{\min}}{4}\right)\alpha_i^2+2\alpha_i\beta_iF_i+\left(\frac{g_i}{c_i}+\frac{r_0\tilde{h}_i}{24}-\frac{a^2(b_i)}{2\tau\mu}(h_{i-1}^{-1}+h_i^{-1})\right)c_i^2\beta_i^2\right].
\end{aligned}
$$

In a similar fashion, we have

$$
(\boldsymbol{\alpha}^T,\boldsymbol{\beta}^T)\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta})\begin{pmatrix}\boldsymbol{\alpha}\\ \boldsymbol{\beta}\end{pmatrix}\geq\sum_{i=1}^n\left[\left(\frac{r_0h_{\min}^3}{96}\right)\alpha_i^2+2\alpha_i\beta_iF_i+\left(\frac{g_i}{c_i}+\frac{r_0\tilde{h}_i}{24}\right)c_i^2\beta_i^2\right]
$$

for the least-squares approximation problem. Notice that $a_1x^2+2a_2xy+a_3y^2>0$ for all $(x,y)\neq(0,0)$ if $a_1>0$, $a_3>0$, and $a_2^2<a_1a_3$. Now, the positive definiteness of $\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta})$ is a direct consequence of (4.16). □

If $c_j\neq 0$, by (4.7), $F_j(\boldsymbol{\theta})$ vanishes at a critical point $\boldsymbol{\theta}^*$. Hence, $\mathcal{R}(\boldsymbol{\theta}^*)=\mathbf{0}$, which gives

$$
\nabla^2_{\boldsymbol{\theta}}F(\boldsymbol{\theta}^*)=\begin{cases}\mathcal{D}(\hat{\mathbf{c}}^*)\mathbb{H}_{\Sigma+\Lambda}(\boldsymbol{\theta}^*)\mathcal{D}(\hat{\mathbf{c}}^*)+\mathcal{D}_0(\boldsymbol{\theta}^*)+\gamma\begin{pmatrix}\mathbf{d}\\ -\hat{\mathbf{c}}^*\end{pmatrix}(\mathbf{d}^T,-(\hat{\mathbf{c}}^*)^T), & \text{DR},\\ \mathcal{D}(\hat{\mathbf{c}}^*)\mathbb{H}_{\Sigma}(r;\boldsymbol{\theta}^*)\mathcal{D}(\hat{\mathbf{c}}^*)+\mathcal{D}_0(\boldsymbol{\theta}^*), & \text{LS}.\end{cases}
$$

Theorem 4.1. *For an open set $\mathcal{O}\subseteq\mathbb{R}^{2n+1}$ satisfying the invertibility assumption, let $\boldsymbol{\theta}^*=\begin{pmatrix}\mathbf{c}^*\\ \mathbf{b}^*\end{pmatrix}\in\mathcal{O}$ be a minimizer of problem* (4.2)*. For all $i\in\{1,\ldots,n\}$, assume that $c_i^*\neq 0$ and that*

$$
\frac{g_i^*}{c_i^*}+\frac{r_0\tilde{h}_i^*}{24}>\begin{cases}\dfrac{a^2(b_i^*)}{2\mu}\left(\dfrac{1}{h_{i-1}^*}+\dfrac{1}{h_i^*}\right), & \text{DR},\\ 0, & \text{LS}.\end{cases}\tag{4.17}
$$

Then the Hessian matrix at the critical point $\boldsymbol{\theta}^*$, $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*) = \nabla^2_{\boldsymbol{\theta}} J(u_n(\cdot;\boldsymbol{\theta}^*))$, *is SPD. Moreover, the BN method using either NL-GS or L-GS converges locally to* $\boldsymbol{\theta}^*$ *in the norm induced by* $\nabla^2_{\boldsymbol{\theta}} F(\boldsymbol{\theta}^*)$.

Proof. This is a direct consequence of Lemma 4.3, (4.17), and the fact that $F_i(\boldsymbol{\theta}^*) = 0$ for all $i = 1, \cdots, n$. □

4.3 Feasibility of BN

This section discusses feasibility of the BN with either NL-GS or L-GS, or equivalently, invertibility of $\mathcal{H}_{11}(\boldsymbol{\theta})$ and $\mathcal{H}_{22}(\boldsymbol{\theta})$. Invertibility of $\mathcal{H}_{11}(\boldsymbol{\theta})$ was shown in [6]; moreover, $\mathcal{H}_{11}(\boldsymbol{\theta})$ is SPD.

