Accuracy analysis of uncertain variational problems with analytical and machine learning methods

Vilho Halonen

Mathematics master's thesis

Jyväskylän yliopisto
Department of Mathematics and Statistics
Fall 2021

Abstract Vilho Halonen, *Accuracy analysis of uncertain variational problems with analytical and machine learning methods*, mathematics master's thesis, 55 p., Jyväskylän yliopisto, Department of Mathematics and Statistics, Fall 2021.

In this thesis we compare the performance of analytical methods and neural networks trained with numerically produced data in controlling uncertainty errors of a linear variational problem. We find that neural networks perform well and are feasible to use in practical computations in place of analytical control methods.

Analytical methods for controlling uncertainty errors have been derived for various differential problems (see [1], [2]) in recent decades. The first chapters are devoted to deriving by known methods analytical error bounds for the linear variational problem which we will study. These error bounds are numerically tested and we find that the bounds while they are guaranteed and cheap to compute are not always as sharp as an engineer might hope.

The second part of this thesis consists of creating machine learning models with the goal of approximating the exact error caused by uncertainty in our mathematical model. The chosen type of machine learning model is a deep neural network. The training data used for training the models is generated by numerical computations.

In the final chapter we compare the performance of the analytical methods and machine learning models and we conclude that neural networks can be competitive in this task. If such models are made and found to work for more complicated nonlinear PDE's this method could prove very useful in computer simulations and engineering.

Tiivistelmä Vilho Halonen, *Accuracy analysis of uncertain variational problems with analytical and machine learning methods*, Matematiikan Gradu, 55 s., Jyväskylän yliopisto, Matematiikan ja Tilastotieteen laitos, Syksy 2021.

Tässä tutkielmassa verrataan analyyttisien menetelmien ja koneoppimismallien toimivuutta epätarkkuudesta johtuvien virheiden kontrolloinnissa. Tarkasteltavana matemaattisena esimerkkiongelmana käytetään lineaarista variaatio-ongelmaa. Tuloksena havaitaan, että neuroverkot toimivat hyvin ja ovat käytäntöön mahdollisesti soveltuva keino tehdä virhearviointia.

Monille osittaisdifferentiaaliyhtälöille on johdettu analyyttisia kontrollointikeinoja viime vuosikymmenien aikana (katso [1], [2]). Ensimmäiset luvut käytämme analyyttisien virhearvioiden todistamiseen tunnettujen analyysin työkalujen avulla tarkasteltavalle variaatio-ongelmalle. Virhearvioita testataan numeerisesti ja huomataan, että vaikka analyyttiset rajat ovat varmoja ja halpoja laskennallisesti, ne ovat monesti toivottua epätarkempia.

Tutkielman toisessa osiossa luodaan koneoppimismalleja, joilla pyritään arvioimaan tarkalleen epätarkkuuden aiheuttamaa virhettä. Valittu koneoppimismalli on neuroverkko. Mallien kouluttamiseen käytetty data luodaan itse numeerisilla menetelmillä.

Viimeisessä luvussa verrataan analyyttisien metodien ja luotujen neuroverkkojen toimivuutta. Vertailussa käytetään koulutusdatasta eroavaa generoitua dataa jolle lasketaan analyyttiset rajat, numeeriset approksimaatiot ja neuroverkkojen tulokset. Havaitaan, että neuroverkot suoriutuvat tehtävästä niin hyvin, että voidaan sanoa niiden olevan kilpailullisia analyyttisien metodien kanssa. Jos vastaavia koneoppimismalleja pystytään luomaan vaikeammille moniulotteisille ongelmille, tämä menetelmä voi osoittautua varsin hyödylliseksi simuloinnissa ja insinöörityössä.

Contents

Introduction 1

Chapter 1. Uncertainty Errors in Mathematical Modeling 3
1.1. Sources of Errors in Mathematical Modeling 3
1.2. Notation and Definitions 4
1.3. Model Problem Definition 6

Chapter 2. Quantifying Uncertainty Errors by Analytical Methods 9
2.1. Preliminary Tools 9
2.2. Two-Sided Bounds of the Solution Set Diameter 13

Chapter 3. Sensitivity Analysis and Numerical Approximation of Diam($\mathcal{S}(\mathcal{D})$) 23
3.1. Brute-force Method for Approximating the Diameter 24
3.2. Numerical Tests for Sensitivity 26

Chapter 4. Machine Learning Model for Quantifying Size of the Solution Set 35
4.1. Model Description 35
4.2. Constraining and Generating Training Data 36
4.3. Example Models 37
4.4. Comparing Analytical, Brute-force and ML-Model Methods 42

Conclusions 49

Appendix A. Mathematical Background 51
1.1. Function Spaces 51
1.2. Calculus of variations 52
1.3. Inequalities 54

Bibliography 57

Introduction

Data uncertainty is an ever present problem in mathematical modeling, simulation and computation (see e.g. [8]). In this thesis we present three different methods for controlling errors generated by uncertainty in the case of a linear variational problem. First in Chapter 2 we derive error bounds using mathematical analysis. In Chapter 3 we present a numerical method which uses pure computation to approximate the error quantity of interest. Finally in Chapter 4 we create a machine learning model to do the same. In section 4.4 we test all three methods against one another and see how well each one performs.

