Fast Iterative Solver for Neural Network Method: II. 1D General Elliptic Problems and Data Fitting

Zhiqiang Cai[1] Anastassia Doktorova[1] Robert D. Falgout[2]
Speaker: César Herrera[1]

[1]Department of Mathematics, Purdue University
[2]Lawrence Livermore National Laboratory

Copper Mountain Conference on Iterative Methods
April 2024

Table of Contents

Table of Contents

Shallow Neural Network

Let

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

where $\sigma(t) = \max\{t, 0\}$.
Given $\mathbf{b} = (b_0, b_1, \ldots, b_n)^T$, let

$$\mathbf{H}(x) := (\sigma'(x - b_0), \sigma'(x - b_1), \ldots, \sigma'(x - b_n))^T.$$

Coefficient matrix: $A(\mathbf{b}) = \int_0^1 \mathbf{H}(x)\mathbf{H}(x)^T dx$

Mass matrix

Let

$$\Sigma(x) := (\sigma(x - b_0), \sigma(x - b_1), \ldots, \sigma(x - b_n))^T.$$

Mass matrix: $M(\mathbf{b}) = \int_0^1 \Sigma(x)\Sigma(x)^T dx$

Lemma

The condition number of the mass matrix $M(\mathbf{b})$ is bounded above by $\mathcal{O}\left(n/h_{min}^3\right)$.

Table of Contents

Least-Squares Optimization Problems

Given a function $u(x)$ defined on $I = (0, 1)$, the best least-squares approximation to u in $\mathcal{M}_n(\Omega)$ is to find $u_n \in \mathcal{M}_n(\Omega)$ such that

$$\mathcal{J}(u_n) = \min_{v \in \mathcal{M}_n(\Omega)} \mathcal{J}(v),$$

where $\mathcal{J}$ is the least-squares loss functional given by

$$\mathcal{J}(v) = \frac{1}{2} \int_\Omega ((v(x) - u(x))^2 \, dx.$$

Systems of algebraic equations

Let

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

be the best least-squares NN approximation. Then the linear and nonlinear parameters

$$\mathbf{c} = (c_0, \ldots, c_n)^T \quad \text{and} \quad \mathbf{b} = (b_0, \ldots, b_n)^T$$

satisfy the following system of algebraic equations

$$\nabla_{\mathbf{c}} \mathcal{J}(u_n) = \mathbf{0} \quad \text{and} \quad \nabla_{\mathbf{b}} \mathcal{J}(u_n) = \mathbf{0}.$$

Linear parameters $\mathbf{c}$

The equation $\nabla_{\mathbf{c}}\mathcal{J}(u_n) = 0$ has the form

$$M(\mathbf{b})\mathbf{c} = \mathbf{u}(\mathbf{b}),$$

where $\mathbf{u}(\mathbf{b}) = \left(\int_0^1 u(x)\sigma(x-b_0)dx, \ldots, \int_0^1 u(x)\sigma(x-b_n)dx\right)^T$.

Linear parameters $\mathbf{c}$

Let $h_i = b_i - b_{i-1}$ for $i = 0, \ldots, n$. Define the matrices

$$G = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \quad D_h(\mathbf{b}) = \begin{pmatrix} h_0 & & & \\ & h_1 & & \\ & & \ddots & \\ & & & h_n \end{pmatrix},$$

and the tridiagonal matrices

$$T_1 = T_1(\mathbf{b}) = G D_h(\mathbf{b})^{-1} G,$$

$$T_2 = T_2(\mathbf{b}) = \frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & \\ h_2 & 2(h_2 + h_3) & h_3 & \\ & & \ddots & \\ & & h_{n+1} & 2h_{n+1} \end{pmatrix}.$$

Linear parameters $\mathbf{c}$

The following factorization holds:

$$M(\mathbf{b}) = T_1^{-T} T_2 T_1^{-1}$$

i.e.,

$$M(\mathbf{b})^{-1} = T_1 (T_2)^{-1} T_1.$$

So that the linear system

$$M(\mathbf{b})\mathbf{c} = \mathbf{u}(\mathbf{b}),$$

can be solved in $14(n+1)$ operations.