Regarding $\mathcal{H}_{22}(\boldsymbol{\theta})$, the formula in (4.6) uses derivative of the diffusion coefficient $a(x)$. As discussed in Section 5 of [5], if $a(x)$ is not differential at b_i for some $i \in \{1,\ldots,n\}$, then b_i lies at the physical interface. This means that the (mesh) breakpoint b_i is already at a good location and hence should be fixed without further update.

Below, we discuss invertibility of $\mathcal{H}_{22}(\boldsymbol{\theta})$ at the k^{th} iteration. For simplicity of notation, the parameters are still denoted by $\hat{\mathbf{c}}$ and $\mathbf{b}$ without the superscript (k). The assumption $c_i \neq 0$ for all $i \in \{1,\ldots,n\}$ is a necessary condition for invertibility of $\mathcal{H}_{22}(\boldsymbol{\theta})$. During the iterative process, this condition is not always satisfied. In the case that $c_l = 0$ for some $l \in \{1,\ldots,n\}$, then $\mathcal{H}_{22}(\boldsymbol{\theta})$ is singular. To deal with such singularity, similar to [4–6], the nonlinear parameter b_l is not updated at the current step, because the corresponding neuron has no contribution to the current approximation. When this happens several times to a neuron, then we can either remove or redistribute this neuron. Based on the above discussion, we introduce the following set of indices for the non-contributing neurons

$$S_1 = \big\{ i \in \{1,\cdots,n\} : |c_i| < \tau_1 \text{ or } b_i \notin I \big\}, \tag{4.18}$$

where $0 \leq \tau_1 < 1$ is a small parameter.

Next, at each step, to update the nonlinear parameters of neurons whose indices are not in S_1, let us denote by $\hat{\mathcal{H}}_{22}(\boldsymbol{\theta})$ the reduced Hessian after removing neurons with indices in S_1. For the diffusion problem, $\hat{\mathcal{H}}_{22}(\boldsymbol{\theta})$ is invertible if and only if $g_i \neq 0$ for all $i \notin S_1$. Under the assumption that $u_n(x) \approx u(x)$, i.e., $u_n(x)$ is a good approximation to $u(x)$, it was shown in [5] that

$$g_i \approx a(b_i) u''(b_i), \tag{4.19}$$

where g_i is defined in (4.4). As discussed in [5], $g_i \approx 0$ implies that b_i is either at a nearly optimal location or no need for update. In both cases, b_i is not updated. Hence, we introduce the following set of indices for non-updating neurons

$$S_2 = \left\{ i \in \{1,\cdots,n\} \backslash S_1 : \frac{|g_i|}{a(b_i)} \leq \tau_2 \ \text{ or } \ a'(b_i) \ \text{ DNE} \right\}, \tag{4.20}$$

where $0 \leq \tau_2 < 1$ is a small parameter.

The above strategy does apply to the diffusion-reaction problem because (4.4) and (4.1) imply the validity of (4.19):

$$g_i = r(b_i)u_n(b_i) - f(b_i) - a'(b_i)u_n'(b_i) \approx r(b_i)u(b_i) - f(b_i) - a'(b_i)u'(b_i) = a(b_i)u''(b_i).$$

Nevertheless, the sufficient condition on the positive definiteness of $\mathcal{H}_{22}(\boldsymbol{\theta})$ was given in Lemma 5.2 of [6]:

$$\frac{g_i}{c_i} + \frac{r_0 \tilde{h}_i}{4} > 0 \quad \text{for all } i \in \{1,\cdots,n\} \setminus S_1, \tag{4.21}$$

where $r_0 \geq 0$ for DR and $r_0 > 0$ for LS. Notice that the linear parameter $c_i = u_n'(b_i^+) - u_n'(b_i^-)$ means the change of the slope of $u_n(x)$ at b_i (see Section 5 of [5]). Hence, $c_i \approx u'(b_i^+) - u'(b_i^-)$ and $u''(b_i)$ have the same sign, which implies

$$\frac{g_i}{c_i} \approx \frac{a(b_i)u''(b_i)}{c_i} = \frac{a(b_i)|u''(b_i)|}{|c_i|},$$

which, together with positivity of $a(b_i)$, implies (4.21) is almost always valid.