Consider first an abstract uncertain mathematical problem of the form: Find $u \in V$ such that

$$\mathcal{A}u = f.$$

The operator $\mathcal{A}$ and source term f are objects which are related to some physical phenomenon. For example, in linear elasticity the operator $\mathcal{A}$ depends on the Lamè parameters of the material and the source term f describes the forces that deform our material. In a heat equation the operator $\mathcal{A}$ depends on the thermal diffusivity of the medium and the source term f describes heatflow from outside the system.

In reality we never know the exact value of $\mathcal{A}$ and f. Physical paremeters are always given in some set of indeterminacy. We denote *the set of admissible data* by $\mathcal{D}$. In this case we only know that

$$(\mathcal{A}, f) \in \mathcal{D}.$$

We wish to quantify the distance between solutions generated by different elements of $\mathcal{D}$. Assume $(\mathcal{A}_1, f_1), (\mathcal{A}_2, f_2) \in \mathcal{D}$ and $(\mathcal{A}_1, f_1) \neq (\mathcal{A}_2, f_2)$. Solving the problems

$$\begin{aligned} \mathcal{A}_1 u &= f_1 \quad \text{and} \\ \mathcal{A}_2 u &= f_2, \end{aligned}$$

the solutions u_1 and u_2 may not be the same. The quantity of interest is the distance

$$d = |u_1 - u_2|,$$

where $|\cdot|$ is a suitable norm in the problem setting. It is even more interesting if we can quantify the *maximum distance between any two solutions*. For this we define the diameter quantity

$$\text{Diam} = \sup_{(\mathcal{A}_1, f_1), (\mathcal{A}_2, f_2) \in \mathcal{D}} |u_1 - u_2|.$$

We will derive methods to control the diameter by analytical, numerical and machine learning tools. The analytical control scheme is presented in chapter 2, the numerical brute-force method in chapter 3 and the machine learning model in chapter 4.

In the past, this problem has been dealt with by many different methods such as the probabilistic method [8] and the worst case scenario method [9]. More recently in the works [1] and [2], analytical tools for controlling uncertainty errors in a guaranteed way have been derived for various PDE's and these are the methods we will present in Chapter 2.

Many important effects related to uncertainty can be seen even in simple problem settings. We analyse them using the problem: Find $u \in V$ such that

$$J(u) = \inf_{v \in V} J(v), \quad \text{where} \quad J(v) := \int_a^b \left(\frac{1}{2}\alpha |v'|^2 + \frac{1}{2}\beta |v|^2 + fv \right) \, dx.$$

In this case, the indeterminacy is present in the coefficients α, β and f. They are not exactly known but instead belong to some admissible set $\mathcal{D}$. For this simple problem, analytical methods are well known and sufficiently sharp for practical computations and numerical methods are sufficiently inexpensive to use. Because of this it is not apparent why alternative approaches are required. Our goal is to give a proof of concept which can be extended to linear and nonlinear PDE's. For very difficult PDE's analytical methods do not exist and numerical methods require a lot of computation. For such problems, an inexpensive method to check our accuracy would be very useful in for example computer simulations.

In chapter 3, we present various tests to show that it is not obvious how changes in the indeterminacy of the data affect the numerical approximations and the analytical bounds. In the tests of section 4.4, we find that in some cases our neural network could be preferred over analytical methods even in our constrained problem setting (see Figure 4.9).

CHAPTER 1

Uncertainty Errors in Mathematical Modeling

In this chapter, we introduce the model problem and the important quantities we wish to control when our problem involves uncertainty. Reliably solving an uncertain mathematical model requires knowledge about the distance between all possible solutions of the problem. The maximum distance between two different solutions is the most useful information. To quantify this distance we define the *diameter of the solution set*. In chapters 2, 3 and 4, we create and compare different methods for controlling the diameter quantity.

We start this chapter by describing the main sources of errors in mathematical modeling (see [10, chapter 1]). The theory of error control has been widely studied with various approaches in for example [11]. In section 1.2, we present the model problem and give the required definitions for analysing errors generated by uncertain data.