Nonlinear parameters $\mathbf{b}$

Lemma

The Hessian matrix $\nabla^2_{\mathbf{b}} \mathcal{J}(u_n)$ has the form

$$\mathcal{H}(\mathbf{c}, \mathbf{b}) = D(\mathbf{c})\Lambda(\mathbf{c}, \mathbf{b}) + D(\mathbf{c})A(\mathbf{b})D(\mathbf{c}),$$

where $\Lambda(\mathbf{c}, \mathbf{b}) = \mathrm{diag}(u_n(b_0) - u(b_0), \ldots, u_n(b_n) - u(b_n))$ and $D(\mathbf{c}) = \mathrm{diag}(c_0, c_1, \ldots, c_n)$.

Assume that $c_i \neq 0$ for all $i = 0, 1, \ldots, n$, and $I + A(\mathbf{b})^{-1}D(\mathbf{c})^{-1}\Lambda(\mathbf{c}, \mathbf{b})$ is invertible. Then $\mathcal{H}(\mathbf{c}, \mathbf{b})$ is invertible and

$$\mathcal{H}(\mathbf{c}, \mathbf{b})^{-1} = \left(I + D(\mathbf{c})^{-1}A(\mathbf{b})^{-1}\Lambda(\mathbf{c}, \mathbf{b})\right)^{-1} D(\mathbf{c})^{-1}A(\mathbf{b})^{-1}D(\mathbf{c})^{-1}.$$

Table of Contents

A Damped Block Newton (dBN) Method

Let $\left(\mathbf{c}^{(k)}, \mathbf{b}^{(k)}\right)$ be the previous iterate. We then compute the current state $\left(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k+1)}\right)$ by doing the following:

(i) Compute the current linear parameters $\mathbf{c}^{(k+1)}$ solving

$$M(\mathbf{b}^{(k)})\mathbf{c} = \mathbf{u}(\mathbf{b}^{(k)}).$$

(ii) Assume that the Hessian matrix $\mathcal{H}(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})$ is invertible. Set the search direction

$$\mathbf{p}^{(k)} = -\mathcal{H}(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})^{-1} \nabla_{\mathbf{b}} \mathcal{J}(u_n(x; \mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})).$$

(iii) Compute the stepsize η_k

$$\eta_k = \operatorname*{argmin}_{\eta \in \mathbb{R}_+} \mathcal{J}(u_n(x; \mathbf{c}^{(k+1)}, \mathbf{b}^{(k)} + \eta \mathbf{p}^{(k)})).$$

Set the current nonlinear parameters by

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

Table of Contents

A Damped Block Gauss-Newton (dBGN) Method

Recall that we want to find a minimizer $u_n(x) \in \mathcal{M}_n(I)$ for the loss function

$$\mathcal{J}(v) = \frac{1}{2} \int_{\Omega} ((v(x) - u(x))^2 \, dx.$$

Since this is a least-squares problem, we can get the Gauss-Newton matrix

$$\mathcal{H}_{GN}(\mathbf{c}, \mathbf{b}) = D(\mathbf{c}) A(\mathbf{b}) D(\mathbf{c}),$$

which is positive definite when $c_i \neq 0$, for all $i = 0, \ldots, n$.

A Damped Block Gauss-Newton (dBGN) Method

Let $\left(\mathbf{c}^{(k)}, \mathbf{b}^{(k)}\right)$ be the previous iterate. We then compute the current state $\left(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k+1)}\right)$ by doing the following:

(i) Compute the current linear parameters $\mathbf{c}^{(k+1)}$ solving

$$M(\mathbf{b}^{(k)})\mathbf{c} = \mathbf{u}(\mathbf{b}^{(k)}).$$

(ii) Assume that the Gauss-Newton matrix $\mathcal{H}_{GN}(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})$ is invertible. Set the search direction

$$\mathbf{p}^{(k)} = -\mathcal{H}_{GN}(\mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})^{-1} \nabla_{\mathbf{b}} \mathcal{J}(u_n(x; \mathbf{c}^{(k+1)}, \mathbf{b}^{(k)})).$$

(iii) Compute the stepsize η_k

$$\eta_k = \underset{\eta \in \mathbb{R}_+}{\operatorname{argmin}} \, \mathcal{J}(u_n(x; \mathbf{c}^{(k+1)}, \mathbf{b}^{(k)} + \eta \mathbf{p}^{(k)})).$$

