1 Set up

Let us consider the second order elliptic equation

$$\mathcal{L}v = f \text{ in } \Omega, \tag{1.1}$$

where $\Omega \in \mathbb{R}^d$ is an open subset, $d \geq 1$. $\mathcal{L}$ is the second order elliptic operator defined by

$$\mathcal{L}v = -\sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j}\frac{\partial v}{\partial x_i}\frac{\partial v}{\partial x_j} + \sum_{k=1}^{d} b_k \frac{\partial v}{\partial x_k} + cv. \quad \forall v \in C^2(\Omega) \tag{1.2}$$

In order to ensure the existence and uniqueness of the weak solution for problem (1.1), the operator $\mathcal{L}$ should be assumed to satisfy

$$a_{i,j} \in L^\infty(\Omega),\ 1 \leq i, j \leq d, \tag{1.3}$$

and there exist positive constants λ and Λ, such that

$$\lambda|\boldsymbol{\xi}|^2 \leq \sum_{i=1}^{d}\sum_{j=1}^{d} a_{i,j}(x)\xi_i\xi_j \leq \Lambda|\boldsymbol{\xi}|^2,\ \forall \boldsymbol{\xi} \in \mathbb{R}^d,\ x \in \Omega. \tag{1.4}$$

For the simplicity of notations, we consider the following second-order partial differential equation: (at the end of the article, we need a remark to claim that all the results can be applied on the general form (1.2) with simply Dirichlet BC)

$$-a\Delta v(x) + cv(x) = f(x),\ \text{in } \Omega \tag{1.5}$$

$$\frac{\partial v}{\partial n} = 0,\ \text{on } \partial\Omega \tag{1.6}$$

where $a \geq 0, c \geq 0$ are constants, so that the conditions (1.3)-(1.4) are satisfied. Moreover, we define the admissible set on Ω

$$V = H^1(\Omega). \tag{1.7}$$

Then the PDE (1.5) is equivalent to the following energy minimization:

$$\min_{w \in V} \mathcal{J}(w) = \int_\Omega \left(\frac{a}{2}|\nabla w(x)|^2 + \frac{c}{2}w(x)^2 - f(x)w(x)\right) dx. \tag{1.8}$$

When $a \neq 0$, minimizing (1.8) is equivalent to solving the PDE (1.5); while $a = 0$ but $c \neq 0$, (1.8) degenerates into a function approximation problem. The minimization task can be finished by learning algorithms such as the deep neural network(DNN), which is referred to as the Deep Ritz method in [1]. In practice, the energy integral can be computed by means of the numerical quadratures, such as the Gauss quadrature and the Monte-Carlo method.

2 Preliminaries

In this section, we will give an introduction of the ReLU^k-DNN model and a convergence result of the Deep Ritz method. But the discussion in the following does not specify the form of activation functions. Firstly, let us consider about the energy minimization problem (1.8). Let the minimizer be

$$v = \arg\min_{w \in V} \mathcal{J}(w) = a(w, w) - (f, w), \tag{2.1}$$

where $a(w, z) = \int_\Omega \left(\frac{a}{2} \nabla w \cdot \nabla z + \frac{c}{2} wz \right) dx$ is a symmetric bilinear form.

Lemma 1. *Suppose v is the minimizer of (1.8) or the solution of (1.5), then it holds*

$$\mathcal{J}(u) - \mathcal{J}(v) = \|u - v\|_a^2. \tag{2.2}$$

for $\forall u \in V$. Here

$$\|u\|_a^2 = a(u, u). \tag{2.3}$$

Proof. Let v be the solution of (1.5) and $u \in V$ be arbitrary. Since v is the minimizer, then we have that

$$0 = \delta \mathcal{J}(v)[u] = 2a(v, u) - (f, u), \quad \forall u \in V. \tag{2.4}$$

Then by a simple computation, it is easy to see that

$$\begin{aligned} \mathcal{J}(u) - \mathcal{J}(v) &= a(u, u) - (f, u) - a(v, v) + (f, v) && (2.5)\\ &= a(u, u) - (f, u) + a(v, v) && (2.6)\\ &= a(u, u) - 2a(v, u) + a(v, v) && (2.7)\\ &= \|u - v\|_a^2 && (2.8) \end{aligned}$$