1.1. Sources of Errors in Mathematical Modeling

In any mathematical model of a real-world problem a number of different sources of errors are present (see Figure 1.1).The first error comes from the fact that reality and the mathematical model are not exactly the same. Our models are always looking at an idealized version of what is truly occurring. This error is called the *modeling error* and we denote it E_1. The modeling error describes the distance between the solution U of the real world problem and the solution u of the mathematical model.

$$E_1 = |U - u|.$$

The modeling error arises from uncertainty in measurements and the simplifying assumptions of the model. For example, models of elasticity often make assumptions about elasticity constants of a material being homogenous. Such models may produce good results but in reality any material constant has variance. The modeling error also accounts for uncertainty in measurements. In practice there is always some level of uncertainty in data. For a model to be considered reliable, effects of uncertainty on possible solutions must be understood. The uncertainty error and methods of controlling it are our main interest.

Continual models describing a real world problem are made such that in theory an exact solution exists. Unfortunately, exact solutions are rarely at our disposal. Models describing complicated phenomena often lead to models like nonlinear PDE's for which analytical solutions are unknown. In order to find quantitative answers the model must be reduced to a discrete model which is solvable by numerical methods. The loss of information caused by the discretization is called the *approximation error*.

We denote the approximation error E_2. The approximation error describes the distance between the solution u of the continual model and the solution v of the discrete model.

$$E_2 = |u - v|.$$

After a discrete model has been constructed, it remains to solve it correctly. Numerical methods are subject to numerical errors so we still have to account for the *numerical error* E_3 which consists of roundoff errors, stopping conditions which dont allow arbitrary amounts of computation and bugs in the code of our solver. The numerical error describes the difference between the solution v of the discrete model and the quantitative result $\hat{v}$ produced by some numerical method.

$$E_3 = |v - \hat{v}|.$$

In quantitative research, all of the errors described above are important to have controls for. Our goal in this thesis is to present and compare different methods for controlling the part of the modeling error E_1 which comes from data uncertainty.

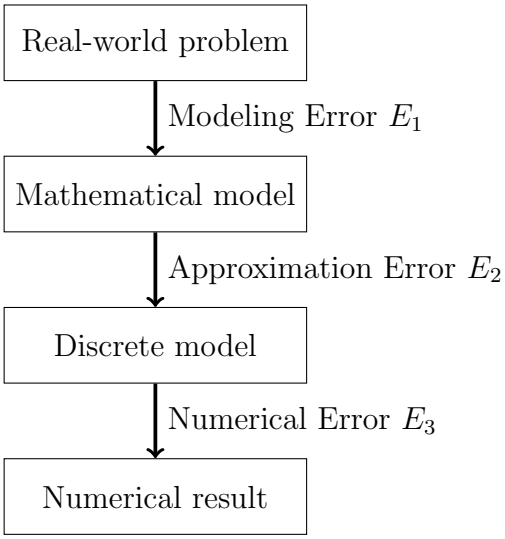


FIGURE 1.1. Error types

1.2. Notation and Definitions

Uncertainty errors in mathematical modeling appear because measuring accuracy in any real-world phenomenon is not exact. In this section we define the set of admissible data, solution sets and the diameter of the solution set. The quantity of interest in later chapters is the diameter. In this section, the definitions are given in abstract form and in the following section we specify them for the model problem.

DEFINITION 1.1. Consider an abstract differential problem of the form: Find $u \in V$ such that

$$\mathcal{A}u = f.$$

The operator $\mathcal{A}$ and source term f may be incompletely known and instead it is known that

$$(\mathcal{A}, f) \in \mathcal{D}.$$

The set $\mathcal{D}$ is called the *set of admissible data*.

Next we define the set of solutions which contains all the different solutions generated by $\mathcal{D}$. In later chapters we derive analytical, numerical and machine learning methods that control the "size" of this set.

DEFINITION 1.2. Let $\mathcal{D}$ be the admissible dataset of a differential problem of the type from Definition 1.1. We define the *solution mapping* $\mathcal{S} : \mathcal{D} \to V$ as the mapping that takes an input $(\mathcal{A}_x, f_x) \in \mathcal{D}$ and outputs the exact solution $u_x \in V$ of the related problem $\mathcal{A}_x u = f_x$. The image $\mathcal{S}(\mathcal{D})$ is called the *Solution Set* of the uncertain problem.

It remains to define a quantity which is suitable to describe the "size" of the solution set. In our case, we define the diameter and radius of the set with respect to some suitable norm.