Set the current nonlinear parameters by

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

Consider the function

$$u(x) = \sqrt{x}$$

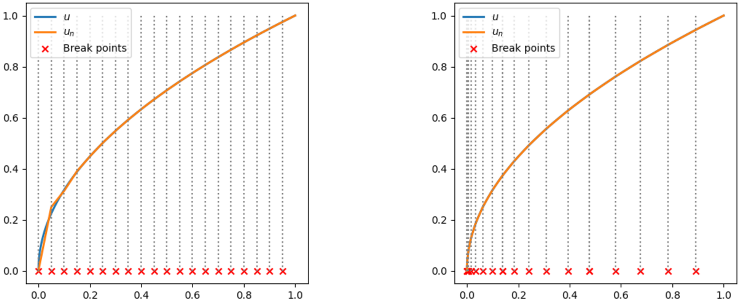
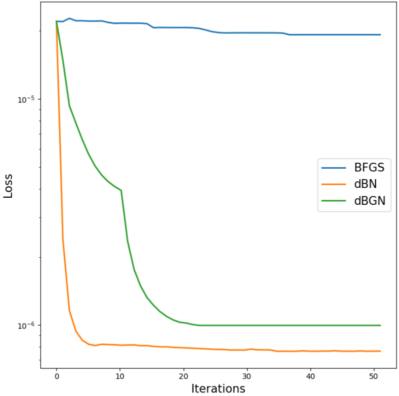
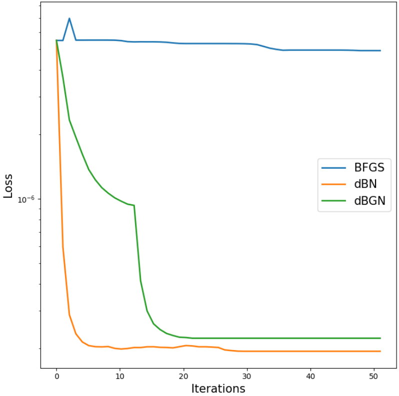


Figure: $u(x)$ approximated by NN. Left: 20 uniform breakpoints, $\mathcal{J}(u_n) = 3.17 \times 10^{-5}$. Right: optimized NN model with 20 breakpoints, 1000 iterations, $\mathcal{J}(u_n) = 6.13 \times 10^{-8}$.

dBN vs dBGN vs BFGS



(a) Loss vs number of iterations using 24 neurons. Final losses: BFGS - 1.92×10^{-5} , dBGN - 9.96×10^{-7}, dBN - 7.66×10^{-7}.



(b) Loss vs number of iterations using 48 neurons. Final losses: BFGS - 4.93×10^{-6} , dBGN - 2.24×10^{-7}, dBN - 1.94×10^{-7}.

Table of Contents

1D Diffusion Reaction Problem

We consider the 1D problem

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

Ritz formulation: find $u \in H^1(I)$ such that

$$u = \underset{\substack{v \in H^1(I) \\ v(0)=\alpha, v(1)=\beta}}{\arg\min} \left\{ \frac{1}{2} \int_0^1 (v'(x))^2 dx + \frac{1}{2} \int_0^1 (v(x))^2 dx - \int_0^1 f(x)v(x)dx \right\}$$

Modified Ritz formulation

Given $\gamma > 0$, let $J : H^1(I) \to \mathbb{R}$ be the modified energy functional given by

$$J(v) = \frac{1}{2}\int_0^1 (v'(x))^2 dx + \frac{1}{2}\int_0^1 (v(x))^2 dx - \int_0^1 f(x)v(x)dx + \frac{\gamma}{2}(v(b) - \beta)^2$$

Ritz neural network approximation: find $u_n(x) \in \mathcal{M}_n(I)$ such that

$$J(u_n) = \min_{\substack{v \in \mathcal{M}_n(I) \\ v(0)=\alpha}} J(v)$$

Error estimate

Proposition

Let u be the exact solution and $u_n \in \mathcal{M}_n(I)$ be the Ritz neural network approximation. There exists a constant C depending on u such that

$$\|u - u_n\|_a \leq C\left(n^{-1} + \gamma^{-1/2}\right),$$

where $\|v\|_a^2 = \int_0^1 (v'(x))^2 dx + \int_0^1 (v(x))^2 dx + \gamma(v(1))^2$.