□

Remark 1. *In general, the Deep Ritz method for second-order partial differential equations asks for an admissible set with at least H^1 regularity. With different training algorithms, more regularity might be required for this method. We therefore need an appropriate choice for the activation function in DNN models to make sure they are H^k functions. For instance, let the ReLU^k-DNN with one hidden layer be denoted by $V_N^k = \{\sum_{i=1}^N a_i ReLU^k (w_i x + b_i), a_i, b_i \in \mathbb{R}^1, w_i \in \mathbb{R}^{1\times d}\}$, then we have $V_N^k(\Omega) \subset H^k(\Omega), k \geq 1$. The gradient method will require first-order derivatives of w_i and b_i inside the integration of (1.8), thus $k \geq 2$ as the power of ReLU.*

Now Let the DNN model be denoted by $u(x, \theta)$, where $x \in \Omega$ and $\theta \in \mathbb{R}^m$. Here the dimension m of the parameter space is related to the number of neurons in specific models. The admissible set on Ω for the energy functional is then changed into the space of all the ReLU^k-DNN functions with J hidden layers, which is denoted by DNN_J. In the following, we always assume that the DNN function space $DNN_J \subset V = H^1$, which suggests that the DNN model we use should be at least a C^0 function. We have the approximation result for DNN_J:

Lemma 2. *For $\forall \epsilon > 0$, there exist $J \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$, such that $u(\cdot, \theta) \in DNN_J$, $\theta \in \mathbb{R}^m$, and for θ^* being the minimizer in $\mathbb{R}^m$:*

$$\theta^* = \arg\min_{\theta} \mathcal{J}(\theta) = \int_\Omega \left(\frac{a}{2} |\nabla u(x,\theta)|^2 + \frac{c}{2} u(x,\theta)^2 - f(x) u(x,\theta) \right) dx, \tag{2.9}$$

it holds

$$\|u(\cdot,\theta^*) - v\|_{L^2(\Omega)} \lesssim \epsilon \tag{2.10}$$

$$\left|u(\cdot,\theta^*) - v\right|_{H^1(\Omega)} \lesssim \epsilon. \tag{2.11}$$

We will use the notation "$A \lesssim B$" to denote $A \leq CB$ for some constant C independent of crucial parameters such as θ.

Proof. From the Weierstrass theorem in Sobolev space $W_p^1 (1 \leq p \leq \infty)$ and the fact that ReLU^k-DNN functions are piecewise polynomials that belong to H^1, it can be verified that there exist $J \in \mathbb{N}^+$, $m \in \mathbb{N}^+$ and $\theta \in \mathbb{R}^m$, such that $u(\cdot,\theta) \in DNN_J$, and

$$\|u(\cdot,\theta) - v\|_a \lesssim \|u(\cdot,\theta) - v\|_{H^1(\Omega)} < \epsilon, \tag{2.12}$$

In particular, when considering ReLU^1-DNN functions, one can refer to [2] for estimations of a sufficiently large depth J and a sufficiently large number of neurons to represent all the piecewise linear functions in $\mathbb{R}^d$.

Next, since $DNN_J \subset V$, it holds that

$$\mathcal{J}(u(\cdot,\theta)) \geq \mathcal{J}(u(\cdot,\theta^*)) \geq \mathcal{J}(v), \tag{2.13}$$

which, together with Lemma 1, implies that

$$\|u(\cdot,\theta^*) - v\|_{L^2(\Omega)} \lesssim \|u(\cdot,\theta^*) - v\|_a \leq \|u(\cdot,\theta) - v\|_a < \epsilon \tag{2.14}$$

$$\left|u(\cdot,\theta^*) - v\right|_{H^1(\Omega)} \lesssim \|u(\cdot,\theta^*) - v\|_a \leq \|u(\cdot,\theta) - v\|_a < \epsilon. \tag{2.15}$$

□

Based on the results above, we have an estimation for the residual function:

Lemma 3. *For $\forall \epsilon > 0$, there exist $\delta > 0$, $J \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$, such that $\theta \in \mathbb{R}^m$, $u(\cdot,\theta) \in DNN_J \subset H^1(\Omega)$, and for θ^* being defined by (2.9), $\|\theta - \theta^*\| < \delta$, it holds*

$$\|u(\cdot,\theta) - v\|_{L^2(\Omega)} < \epsilon \tag{2.16}$$

and

$$\int_\Omega (a \nabla u(x,\theta) \cdot \nabla w(x) + c u(x,\theta) w(x) - f(x) w(x))\, dx < C\epsilon, \quad \forall w \in H^1(\Omega) \tag{2.17}$$

where $C > 0$ is independent of θ.