For the LS problem, (4.21) is sufficient but not necessary for the positive definiteness of $\mathcal{H}_{22}(\boldsymbol{\theta})$. Moreover, positive definiteness differs from invertibility. Since $g_i = r(b_i)(u_n(b_i) - u(b_i))$ is assumed to be small, for implementation purposes, we do not update the i^{th} neuron if $g_i/c_i < -1/\tau_3$ is negatively large, where parameter $\tau_3 \in (0,1)$. In this case, $|c_i| < \tau_3 |g_i|$ is small if g_i is small. Hence, let

$$S_2 = \left\{ i \in \{1,\cdots,n\} \setminus S_1 : \frac{g_i}{c_i} < 0 \text{ and } |c_i| < \tau_3 |g_i| \right\}. \tag{4.22}$$

Notice that the reduced Hessian $\hat{\mathcal{H}}_{22}(\boldsymbol{\theta})$ is not guaranteed to be nonsingular after removing neurons with indices in $S_1 \cup S_2$. In cases where the reduced Hessian $\hat{\mathcal{H}}_{22}(\boldsymbol{\theta})$ is not invertible, one might consider the Gauss-Newton matrix

$$\mathcal{G}_{22}(\boldsymbol{\theta}) = \mathcal{D}(\hat{\mathbf{c}}) \mathbb{H}_{\Sigma}(r;\boldsymbol{\theta}) \mathcal{D}(\hat{\mathbf{c}}),$$

which is SPD under the assumption that $c_i \neq 0$ for all $i \in \{1,\cdots,n\}$.

4.4 Implementation

In this section, we outline the implementation of the BN method using a reduced nonlinear system constructed based on the sets (4.18), (4.20), and (4.22). The NL-GS scheme, implemented in earlier works [5,6], is used to illustrate the implementation, which is also applicable to the other two schemes described in Section 2.

To this end, let

$$S = \{1,\cdots,n\} \setminus (S_1 \cup S_2)$$

be the set of indices for the neurons which remain in the system, and denote by $S^c = S_1 \cup S_2$ the complement of S, the set of indices to be removed. Next, for a given vector $\mathbf{v} \in \mathbb{R}^n$,

denote by $\mathbf{v}_S \in \mathbb{R}^{n-|S^c|}$ as the vector obtained by removing entries whose indices belong to S^c, where $|S^c|$ denotes the number of indices in S^c. Similarly, for a matrix $\mathbf{B} \in \mathbb{R}^{n\times n}$, define $\mathbf{B}_S \in \mathbb{R}^{(n-|S^c|)\times(n-|S^c|)}$ as the matrix obtained by removing the rows of columns of $\mathbf{B}$ corresponding to the indices in S^c. Then the *reduced* search direction vector for the Newton step is defined as

$$\mathbf{d}_R(\mathbf{c},\mathbf{b}) = \mathbf{d}_R(\boldsymbol{\theta}) = -\Big(\mathcal{H}_{22}(\boldsymbol{\theta})_S\Big)^{-1}\Big(\nabla_{\mathbf{b}}F(\boldsymbol{\theta})\Big)_S. \tag{4.23}$$

We are now ready to present the reduced block Newton (rBN) algorithm with the NL-GS scheme (see Algorithm 4.1 for pseudocode). Given the previous iterate $\mathbf{b}^{(k)}$. Then the current iterate $\boldsymbol{\theta}^{(k+1)} = \begin{pmatrix}\mathbf{c}^{(k+1)}\\ \mathbf{b}^{(k+1)}\end{pmatrix}$ is computed as follows:

(i) Compute the linear parameters

$$\mathbf{c}^{(k+1)} = \mathbf{c}^{(k)} - \mathcal{H}_{11}(\mathbf{c}^{(k)},\mathbf{b}^{(k)})^{-1}\nabla_{\mathbf{c}}F(\mathbf{c}^{(k)},\mathbf{b}^{(k)}).$$

(ii) Compute the search direction $\mathbf{p}^{(k)} = \Big(p_1^{(k)},\cdots,p_n^{(k)}\Big)^T$ by

$$\big(p_i^{(k)}\big)_{i\in S} = \mathbf{d}_R(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)}) \quad \text{and} \quad \big(p_i^{(k)}\big)_{i\notin S} = \mathbf{0}, \tag{4.24}$$

where $\mathbf{d}_R(\mathbf{c}^{(k+1)},\mathbf{b}^{(k)})$ is the *reduced* Newton's direction vector defined in (4.23).