DEFINITION 1.3. The diameter of the solution set $\mathcal{S}(\mathcal{D})$ is

$$\text{Diam}(\mathcal{S}(\mathcal{D})) := \sup_{u_1, u_2 \in \mathcal{S}(\mathcal{D})} |||u_1 - u_2|||,$$

where $|||\cdot|||$ is a suitable norm in the Banach space V.

The radius is defined with respect to a "mean" element of the admissible data. The admissible dataset may be not symmetric so "mean" in this case is simply an element of our choice.

DEFINITION 1.4. The radius of the solution set $\mathcal{S}(\mathcal{D})$ is

$$r := \sup_{u \in \mathcal{S}(\mathcal{D})} |||u_\circ - u|||,$$

where $u_\circ = \mathcal{S}(\alpha_\circ, \beta_\circ, f_\circ)$ is the "mean" solution and $|||\cdot|||$ is a suitable norm for the problem.

Note that $r \leq \text{Diam}(\mathcal{S}(\mathcal{D})) \leq 2r$. The quantity $\text{Diam}(\mathcal{S}(\mathcal{D}))$ is useful in practical computations for two reasons:

- It gives us an idea of whether or not our measurements are sufficiently accurate.
- When we approximate an exact solution u by some v and have a way of estimating the approximation error $|||u - v|||$ it gives us an accuracy limit on this error (see Figure 1.2).

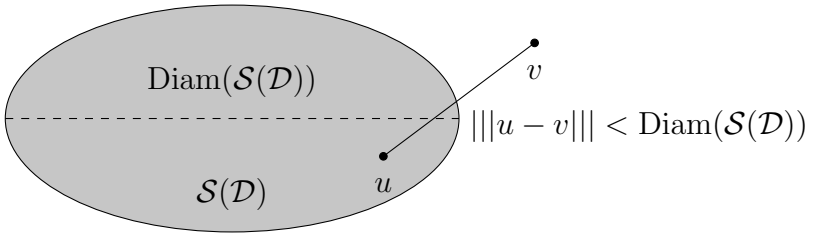


FIGURE 1.2. If v is the approximation of u, in this case v could already be inside the solution set! There is no point in sharpening the approximation scheme.

1.3. Model Problem Definition

The model problem is a linear variational minimization problem. It is referred to in all chapters as problem $\mathcal{P}$ or $\mathcal{P}(\alpha, \beta, f)$ if there is possibility of confusion with related problem data.

Problem $\mathcal{P}(\alpha, \beta, f)$: Find $u \in H^1(a, b)$ (see definition A.2) such that

$$J(u) = \inf_{v \in H^1} J(v), \quad \text{where} \quad J(v) := \int_a^b \frac{1}{2}\alpha|v'|^2 + \frac{1}{2}\beta|v|^2 + fv \, dx, \tag{1.1}$$

$$0 < \alpha(x) \leq \alpha_\oplus, \quad 0 < \beta(x) \leq \beta_\oplus, \quad f \in L^2(a, b), \tag{1.2}$$

$$u(a) = A, \quad \text{and} \quad u(b) = B. \tag{1.3}$$

The functional J is convex so problem $\mathcal{P}$ is equivalent to the boundary value problem

$$\begin{aligned} &(\alpha u')' - \beta u = f, \\ &u(a) = A, \quad u(b) = B. \end{aligned} \tag{1.4}$$

Problem $\mathcal{P}$ can be understood as an uncertain problem when the coefficient functions α, β and f are incompletely known.

DEFINITION 1.5. The admissible dataset of the uncertain problem $\mathcal{P}$ is

$$\mathcal{D} := \mathcal{D}_\alpha \times \mathcal{D}_\beta \times \mathcal{D}_f,$$

where

$$\begin{aligned} \mathcal{D}_\alpha &:= \{\alpha = \alpha_\circ + \delta_1 g \ : \ ||g||_\infty \leq 1\}, \\ \mathcal{D}_\beta &:= \{\beta = \beta_\circ + \delta_2 g \ : \ ||g||_\infty \leq 1\}, \\ \mathcal{D}_f &:= \{f = f_\circ + \delta_3 g \ : \ ||g||_\infty \leq 1\}. \end{aligned}$$

The functions $\alpha_\circ, \beta_\circ$ and $f_\circ$ are the "mean" elements of the sets and the functions $\delta_i(x) \geq 0$ are the maximum perturbations from the central elements.

Defining the admissible data in this way is convenient for our later analysis and it is a common way to describe uncertainty in practice.