Proof. By the continuity of $u(\cdot,\theta)$ with respect to θ, it is easy to see from Lemma 2 that

$$\|u(\cdot,\theta)-v\|_{L^2(\Omega)} \le \|u(\cdot,\theta)-u(\cdot,\theta^*)\|_{L^2(\Omega)} + \|u(\cdot,\theta^*)-v\|_{L^2(\Omega)} < \epsilon. \tag{2.18}$$

Next, we have for $\forall w \in H^1(\Omega)$,

$$\begin{aligned}(r(u(\cdot,\theta)),w)_{L^2(\Omega)} &= \int_\Omega (a\nabla u(x,\theta)\cdot\nabla w(x) + cu(x,\theta)w(x) - f(x)w(x))\,dx &(2.19)\\ &= \int_\Omega (a\nabla(u(x,\theta)-v(x))\cdot\nabla w(x) + c(u(x,\theta)-v(x))w(x))\,dx &(2.20)\\ &\lesssim \big|u(\cdot,\theta)-v\big|_{H^1(\Omega)}\big|w\big|_{H^1(\Omega)} + \|u(\cdot,\theta)-v\|_{L^2(\Omega)}\|w\|_{L^2(\Omega)}. &(2.21)\end{aligned}$$

Since $w \in H^1(\Omega)$, we can conclude form Lemma 2 that the estimation (2.17) is valid. □

3 Gauss-Newton Iteration with numerical quadrature

3.1 The collocation problem

Let $u(x,\theta)$ be a learning model to approximate the exact solution $v(x)$. In order to solve the PDE with learning algorithms, one possible way is to construct an alternative problem with collocation points $\{x_1,x_2,\cdots,x_N\}\subset\Omega$ and $\{x_1^b,x_2^b,\cdots,x_n^b\}\subset\partial\Omega$, such that :

$$\boldsymbol{F}(x,\theta) = \begin{pmatrix} -a\Delta u(x_1,\theta)+cu(x_1,\theta)-f(x_1)\\ -a\Delta u(x_2,\theta)+cu(x_2,\theta)-f(x_2)\\ \vdots\\ -a\Delta u(x_N,\theta)+cu(x_N,\theta)-f(x_N)\\ \nabla u(x_1^b,\theta)\cdot\boldsymbol{n}\\ \nabla u(x_2^b,\theta)\cdot\boldsymbol{n}\\ \vdots\\ \nabla u(x_n^b,\theta)\cdot\boldsymbol{n}\end{pmatrix} = \mathbf{0}. \tag{3.1}$$

where $\boldsymbol{n}$ is the outer normal unit vector of $\partial\Omega$ and $\theta=\{\theta_1,\theta_2,\cdots,\theta_m\}$ represents the parameters in the learning model $u(x,\theta)$. Therefore, an optimal choice of θ should be the one that minimizes $\|\boldsymbol{F}(x,\theta)\|$. Let us consider the Gauss-Newton method for solving the collocation point problem (3.1), which can be written as

$$\theta_{k+1} = \theta_k - (\mathbf{JF}(x,\theta_k))^\dagger \boldsymbol{F}(x,\theta_k), \quad k=0,1,2,\cdots \tag{3.2}$$

where $\mathbf{JF}$ is the Jacobi matrix, $\epsilon_1>0$ is a sufficiently small constant and the matrix inverse is the Moore-Penrose inverse. Hereafter the gradient operator will always be regarded as a column vector for both θ and $\boldsymbol{x}$. As for the matrix of mixed second-order derivatives, we define $\nabla_\theta\nabla$ to be

$$\begin{bmatrix} \partial_{x_1}\partial_{\theta_1} & \partial_{x_2}\partial_{\theta_1} & \cdots & \partial_{x_d}\partial_{\theta_1}\\ \partial_{x_1}\partial_{\theta_2} & \partial_{x_2}\partial_{\theta_2} & \cdots & \partial_{x_d}\partial_{\theta_2}\\ \vdots & \vdots & \cdots & \vdots\\ \partial_{x_1}\partial_{\theta_m} & \partial_{x_2}\partial_{\theta_m} & \cdots & \partial_{x_d}\partial_{\theta_m}\end{bmatrix}. \tag{3.3}$$

So we have

$$\mathbf{JF} = \begin{bmatrix} -\nabla_\theta \Delta u(x,\theta)^T + \nabla_\theta u(x,\theta)^T \\ \left(\nabla_\theta \nabla u(x^b,\theta)\cdot \boldsymbol{n}\right)^T \end{bmatrix} \in \mathbb{R}^{(N+n)\times m}. \tag{3.4}$$

3.2 The variational problem

Next let us write the Newton iteration of the Deep Ritz method with Gauss quadrature rule. The loss function is defined as follows

$$L(\theta) = \int_\Omega \frac{1}{2}|\nabla u(x,\theta)|^2 + \frac{1}{2}u(x,\theta)^2 - f(x)u(x,\theta)dx. \tag{3.5}$$