(iii) Compute the nonlinear parameters

$$\mathbf{b}^{(k+1)} = \mathbf{b}^{(k)} + \mathbf{p}^{(k)}.$$

(iv) Redistribute non-contributing breakpoints $b_i^{(k+1)}$ for all $i \in S_1$ and sort $\mathbf{b}^{(k+1)}$.

Remark 4.1. The redistribution implemented for the numerical results shown in [5,6] in step (iv) was carried out as follows: For a neuron $b_l^{(k+1)}$ satisfying $l \in S_1$, we set

$$b_l^{(k+1)} \leftarrow \frac{b_{m-1}^{(k+1)} + b_m^{(k+1)}}{2},$$

where $m \in \{1,\cdots,n+1\}$ is an integer chosen uniformly at random.

Algorithm 4.1 Reduced block Newton (rBN) method for (4.2).

Require: Initial network parameters $\mathbf{b}^{(0)}$
Ensure: Network parameters $\mathbf{c}$, $\mathbf{b}$
for $k=0,1,\ldots,$ **do**
▷ Linear parameters
$\mathbf{c}^{(k+1)} \leftarrow \mathbf{c}^{(k)} - \mathcal{H}_{11}(\mathbf{c}^{(k)},\mathbf{b}^{(k)})^{-1}\nabla_{\mathbf{c}}F(\mathbf{c}^{(k)},\mathbf{b}^{(k)})$
▷ nonlinear parameters
Compute the search direction $\mathbf{p}^{(k)}$ *as in* (4.24)
$\mathbf{b}^{(k+1)} \leftarrow \mathbf{b}^{(k)} + \mathbf{p}^{(k)}$
▷ Redistribute non-contributing neurons and sort $\mathbf{b}^{(k+1)}$
end for

4.4.1 Numerical experiment

To illustrate the performance of the reduced BN method outlined above, consider a singularly perturbed reaction-diffusion equation:

$$\begin{cases} -\varepsilon^2 u''(x) + u(x) = f(x), & x \in (-1,1), \\ u(-1) = u(1) = 0. \end{cases} \tag{4.25}$$

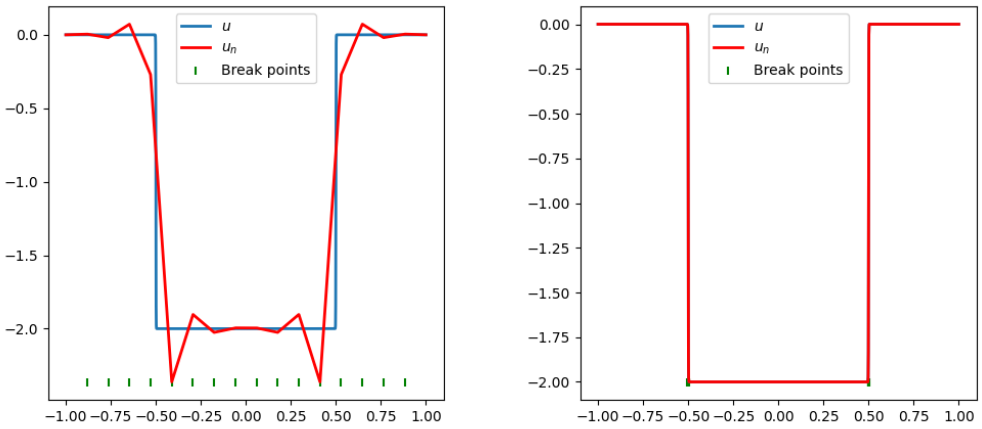
For

$$\begin{aligned} f(x) = -2\left(\varepsilon - 4x^2\tanh\left(\frac{1}{\varepsilon}\left(x^2-\frac{1}{4}\right)\right)\right)\left(1\Big/\cosh\left(\frac{1}{\varepsilon}\left(x^2-\frac{1}{4}\right)\right)\right)^2 \\ + \tanh\left(\frac{1}{\varepsilon}\left(x^2-\frac{1}{4}\right)\right) - \tanh\left(\frac{3}{4\varepsilon}\right), \end{aligned}$$

problem (4.25) has the following exact solution:

$$u(x) = \tanh\left(\frac{1}{\varepsilon}\left(x^2-\frac{1}{4}\right)\right) - \tanh\left(\frac{3}{4\varepsilon}\right). \tag{4.26}$$