Solutions of problem $\mathcal{P}$ are analysed with the so called *energy norm*

$$||v||_{\alpha,\beta} = \left(\int_a^b \alpha |v'|^2 + \beta |v|^2 \, dx \right)^{\frac{1}{2}}. \tag{1.5}$$

The natural choice to compare distances between different solutions of the uncertain problem $\mathcal{P}$ is the *mean energy norm*

$$||w||_\circ := ||w||_{\alpha_\circ,\beta_\circ} = \left(\int_a^b \alpha_\circ |w'|^2 + \beta_\circ |w|^2 \, dx \right)^{\frac{1}{2}}, \tag{1.6}$$

where $\alpha_\circ$, $\beta_\circ$ and f are from definition 1.5. The diameter and radius of the solution set of problem $\mathcal{P}$ are defined using the mean energy norm with definitions 1.3 and 1.4.

CHAPTER 2

Quantifying Uncertainty Errors by Analytical Methods

In this chapter we derive a lower bound and four different upper bounds for the quantity $\text{Diam}(\mathcal{S}(\mathcal{D}))$ of problem $\mathcal{P}$. In section 2.1 some preliminary lemmas are introduced and in section 2.2 the bounds are derived.

2.1. Preliminary Tools

The most important tool needed for deriving the bounds are the *functional a posteriori error minorant and majorants*. These bounds control the approximation and numerical error of a problem in a guaranteed way (see Figure 1.1). This theory is discussed at length in [1] and [2] where bounds for various differential problems are derived.

Assume that u is the exact solution of $\mathcal{P}(\alpha, \beta, f)$ and v is an approximation of u computed by any method. A posteriori bounds are quantities $\underline{M}$ and $\overline{M}$ which do not depend on the exact solution u and control the errors E_2 and E_3 by

$$\underline{M}_{(\alpha,\beta,f)}(v, y) \leq ||u - v||_{\alpha,\beta} \leq \overline{M}_{(\alpha,\beta,f)}(v, y). \tag{2.1}$$

These quantities depend only on the approximation v, the problem data (α, β, f) and a free function $y \in H_0^1(a, b)$. The tricky part is choosing the free function y such that the bounds are sharp. If we know the exact solution u, we can pick y such that 2.1 is an equality for the minorant and one of the majorants. In practice, the bounds can be optimized with respect to y by any numerical method.

The error minorant $\underline{M}$ follows directly from the definition of problem $\mathcal{P}$ with some algebraic manipulations.

LEMMA 2.1. *Let $u \in V$ be an exact solution of $\mathcal{P}$ and $v \in H^1(a, b)$ be an approximation of u. Then*

$$||u - v||_{\alpha,\beta}^2 \geq \underline{M}_{\alpha,\beta,f}(v, w) := -||w||_{\alpha,\beta}^2 - 2\int_a^b \alpha v'w' + \beta vw + fw \, dx,$$

where $w \in H_0^1(a, b)$.

PROOF. Notice that $\frac{1}{2}||u - v||_{\alpha,\beta}^2 = J(v) - J(u)$. This is shown as follows:

$$J(v) - J(u) = \int_a^b \frac{1}{2}\alpha|v'|^2 + \frac{1}{2}\beta|v|^2 + fv - \frac{1}{2}\alpha|u'|^2 - \frac{1}{2}\beta|u|^2 - fu \, dx. \tag{2.2}$$

Since u is a solution to $\mathcal{P}$, for any $g \in H_0^1(a, b)$ we have

$$\int_a^b \alpha u'g' + \beta ug \, dx = \int_a^b -fg \, dx.$$

Substituting fv and fu in (2.2) we have

$$\begin{aligned} J(v) - J(u) &= \int_a^b \frac{1}{2}\alpha|v'|^2 + \frac{1}{2}\beta|v|^2 - \alpha u'v' - \beta uv + \frac{1}{2}\alpha|u'|^2 + \frac{1}{2}\beta|u|^2 \, dx \\ &= \frac{1}{2}\int \alpha|u' - v'|^2 + \beta|u - v|^2 \, dx \\ &= \frac{1}{2}||u - v||_{\alpha,\beta}. \end{aligned}$$

Since u is a minimizer of J, for any $v + w \in H_0^1(a, b)$ we find

$$\begin{aligned} \frac{1}{2}||u - v||_{\alpha,\beta}^2 &= J(v) - J(u) \geq J(v) - J(v + w) \\ &= \int_a^b \frac{1}{2}\alpha|v'|^2 + \frac{1}{2}\beta|v|^2 - \frac{1}{2}\alpha|v' + w'|^2 - \frac{1}{2}\beta|v + w|^2 - fw \, dx \\ &= \int_a^b -\frac{1}{2}\alpha|w'|^2 - \frac{1}{2}\beta|w|^2 - \alpha v'w' - \beta vw - fw \, dx \\ &= -\frac{1}{2}||w||_{\alpha,\beta}^2 - \int_a^b \alpha v'w' + \beta vw + fw \, dx. \end{aligned}$$

Multiplying both sides by 2 completes the proof. □

The minorant $\underline{M}_{(\alpha,\beta,f)}(v, w)$ is sharp. If we know the exact solution we can pick $w = u - v$ and have

$$||u - v||_{\alpha,\beta}^2 = \underline{M}_{\alpha,\beta,f}(v, u - v).$$

The majorants are derived similarily as in [1, Section 3.1]. There are two variations of the upper bound and depending on the values of the coefficient β in $\mathcal{P}$ one of them is better. In practical computations both should be computed and the better one used.