So the minimization will be considered within a smaller admissible set, i.e., the DNN function space. We can compute the gradient of $L(\theta)$:

$$\nabla_\theta L(\theta) = \int_\Omega \nabla_\theta \nabla u(x,\theta)\nabla u(x,\theta) + u(x,\theta)\nabla_\theta u(x,\theta) - f(x)\nabla_\theta u(x,\theta)dx \tag{3.6}$$

Then it can be verified that the formula of the Hessian of $L(\theta)$, which is denoted by $\mathbf{HL}(\theta)$, is as follows:

$$\mathbf{HL}(\theta) = \int_\Omega \nabla_\theta \nabla u(x,\theta)\cdot\nabla_\theta \nabla u(x,\theta)^T + \nabla_\theta u(x,\theta)\cdot\nabla_\theta u(x,\theta)^T dx \tag{3.7}$$

$$+ \int_\Omega \nabla_\theta^2 \nabla u(x,\theta)\cdot\nabla u(x,\theta) + u(x,\theta)\nabla_\theta^2 u(x,\theta) - f(x)\nabla_\theta^2 u(x,\theta)dx \tag{3.8}$$

$$:= \boldsymbol{J}(\theta) + \boldsymbol{Q}(\theta), \tag{3.9}$$

The Hessian can be taken apart into two matrix $\boldsymbol{J}(\theta)$ and $\boldsymbol{Q}(\theta)$, standing for the first-order and second-order component of θ respectively. As a straight consequence of Lemma 3, we have

Lemma 4. *For $\forall \epsilon > 0$, there exist $\delta > 0$, $J \in \mathbb{N}^+$ and $m \in \mathbb{N}^+$, such that $\theta \in \mathbb{R}^m$, $DNN_J \subset H^1(\Omega)$, and for $\|\theta - \theta^*\| < \delta$, θ^* being defined by (2.9), if $D_\theta^2 u(x,\theta) \in H^1(\Omega)$, then it holds*

$$\|\boldsymbol{Q}(\theta)\| < \epsilon. \tag{3.10}$$

Therefore, it is reasonable to consider the first-order approximation of the Hessian $\boldsymbol{J}(\theta) \approx \mathbf{HL}(\theta)$. The Gauss-Newton iteration for the variational problem is then written as

$$\theta_{k+1} = \theta_k - \boldsymbol{J}(\theta_k)^\dagger \nabla_\theta L(\theta_k), \quad k = 0, 1, 2, \cdots \tag{3.11}$$

where the matrix inverse is the Moore-Penrose inverse. Again, we will present an appropriate perturbed Gauss-Newton iteration for the variational problem in the following section.

3.3 The consistency between the collocation and variational problem

Again, by the divergence theorem on $\boldsymbol{J}(\theta)$, we have

$$\boldsymbol{J}(\theta) = \int_\Omega \nabla_\theta u(x,\theta)\cdot\left(-\nabla_\theta \Delta u(x,\theta) + \nabla_\theta u(x,\theta)\right)^T dx$$

$$+ \int_{\partial\Omega} \nabla_\theta u(x,\theta)\cdot \frac{\partial \nabla_\theta u(x,\theta)^T}{\partial \boldsymbol{n}} dS. \tag{3.12}$$

Therefore, if all the integrals are computed by the Gauss quadrature rule, then the iteration for the Deep Ritz method is consistent with the iteration of the collocation problem (3.2). In fact, with the quadrature points $\boldsymbol{x} = (x_1, x_2, \cdots, x_N)^T$ and the quadrature weights $\boldsymbol{w} = (w_1, w_2, \ldots, w_N)^T$, the first part of (3.12) can be computed through

$$\int_\Omega \nabla_\theta u(x,\theta) \cdot (-\nabla_\theta \Delta u(x,\theta) + \nabla_\theta u(x,\theta))^T \, dx = \sum_{i=1}^{N} w_i \nabla_\theta u(x_i,\theta) \cdot (-\nabla_\theta \Delta u(x_i,\theta) + \nabla_\theta u(x_i,\theta))^T . \tag{3.13}$$