Denote by $|\cdot|_1$ the H^1 seminorm on $(-1,1)$. For $\nu=\varepsilon^2=10^{-6}$, the solution of (4.26) exhibits sharp interior layers. It is well-known that overshooting and oscillations occur when using continuous piecewise linear approximation on a uniform partition (see Fig. 1(a) for $n=16$). Using these uniform mesh points as an initial for the nonlinear parameters, 100 iterations of the BN method moves them efficiently toward the interior layers, and the resulting approximation is greatly improved (see Fig. 1(b)). This example demonstrates importance of non-uniform mesh in approximation and efficiency of the BN method in non-convex optimization. For more numerical experiments, we refer readers to [5] and [6].



(a) (b)

Figure 1: For $\nu=\varepsilon^2=10^{-6}$: (a) initial NN model with 16 uniform breakpoints; $\frac{|u-u_n|_1}{|u|_1}=0.988$, (b) optimized NN model with 16 breakpoints, 100 iterations, $\frac{|u-u_n|_1}{|u|_1}=0.173$.

4.5 Reduced nonlinear system

This section studies the rBN method introduced in Section 4.4. As explained in Section 4.4, the iterative process may fix the location of certain breakpoints. We analyze the local convergence of the method under the assumption that no additional breakpoints will be fixed after a finite number of iterations.

To this end, for a given $\alpha\in\mathbb{R}$, assume that there are $k<n$ fixed breakpoints denoted by

$$0=\tilde{b}_0<\tilde{b}_1<\cdots<\tilde{b}_k<1.$$

Define the set of shallow ReLU neural networks with k fixed and $n-k$ moving breakpoints as follows:

$$\mathcal{M}_{n,k}(I)=\left\{\alpha+\sum_{i=0}^{k}c_i\sigma(x-\tilde{b}_i)+\sum_{j=1}^{n-k}c_{k+j}\sigma(x-b_j):c_l\in\mathbb{R},0<b_1<\cdots<b_{n-k}<1\right\}.$$

For any $u_{n,k}(x)\in\mathcal{M}_{n,k}(I)$, denote the corresponding parameters by

$$\boldsymbol{\theta}_R=\begin{pmatrix}\mathbf{c}\\ \mathbf{b}_R\end{pmatrix}\in\mathbb{R}^{2n+1-k},$$

where $\mathbf{c}=(c_0,\cdots,c_n)^T\in\mathbb{R}^{n+1}$ are again the linear parameters and $\mathbf{b}_R=(b_1,\cdots,b_{n-k})^T\in\mathbb{R}^{n-k}$ are the nonlinear parameters. After a certain number of the BN iterations, the optimization becomes to seek $u^*_{n,k}(x)=u_{n,k}(x;\boldsymbol{\theta}^*_R)\in\mathcal{M}_{n,k}(I)$ such that

$$u^*_{n,k}(x)=u^*_{n,k}(x;\boldsymbol{\theta}^*_R)=\underset{u_{n,k}\in\mathcal{M}_{n,k}(I)}{\operatorname{argmin}} J(u_{n,k}(\cdot;\boldsymbol{\theta}_R)).$$

The analysis presented in the preceding sections extends to the rBN method. Furthermore, following a similar reasoning as in Theorem 4.1, if $c^*_{k+i}\neq 0$ and

$$\frac{g^*_i}{c^*_{i+k}}+\frac{r_0\tilde{h}^*_i}{24}>\begin{cases}\dfrac{a^2(b^*_i)}{2\mu}\left(\dfrac{1}{h^*_{i-1}}+\dfrac{1}{h^*_i}\right), & \text{DR},\\ 0, & \text{LS},\end{cases}$$

for all $i\in\{1,\cdots,n-k\}$, then the BN method applied to this reduced parameter set converges locally to $\boldsymbol{\theta}^*_R=\begin{pmatrix}\mathbf{c}^*\\ \mathbf{b}^*_R\end{pmatrix}$.