LEMMA 2.2. *Let $u \in V$ be an exact solution of $\mathcal{P}$ and $v \in V$ be an approximation of u computed by some method. Then*

$$||u - v||_{\alpha,\beta} \leq \overline{M}_{1(\alpha,\beta,f)}(v, y) := \left(\int_a^b \frac{1}{\alpha}(y - \alpha v')^2 dx\right)^{1/2} + \overline{C}_F||y' - \beta v - f||,$$

where $\overline{C}_F = \frac{b-a}{\pi}(\operatorname{ess\,sup} \alpha(x)^{-1})^{\frac{1}{2}}$ and $y \in H^1(a, b)$.

PROOF. A minimizer of (1.1) satisfies the weak Euler-Lagrange equation (see A.5)

$$\int_a^b \alpha u'\varphi' + \beta u\varphi \, dx = \int_a^b (-f\varphi) \, dx \quad \forall \varphi \in \mathrm{H}_0^1(a, b). \tag{2.3}$$

Adding $-\alpha v'\varphi' - \beta v\varphi$ to both sides of (2.3) gives

$$\int_a^b \alpha(u' - v')\varphi' + \beta(u - v)\varphi \, dx = \int_a^b (-f\varphi - \beta\varphi v - \alpha v'\varphi') \, dx. \tag{2.4}$$

Since

$$\int_a^b (y\varphi)' dx = \Big|_a^b (y\varphi) = 0 \quad \text{for any} \;\; y \in \mathrm{H}^1(a, b),$$

we have

$$\int_a^b \alpha(u'-v')\varphi' + \beta(u-v)\varphi \, dx = \int_a^b \varphi(y'-f-\beta v) + \varphi'(y-\alpha v') \, dx. \tag{2.5}$$

Both u and v satisfy the boundary conditions (1.3), so $u - v \in \mathrm{H}_0^1$. Substituting $\varphi = u - v$ in (2.5) we have

$$||u-v||_{\alpha,\beta}^2 = \int_a^b (u-v)(y'-f-\beta v) + (u-v)'(y-\alpha v') \, dx. \tag{2.6}$$

We multiply the second term the right hand side of (2.6) by $\sqrt{\frac{\alpha}{\alpha}}$ (α is strictly positive). Afterwards we use Hölder's inequality and have

(2.7)

$$\begin{aligned}\int_a^b (u-v)'(y-\alpha v') &\leq \left(\int_a^b \alpha(u'-v')^2 dx\right)^{\frac{1}{2}} \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 dx\right)^{\frac{1}{2}} \\ &\leq \left(\int_a^b \alpha(u'-v')^2 + \beta(u-v)^2 dx\right)^{\frac{1}{2}} \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 dx\right)^{\frac{1}{2}} \\ &= ||u-v||_{\alpha,\beta} \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 dx\right)^{\frac{1}{2}}.\end{aligned}$$

Using Hölder's inequality and Theorem A.8 for the first term on the right hand side of (2.6) we have

$$\begin{aligned}\int_a^b (u-v)(y'-f-\beta v) &\leq ||u-v||\ ||y'-f-\beta v|| \\ &\leq \overline{C}_F ||u-v||_{\alpha,\beta} ||y'-f-\beta v||.\end{aligned} \tag{2.8}$$

Using the estimates (2.7) and (2.8) for (2.6) we obtain

$$||u-v||_{\alpha,\beta}^2 \leq ||u-v||_{\alpha,\beta} \left(\left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 dx\right)^{\frac{1}{2}} + \overline{C}_F ||y'-f-\beta v|| \right).$$

Dividing by $||u-v||_{\alpha,\beta}$ completes the proof. □

The second majorant is derived similarly but we do not use Theorem A.8.

LEMMA 2.3. *Let $u \in V$ be an exact solution of $\mathcal{P}$ and $v \in V$ be an approximation of u computed by some method. Then*

$$||u-v||_{\alpha,\beta} \leq \overline{M}_{2(\alpha,\beta,f)}(v,y) := \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 + \frac{1}{\beta}(y'-f-\beta v)^2 \, dx\right)^{\frac{1}{2}},$$

where $y \in H^1(a,b)$.