With the boundary quadrature points $\boldsymbol{x^b} = (x_1^b, x_2^b, \cdots, x_n^b)^T$ and the weights $\boldsymbol{w^b} = (w_1^b, w_2^b, \cdots, w_n^b)^T$, the second part of (3.12) can be computed by

$$\int_{\partial\Omega} \nabla_\theta u(x,\theta) \cdot \frac{\partial \nabla_\theta u(x,\theta)}{\partial \boldsymbol{n}}^T \, dS = \sum_{j=1}^{n} w_j^b \nabla_\theta u(x_j,\theta) \cdot \frac{\partial \nabla_\theta u(x_j,\theta)}{\partial \boldsymbol{n}}^T \tag{3.14}$$

Therefore, we have

$$\boldsymbol{J}(\theta) = \begin{bmatrix} w_1 \nabla_\theta u(x_1,\theta) & \cdots & w_N \nabla_\theta u(x_N,\theta) & w_1^b \nabla_\theta u(x_1^b,\theta) & \cdots & w_n^b \nabla_\theta u(x_n^b,\theta) \end{bmatrix} \mathbf{JF} := \boldsymbol{G} \cdot \mathbf{JF} \tag{3.15}$$

where **JF** is the Jacobian defined in (1.3). Similarly, we can compute the integral (1.5) by integral by parts and the quadrature rule:

$$\nabla_\theta L(\theta) = \int_\Omega \nabla_\theta u(x,\theta) \left(-\Delta u(x,\theta) + u(x,\theta) - f(x)\right) dx \tag{3.16}$$

$$+ \int_{\partial\Omega} \nabla_\theta u(x,\theta) \frac{\partial \nabla_\theta u(x,\theta)}{\partial \boldsymbol{n}} dS \tag{3.17}$$

$$= \boldsymbol{G} \cdot \boldsymbol{F}(\boldsymbol{x},\theta) \tag{3.18}$$

Therefore, in the sense of the numerical quadrature and the pseudo-inverse, we have

$$(\boldsymbol{J}(\theta))^\dagger \cdot \nabla_\theta L(\theta) = (\boldsymbol{G} \cdot \mathbf{JF})^\dagger \cdot \boldsymbol{G} \cdot \boldsymbol{F}(\boldsymbol{x},\theta) = (\mathbf{JF})^\dagger \cdot (\boldsymbol{G}^\dagger \boldsymbol{G}) \cdot \boldsymbol{F}(\boldsymbol{x},\theta), \tag{3.19}$$

which shows that (3.11) for the variational problem is consistent with (3.2) for the collocation problem.

Remark 2. *As we refer to the Moore-Penrose inverse, the assumption that G has linearly independent columns is needed in order to guarantee $G^\dagger G = I \in \mathbb{R}^{(N+n)\times(N+n)}$. This is possible in the case that the number of quadrature points $N + n$ is less equal than that of θ.*

Remark 3. *The consistency between (3.2) and (3.11) reveals that the Gauss quadrature points will be a good choice for the collocation problem, since the equivalent iteration comes form a numerical integration which is much more accurate under the Gauss quadrature rule.*

3.4 Convergence analysis

In this section, we will analyse the convergence property of the following Gauss-Newton iteration:

$$\theta_{k+1} = \theta_k - \boldsymbol{J}(\theta_k)^\dagger \nabla_\theta L(\theta_k), \quad k = 0, 1, 2, \cdots \tag{3.20}$$

Firstly, we need the following Wedin's Theorem about a perturbed matrix.

Lemma 5. *(Wedin, see [3] or [4]) Let $A \in \mathbb{R}^{m\times n}$ and $B = A + E$ with $rank(B) = rank(A)$. Then in any unitary invariant norm $\|\cdot\|$,*

$$\|B^\dagger - A^\dagger\| \le \mu \|A^\dagger\|_2 \|B^\dagger\|_2 \|E\| \tag{3.21}$$

for some moderate constant $\mu > 0$ that depends on the norm used. Moreover, if

$$\|E\|_2 < \frac{1}{\|A^\dagger\|_2}, \tag{3.22}$$

then the inverse of B can be bounded by

$$\|B^\dagger\|_2 \le \frac{\|A^\dagger\|_2}{1 - \|A^\dagger\|_2 \|E\|_2} \tag{3.23}$$

Remark 4. *An unitary invariant norm $\|\cdot\|$ satisfies $\|A\| = \|UAV\|$ for any unitary matrices U and V. Specifically, the 2-norm, the Frobenius norm and many other norms related to the singular value are unitary invariant norms.*