Systems of algebraic equations

Let

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

be a solution of the previous minimization problem. Then the linear and nonlinear parameters

$$\mathbf{c} = (c_0, \ldots, c_n)^T \quad \text{and} \quad \mathbf{b} = (b_0, \ldots, b_n)^T$$

satisfy the following system of algebraic equations

$$\nabla_{\mathbf{c}} J(u_n) = \mathbf{0} \quad \text{and} \quad \nabla_{\mathbf{b}} J(u_n) = \mathbf{0}.$$

Linear parameters $\mathbf{c}$

The equation $\nabla_{\mathbf{c}} J(u_n) = 0$ has the form

$$\left(A(\mathbf{b}) + M(\mathbf{b}) + \gamma \mathbf{d}\mathbf{d}^T\right) \mathbf{c} = \mathbf{f}(\mathbf{b}) + \gamma(\beta - \alpha)\mathbf{d}.$$

where

- $A(\mathbf{b})$ is the coefficient matrix
- $M(\mathbf{b})$ is the mass matrix
- $\mathbf{f}(\mathbf{b}) = \left(\int_0^1 f(x)\sigma(x - b_0)dx, \ldots, \int_0^1 f(x)\sigma(x - b_n)dx\right)^T$
- $\mathbf{d} = (b - b_0, \ldots, b - b_n)^T$

Linear parameters c

Let $h_i = b_i - b_{i-1}$ for $i = 0, \dots, n$. Define the matrices

$$G = \begin{pmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & & \ddots & \\ & & & -1 & 1 \end{pmatrix}, \quad D_h(\mathbf{b}) = \begin{pmatrix} h_0 & & & \\ & h_1 & & \\ & & \ddots & \\ & & & h_n \end{pmatrix},$$

and the tridiagonal matrices

$$T_1 = T_1(\mathbf{b}) = G D_h(\mathbf{b})^{-1} G, \quad T_3 = T_3(\mathbf{b}) = G^T D_h(\mathbf{b})^{-1} G \ ,$$

$$T_2 = T_2(\mathbf{b}) = \frac{1}{6} \begin{pmatrix} 2(h_1 + h_2) & h_2 & & \\ h_2 & 2(h_2 + h_3) & h_3 & \\ & & \ddots & \\ & & h_{n+1} & 2h_{n+1} \end{pmatrix}.$$

Linear parameters c

The following factorization holds:

$$M(\mathbf{b}) + A(\mathbf{b}) = T_1^{-T}(T_2 + T_3)T_1^{-1},$$

i.e.,

$$(M(\mathbf{b}) + A(\mathbf{b}))^{-1} = T_1(T_2 + T_3)^{-1}T_1.$$

So that the linear system

$$\left(A(\mathbf{b}) + M(\mathbf{b}) + \gamma\mathbf{d}\mathbf{d}^T\right)\mathbf{c} = \mathbf{f}(\mathbf{b}) + \gamma(\beta - \alpha)\mathbf{d}.$$

can be solved in $30(n+1)$ operations.

Nonlinear parameters $\mathbf{b}$

Lemma

For $j = 0, 1, \ldots, n$, let

$$g(b_j) = u_n(b_j) - f(b_j),$$

and $\mathbf{B}(\mathbf{c}, \mathbf{b}) = \mathrm{diag}(g(b_0), \ldots, g(b_n))$. Let $D(\mathbf{c}) = \mathrm{diag}(c_0, c_1, \ldots, c_n)$. Then the Hessian matrix $\nabla^2_{\mathbf{b}} J(u_n)$ has the form

$$\mathbf{H}(\mathbf{c}, \mathbf{b}) = \tilde{\mathbf{H}}(\mathbf{c}, \mathbf{b}) + \gamma \mathbf{c}\mathbf{c}^T = D(\mathbf{c})\mathbf{B}(\mathbf{c}, \mathbf{b}) + D(\mathbf{c})A(\mathbf{b})D(\mathbf{c}) + \gamma \mathbf{c}\mathbf{c}^T.$$

Assume that $c_i \neq 0$ for all $i = 0, 1, \ldots, n$, and $I + A(\mathbf{b})^{-1}D(\mathbf{c})^{-1}\mathbf{B}(\mathbf{c}, \mathbf{b})$ is invertible. Then $\tilde{\mathbf{H}}$ is invertible and

$$\tilde{\mathbf{H}}^{-1} = \left(I + D(\mathbf{c})^{-1}A(\mathbf{b})^{-1}\mathbf{B}(\mathbf{c}, \mathbf{b})\right)^{-1} D(\mathbf{c})^{-1}A(\mathbf{b})^{-1}D(\mathbf{c})^{-1}.$$

The first test problem involves the function

$$u(x) = x\left(\exp\left(-\frac{(x-\frac{1}{3})^2}{0.01}\right) - \exp\left(-\frac{4}{9\times 0.01}\right)\right)$$