PROOF. We follow the proof of Lemma 2.2 closely. We only modify the estimate in (2.8). Looking at the first term on the right hand side of (2.6), we estimate using

Hölder's inequality and multiplying by $\frac{\sqrt{\beta}}{\sqrt{\beta}}$

$$\int_a^b (u-v)(y'-f-\beta v) \leq \left(\int_a^b \beta(u-v)^2\,dx\right)^{\frac{1}{2}} \left(\int_a^b \frac{1}{\beta}(y'-f-\beta v)^2\,dx\right)^{\frac{1}{2}} \tag{2.9}$$

$$\leq ||u-v||_{\alpha,\beta} \left(\int_a^b \frac{1}{\beta}(y'-f-\beta v)^2\,dx\right)^{\frac{1}{2}} \tag{2.10}$$

Now using (2.7), (2.10) and the algebraic inequality $t_1\lambda_1+t_2\lambda_2 \leq \sqrt{\lambda_1^2+\lambda_2^2}\sqrt{t_1^2+t_2^2}$, we estimate (2.6) and find

$$||u-v||_{\alpha,\beta}^2 = \int_a^b (u-v)(y'-f-\beta v) + (u-v)'(y-\alpha v')\,dx$$

$$\leq \left(\int_a^b \alpha(u'-v')^2 dx\right)^{\frac{1}{2}} \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 dx\right)^{\frac{1}{2}}$$
$$+ \left(\int_a^b \beta(u-v)^2\,dx\right)^{\frac{1}{2}} \left(\int_a^b \frac{1}{\beta}(y'-f-\beta v)^2\,dx\right)^{\frac{1}{2}}$$
$$\leq ||u-v||_{\alpha,\beta} \left(\int_a^b \frac{1}{\alpha}(y-\alpha v')^2 + \frac{1}{\beta}(y'-f-\beta v)^2\,dx\right)^{\frac{1}{2}}.$$

Dividing by $||u-v||_{\alpha,\beta}$ completes the proof. □

If the coefficient β has very small values the majorant $\overline{M}_1$ will often outperform $\overline{M}_2$. If we know the exact solution u we can pick $y=\alpha u'$ and find that

$$||u-v||_{\alpha,\beta} = \overline{M}_{2(\alpha,\beta,f)}(v,\alpha u').$$

The majorant $\overline{M}_1$ is not sharp in the same sense since

$$\overline{M}_{1(\alpha,\beta,f)}(v,\alpha u') = \left(\int_a^b \alpha|u'-v'|^2\,dx\right)^{\frac{1}{2}} + \overline{C}_F||\beta(u-v)||.$$

For $\overline{M}_1$ to be equal to the norm, the approximation has to agree with the exact solution $v=u$.

Next we derive the *energy estimate* which we will use for two of the bounds. This only works in the special case where the boundary conditions are zeros.

LEMMA 2.4. *Let $u=(\mathcal{S}(\alpha,\beta,f))$ and let the boundary conditions of $\mathcal{P}$ be zeros. Then*

$$||u||_{\alpha,\beta} \leq \overline{C}_\oplus \left|\left||f_\circ| + \delta_3\right|\right|, \quad \textit{where} \quad \overline{C}_\oplus := \sup_{\alpha\in\mathcal{D}_\alpha} \left\{\frac{b-a}{\pi}(\operatorname{ess\,sup}\{\alpha(x)^{-1}\})^{\frac{1}{2}}\right\}.$$

This is referred to as the energy estimate.

PROOF. The function u satisfies the weak Euler-Lagrange equation

$$\int_a^b \varphi' u'\alpha + \varphi\beta u\,dx = \int_a^b (-f\varphi)\,dx \quad \forall\varphi\in \mathrm{H}_0^1(a,b).$$

Since the boundary conditions are zeros, we can substitute $\varphi = u$ and obtain

$$\int_a^b u'u'\alpha + u\beta u \, dx = ||u||^2_{\alpha,\beta} = \int_a^b (-fu) \, dx.$$

Using Hölder inequality and Theorem A.8 we have

$$||u||^2_{\alpha,\beta} \leq ||f||_2||u||_2 \leq ||f||_2\overline{C}_F||u||_{\alpha,\beta} \leq \overline{C}_F\Big|\Big||f_\circ| + \delta_3\Big|\Big||u||_{\alpha,\beta}.$$