Therefore, we can estimate the convergence rate of the Gauss-Newton method based on the Lemmas and thanks to the consistency between two problems, we can consider it for the variational problem only. Now assume that the variational problem has stationary point θ^* such that $\nabla_\theta L(\theta^*) = 0$, and assume rank $\boldsymbol{J}(\theta^*) = r \le m$. From the singular decomposition, we have $J(\theta^*) = U\Sigma V^T$ and $J(\theta^*)$ is a r-rank semi-positive definite matrix, i.e.,

$$\Sigma = diag\left([\sigma_1, \cdots, \sigma_r, 0, \cdots, 0]\right), \text{ with } \sigma_1 \ge \sigma_2 \ge \cdots \ge \sigma_r > 0,\ r \le m. \tag{3.24}$$

Then the pseudo-inverse can be represented by

$$\boldsymbol{J}(\theta^*)^\dagger = V\Sigma^\dagger U^T, \text{ with } \Sigma^\dagger = diag\left(\left[\frac{1}{\sigma_1}, \cdots, \frac{1}{\sigma_r}, 0, \cdots, 0\right]\right). \tag{3.25}$$

Lemma 6. *Let $\boldsymbol{J}(\theta)$ be the approximated Hessian defined in (3.7) and θ^* be the stationary point. Then there exits a small open set $\Omega_* \ni \theta^*$ and constants $\epsilon, \zeta, \alpha > 0$ such that for every $\theta_1, \theta_2 \in \Omega_*$, the following inequalities holds:*

$$\|\boldsymbol{J}(\theta_2)\boldsymbol{J}(\theta_2)^\dagger - \boldsymbol{J}(\theta_1)\boldsymbol{J}(\theta_1)^\dagger\| \le \zeta\|\theta_2 - \theta_1\|, \tag{3.26}$$

$$\|\nabla_\theta L(\theta_2) - \nabla_\theta L(\theta_1) - \boldsymbol{J}(\theta_1)(\theta_2 - \theta_1)\| \le \epsilon\|\theta_2 - \theta_1\| + \alpha\|\theta_2 - \theta_1\|^2. \tag{3.27}$$

Proof. From the Taylor's expansion, we have

$$\nabla_\theta L(\theta_2) - \nabla_\theta L(\theta_1) - \boldsymbol{J}(\theta_1)(\theta_2 - \theta_1) \tag{3.28}$$

$$= \mathbf{HL}(\theta_1)(\theta_2 - \theta_1) - \boldsymbol{J}(\theta_1)(\theta_2 - \theta_1) + \mathcal{O}(\|\theta_2 - \theta_1\|^2) \tag{3.29}$$

$$= \boldsymbol{Q}(\theta_1)(\theta_2 - \theta_1) + \mathcal{O}(\|\theta_2 - \theta_1\|^2). \tag{3.30}$$

Therefore, by the smoothness of the Hessian with respect to θ and Lemma 4, there exists a small neighbour Ω_* of θ^* and a constant $\alpha > 0$, such that $\|\boldsymbol{Q}(\theta_1)\| \le \epsilon$, and

$$\|\nabla_\theta L(\theta_2) - \nabla_\theta L(\theta_1) - \boldsymbol{J}(\theta_1)(\theta_2 - \theta_1)\| \le \epsilon\|\theta_2 - \theta_1\| + \alpha\|\theta_2 - \theta_1\|^2. \tag{3.31}$$

for every $\theta_1, \theta_2 \in \Omega_*$. On the other hand, Weyl's Theorem guarantees the singular value to be continuous with respect to the matrix entries. Therefore, as the open set Ω_* being sufficiently small, it holds

$$\sup_{\theta\in\Omega_*} \|\boldsymbol{J}(\theta)\| = \sup_{\theta\in\Omega_*} \sigma_1(\boldsymbol{J}(\theta)) \le C\|\boldsymbol{J}(\theta^*)\| \tag{3.32}$$

$$\sup_{\theta\in\Omega_*} \|\boldsymbol{J}(\theta)^\dagger\| = \frac{1}{\sup_{\theta\in\Omega_*} \sigma_r(\boldsymbol{J}(\theta))} \le C\|\boldsymbol{J}(\theta^*)^\dagger\| \tag{3.33}$$

for all $\theta \in \Omega_*$. By the smoothness of $\boldsymbol{J}(\theta)$ with respect to θ and the estimation in Lemma 5, we have

$$\|\boldsymbol{J}(\theta_2)\boldsymbol{J}(\theta_2)^\dagger - \boldsymbol{J}(\theta_1)\boldsymbol{J}(\theta_1)^\dagger\| \tag{3.34}$$

$$\le \|\boldsymbol{J}(\theta_2)^\dagger\|\|\boldsymbol{J}(\theta_2) - \boldsymbol{J}(\theta_1)\| + \|\boldsymbol{J}(\theta_1)\|\|\boldsymbol{J}(\theta_2)^\dagger - \boldsymbol{J}(\theta_1)^\dagger\| \tag{3.35}$$