Remark 4.2. In Section 3, we presented the local convergence analysis assuming the dimensions of the linear and nonlinear parameters are $n+1$ and n, respectively. For the rBN method discussed here, the dimensions are $n+1$ and $n-k$. However, the proofs of Lemma 3.2 and the other lemmas in that section do not depend on the specific dimension of the parameter space. Therefore, the local convergence analysis extends directly to the rBN method.

5 Conclusions

A local convergence analysis for the BN methods including the reduced BN (rBN) method is presented in this paper. The rBN differs from the common optimization methods in reducing the number of parameters during the optimization process. We established a local convergence of the BN methods under some reasonable contiditions for one-dimensional diffusion-reaction problems and least-squares function approximation. The analysis may be extended to higher dimensions for shallow neural network approximation problems.

Acknowledgements

This work was performed under the auspices of the U.S. Department of Energy by Lawrence Livermore National Laboratory under Contract DE-AC52-07NA27344. This work was supported by the NNSA Advanced Simulation and Computing (ASC) Program LLNL-JRNL-2015820.

References

[1] M. Ainsworth and Y. Shin, Plateau phenomenon in gradient descent training of ReLU networks: Explanation, quantification and avoidance, SIAM J. Sci. Comput., 43 (2020), A3438–A3468.

[2] M. Ainsworth and Y. Shin, Active neuron least squares: A training method for multivariate rectified neural networks, SIAM J. Sci. Comput., 44(4) (2022), A2253–A2275.

[3] H. G. Burchard, Splines (with optimal knots) are better, Appl. Anal., 3 (1974), 309–319.

[4] Z. Cai, T. Ding, M. Liu, X. Liu, and J. Xia, A structure-guided gauss-newton method for shallow ReLU neural network, arXiv:2404.05064v1 [cs.LG], 2024.

[5] Z. Cai, A. Doktorova, R. D. Falgout, and C. Herrera, Efficient shallow Ritz method for 1D diffusion problems, Comput. Math. Appl., 200 (2025), 349–363.

[6] Z. Cai, A. Doktorova, R. D. Falgout, and C. Herrera, Efficient shallow Ritz method for 1D diffusion-reaction problems, SIAM J. Sci. Comput., (2025), S414–S435.

[7] E. C. Cyr, M. A. Gulian, R. G. Patel, M. Perego, and N. A. Trask, Robust training and initialization of deep neural networks: An adaptive basis viewpoint, Proceedings of Machine Learning Research, 107 (2020), 512–536.

[8] I. Daubechies, R. DeVore, S. Foucart, B. Hanin, and G. Petrova, Nonlinear approximation and (deep) ReLU networks, Constr. Approx., 55(1) (2022), 127–172.

[9] R. Falgout and P. Vassilevski, On generalizing the algebraic multigrid framework, SIAM J. Numer. Anal., 42 (2004), 1669–1693.

[10] G. H. Golub and V. Pereyra, The differentiation of pseudo-inverses and nonlinear least squares problems whose variables separate, SIAM J. Numer. Anal., 10(2) (1973), 413–432.

[11] M. Liu, Z. Cai, and J. Chen. Adaptive two-layer ReLU neural network: I. best least-squares approximation. Comput. Math. Appl., 113 (2022), 34–44.

[12] X. Liu and Y. Yuan, On the separable nonlinear least squares problems, J. Comput. Math., 26(3) (2008), 390–403.

[13] J. M. Ortega and W. C. Rheinboldt, Iterative Solution of Nonlinear Equations in Several Variables, Classics in Applied Mathematics, Society for Industrial and Applied Mathematics, 1970.

[14] J. Park, J. Xu, and X. Xu, A neuron-wise subspace correction method for the finite neuron method, arXiv:2211.12031 [math.NA], 2022.

[15] A. Ruhe and P. A. Wedin. Algorithms for separable nonlinear least squares problems. SIAM Rev., 22(3) (1980), 318–337.

[16] L. Schumaker, Spline Functions: Basic Theory, Wiley, New York, 1981.