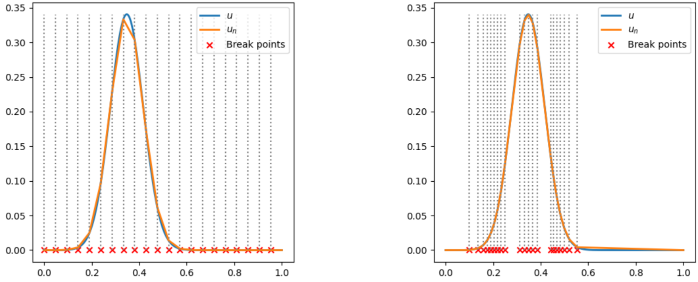
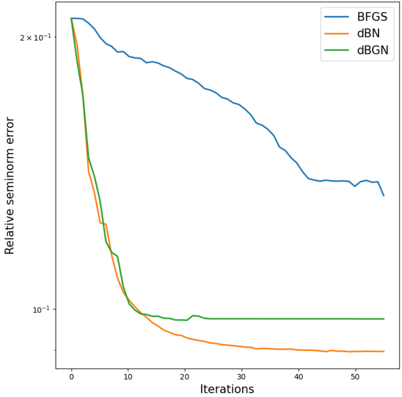
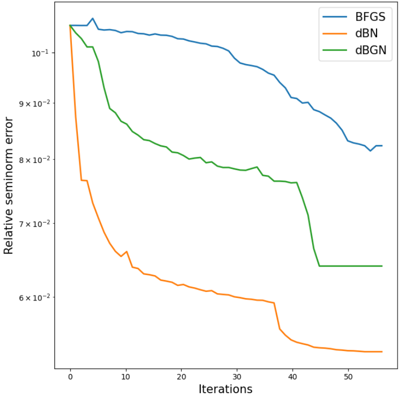


Figure: $u(x)$ approximated by NN. Left: 21 uniform breakpoints, $e_n = 0.238$. Right: optimized NN model with 21 breakpoints, 500 iterations, $e_n = 0.101$.



(a) $\frac{|u-u_n|_1}{|u|_1}$ vs number of iterations using 24 neurons. Final relative errors: BFGS - 0.133, dBGN - 0.097, dBN - 0.090.



(b) $\frac{|u-u_n|_1}{|u|_1}$ vs number of iterations using 48 neurons. Final relative errors: BFGS - 0.082, dBGN - 0.064, dBN - 0.053.

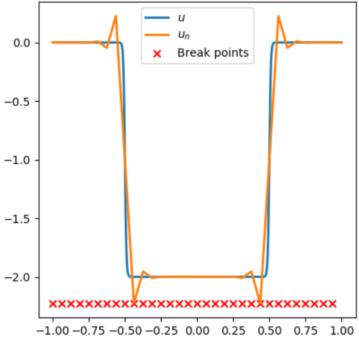
Singularly Perturbed Reaction-Diffusion Equation

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

For $f(x) = -2\left(\varepsilon - 4x^2 \tanh\left(\frac{1}{\varepsilon}(x^2 - \frac{1}{4})\right)\right)\left(1/\cosh\left(\frac{1}{\varepsilon}(x^2 - \frac{1}{4})\right)\right)^2 + \tanh\left(\frac{1}{\varepsilon}(x^2 - \frac{1}{4})\right) - \tanh\left(\frac{3}{4\varepsilon}\right)$, this problem has the following exact solution

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

Singularly Perturbed Reaction-Diffusion Equation



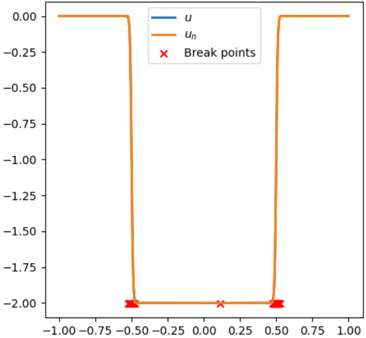


Figure: $u(x)$ approximated by NN. $\varepsilon = 0.01$ Left: 32 uniform breakpoints, $e_n = 0.889$. Right: optimized NN model with 32 breakpoints, 500 iterations, $e_n = 0.090$.

Summary/Future work

Key components of our methods:

- We know how to invert the coefficient matrix and the mass matrix.
- We utilize the geometric meaning of the nonlinear parameters to obtain a good initial approximation.

Future work: Two-dimensional problems.

Thanks!