$$\le \zeta\|\theta_2 - \theta_1\|. \tag{3.36}$$

□

Theorem 1. *(Convergence Theorem) Let $L(\theta)$ be a sufficiently smooth target function of θ and $\boldsymbol{J}(\theta)$ be the approximated Hessian of $L(\theta)$ defined in (3.7). Then for every open neighbourhood Ω_1 of θ^*, there exists another neighbourhood $\Omega_2 \ni \theta^*$ such that, from every initial guess $\theta_0 \in \Omega_2$, the sequence $\{\theta_k\}_{k=1}^\infty$ generated by the iteration (3.11) converges in Ω_1. Furthermore, $\{\theta_k\}_{k=1}^\infty$ has at least linear convergence with a coefficient $\gamma \le 2\epsilon$ when k is large enough.*

Proof. Firstly, Let Ω_* be the small open neighbourhood in Lemma 6 such that $\|\boldsymbol{Q}(\theta)\| \le \epsilon < \frac{1}{2}$. For any open neighbourhood Ω_1 of θ^*, there exists a constant $0 < \delta < 2$ such that $B(\theta^*, \delta) \subset \Omega_1 \cap \Omega_*$ and for any $\theta_1, \theta_2 \in B(\theta^*, \delta)$, it holds

$$\|\boldsymbol{J}(\theta_2)^\dagger\| \left(\alpha\|\theta_2 - \theta_1\| + \zeta\|\nabla_\theta L(\theta_1)\|\right) \le h < 1. \tag{3.37}$$

On the other hand, there also exists $0 < \tau < \frac{\delta}{2}$ such that

$$\|\boldsymbol{J}(\theta)^\dagger\|\|\nabla_\theta L(\theta)\| \le \frac{1-h}{2}\delta < \frac{\delta}{2} < 1 \tag{3.38}$$

holds for any $\theta \in B(\theta^*, \tau)$. Let $\Omega_2 = B(\theta^*, \tau)$, then for $\forall\theta_0 \in \Omega_2$, the iteration implies

$$\|\theta_1 - \theta^*\| \le \|\theta_1 - \theta_0\| + \|\theta_0 - \theta^*\| \le \|\boldsymbol{J}(\theta_0)^\dagger\|\|\nabla_\theta L(\theta_0)\| + \tau < \delta. \tag{3.39}$$

Next we assume that $\theta_k \in B(\theta^*, \delta)$ for some integer $k \ge 1$. From Lemma 6 we have the following estimation

$$\|\theta_{k+1} - \theta_k\| = \|\boldsymbol{J}(\theta_k)^\dagger\nabla_\theta L(\theta_k)\| \tag{3.40}$$

$$= \|\boldsymbol{J}(\theta_k)^\dagger \left(\nabla_\theta L(\theta_k) - \boldsymbol{J}(\theta_{k-1}) \left(\theta_k - \theta_{k-1} + \boldsymbol{J}(\theta_{k-1})^\dagger\nabla_\theta L(\theta_{k-1})\right)\right) \| \tag{3.41}$$

$$\le \|\boldsymbol{J}(\theta_k)^\dagger \left(\nabla_\theta L(\theta_k) - \nabla_\theta L(\theta_{k-1}) - \boldsymbol{J}(\theta_{k-1})(\theta_k - \theta_{k-1})\right) \| \tag{3.42}$$

$$+ \|\boldsymbol{J}(\theta_k)^\dagger \left(\boldsymbol{J}(\theta_k)\boldsymbol{J}(\theta_k)^\dagger - \boldsymbol{J}(\theta_{k-1})\boldsymbol{J}(\theta_{k-1})^\dagger\right) \nabla_\theta L(\theta_{k-1})\| \tag{3.43}$$

$$\le \|\boldsymbol{J}(\theta_k)^\dagger\| \left(\alpha\|\theta_k - \theta_{k-1}\| + \zeta\|\nabla_\theta L(\theta_{k-1})\| + \epsilon\right) \|\theta_k - \theta_{k-1}\|. \tag{3.44}$$

Since $\theta_k \in B(\theta^*, \delta)$, then it leads to

$$\|\theta_{k+1} - \theta_k\| < h\|\theta_k - \theta_{k-1}\| < \|\theta_k - \theta_{k-1}\| \tag{3.45}$$

so the convergence is guaranteed. Therefore, we obtain that $\|\theta_j - \theta_{j-1}\| \leq h^j\|\theta_1 - \theta_0\|$ for $1 \leq j \leq k+1$, and

$$\|\theta_{k+1} - \theta^*\| \leq \|\theta_0 - \theta^*\| + \sum_{j=0}^{k} \|\theta_{k-j+1} - \theta_{k-j}\| \tag{3.46}$$

$$\leq \|\theta_0 - \theta^*\| + \sum_{j=0}^{k} h^j\|\theta_0 - \theta^*\| \tag{3.47}$$

$$< \frac{1}{1-h}\|\theta_1 - \theta_0\| + \|\theta_0 - \theta^*\| \tag{3.48}$$

$$< \frac{1}{1-h}\frac{h-1}{2}\delta + \frac{1}{2}\delta = \delta, \tag{3.49}$$

which completes the induction. Thus we conclude that the sequence $\{\theta_k\}_{k=0}^{\infty} \subset B(\theta^*, \delta) \subset \Omega_1$ as long as the initial iterate $\theta_0 \in B(\theta^*, \tau) = \Omega_2$.

Secondly, let us define $\hat{\theta} = \lim_{k\to\infty} \theta_k$ so that $\hat{\theta} \in \Omega_1$. By the smoothness of $\nabla_\theta L(\theta)$ and the convergence property (3.45), there exists a constant $\mu > 0$ such that

$$\|\nabla_\theta L(\theta_{k-1})\| = \|\nabla_\theta L(\theta_{k-1}) - \nabla_\theta L(\hat{\theta})\| \tag{3.50}$$

$$\leq \mu\|\theta_{k-1} - \hat{\theta}\| \tag{3.51}$$

$$\leq \mu\left(\|\theta_{k-1} - \theta_k\| + \|\theta_k - \theta_{k+1}\| + \cdots\right) \tag{3.52}$$

$$\leq \frac{\mu}{1-h}\|\theta_k - \theta_{k-1}\|. \tag{3.53}$$

Now let us combine (3.44) and (3.53) to find that

$$\|\theta_{k+1} - \theta_k\| \leq \beta\|\theta_k - \theta_{k-1}\|^2 + \epsilon\|\theta_k - \theta_{k-1}\| \tag{3.54}$$

$$= \left(\beta\|\theta_k - \theta_{k-1}\| + \epsilon\right)\|\theta_k - \theta_{k-1}\| \tag{3.55}$$

for some constant $\beta > 0$. As k grows sufficiently large, the convergence implies that

$$\left(\beta\|\theta_k - \theta_{k-1}\| + \epsilon\right) \leq \gamma \leq 2\epsilon < 1. \tag{3.56}$$

Therefore

$$\|\theta_{k+1} - \theta_k\| \leq \gamma\|\theta_k - \theta_{k-1}\| \tag{3.57}$$

$$\leq \gamma^k\|\theta^1 - \theta^0\| \tag{3.58}$$

$$\leq \gamma^k|\Omega_*|. \tag{3.59}$$

On the other hand, it can be easily seen that

$$\|\theta_{k+1} - \hat{\theta}\| = \|\theta_{k+1} - \theta^{k+2}\| + \|\theta^{k+2} - \theta^{k+3}\| + \|\theta^{k+3} - \theta^{k+4}\| + \cdots \tag{3.60}$$

$$\leq \left(h + h^2 + h^3 + \cdots\right)\|\theta_{k+1} - \theta_k\| \tag{3.61}$$

$$\leq \frac{h}{1-h}\|\theta_{k+1} - \theta_k\|. \tag{3.62}$$

Hence

$$\|\theta_{k+1} - \hat{\theta}\| \leq \frac{h}{1-h}|\Omega_*|\gamma^k. \tag{3.63}$$

The estimation above shows that $\{\|\theta_{k+1}-\hat{\theta}\|\}_{k=1}^{\infty}$ is a decreasing and is at least a linear convergence sequence with coefficient γ. Since γ can be very small due to the construction of this iteration, the parameter sequence $\{\theta_k\}_{k=1}^{\infty}$ will be observed to converge rapidly in the numerical experiments.

□

4 Gauss-Newton Iteration with Monte-Carlo method

References

[1] The Deep Ritz method: A deep learning-based numerical algorithm for solving variational problems. Weinan E and Bing Yu.

[2] RELU DEEP NEURAL NETWORKS AND LINEAR FINITE ELEMENTS*

[3] PERTURBATION THEORY FOR PSEUDO-INVERSES

[4] Nieves Castro-González a, Froilán M. Dopico b, Juan M. Molera b. Multiplicative perturbation theory of the Moore-Penrose inverse and the least squares problem.