Dividing by $||u||_{\alpha,\beta}$ we have

$$||u||_{\alpha,\beta} \leq \overline{C}_F\Big|\Big||f_\circ| + \delta_3\Big|\Big|$$

Choosing the supremum of C_F over $\alpha \in \mathcal{D}_\alpha$ makes this hold regardless of which energy norm is on the left hand side. □

2.2. Two-Sided Bounds of the Solution Set Diameter

Using the error minorant $\underline{M}$ and the majorants $\overline{M}_1$ and $\overline{M}_2$, two-sided bounds which control errors generated by data uncertainty can be derived. The quantity we wish to control is the diameter of the solution cloud (see Definition 1.3). The bounds control this quantity by

$$\underline{\mathcal{B}} \leq \mathrm{Diam}(\mathcal{S}(\mathcal{D})) \leq \overline{\mathcal{B}}.$$

First we follow the methods in [1, Chapter 5] to derive two upper bounds and a lower bound which can be computed if the "mean" solution $u_\circ = \mathcal{P}(\alpha_\circ, \beta_\circ, f_\circ)$ is at our disposal. Afterwards, we derive two upper bounds using the energy estimate from Lemma 2.4. The bounds that use energy estimates depend only on the problem data but usually the bounds which utilize the "mean" solution are sharper. In chapter 3, we numerically test how each of the bounds behave in a few example cases. It is not easy to say which bound is optimal in what situation but in practice one can compute all of them and pick the best one.

In order for the following analysis to be sensible the mean coefficients and indeterminacy functions must be constrained as follows. From now on we assume that

$$\frac{\delta_1(x)}{\alpha_\circ(x) - \delta_1(x)} < 1 \quad \text{and} \quad \frac{\delta_2}{\beta_\circ(x) - \delta_2(x)} < 1, \quad \forall \quad x \in (a,b). \tag{2.11}$$

If (2.11) does not hold the bounds are not defined as real numbers. In practice, this is a reasonable assumption since relative errors of more then 50% are rarely interesting.

For abuse of notation we denote

$$c_1 := \min_{x\in(a,b)} \frac{\delta_1(x)}{\alpha_\circ(x) - \delta_1(x)} \quad \text{and} \quad c_2 := \min_{x\in(a,b)} \frac{\delta_2(x)}{\beta_\circ(x) - \delta_2(x)}.$$

Before deriving the bounds we need one more technical result.

THEOREM 2.5. *Let $w \in V$ and the uncertain problem $\mathcal{P}$ be such that it satisfies (2.11). Then, for all elements of the data $(\alpha, \beta) \in \mathcal{D}_\alpha \times \mathcal{D}_\beta$,*

$$\underline{K}||w||_{\alpha,\beta} \leq ||w||_\circ \leq \overline{K}||w||_{\alpha,\beta},$$

where

$$\overline{K} := \sqrt{1 + \max\{c_1, c_2\}} \qquad \textit{and} \qquad \underline{K} := \sqrt{1 - \max\{c_1, c_2\}}.$$

PROOF. First the lower bound

$$\begin{aligned}
||w||_\circ^2 &= \int_a^b \alpha_\circ |w'|^2 + \beta_\circ |w|^2 \, dx \\
&\geq \int_a^b (\alpha - \delta_1)|w'|^2 + (\beta - \delta_2)|w|^2 \, dx \\
&\geq \int_a^b (\alpha - \delta_1)|w'|^2 + (\beta - \delta_2)|w|^2 \, dx \\
&= ||w||_{\alpha,\beta} - \int_a^b \delta_1 |w'|^2 + \delta_2 |w|^2 \, dx \\
&\geq ||w||_{\alpha,\beta} - \int_a^b \frac{\delta_1}{\alpha_\circ - \delta_1} \alpha |w'|^2 + \frac{\delta_2}{\beta_\circ - \delta_2} \beta |w|^2 \, dx \\
&\geq ||w||_{\alpha,\beta} - \int_a^b c_1 \alpha |w'|^2 + c_2 \beta |w|^2 \, dx \\
&\geq ||w||_{\alpha,\beta} \left(1 - \max\{c_1, c_2\}\right).
\end{aligned}$$

Taking square root completes the lower bound. Similarily we have

$$\begin{aligned}
||w||_\circ^2 &\leq \int_a^b (\alpha + \delta_1)|w'|^2 + (\beta + \delta_2)|w|^2 \, dx \\
&= ||w||_{\alpha,\beta} + \int_a^b \delta_1 |w'|^2 + \delta_2 |w|^2 \, dx \\
&\leq ||w||_{\alpha,\beta} + \int_a^b \frac{\delta_1}{\alpha_\circ - \delta_1} \alpha |w'|^2 + \frac{\delta_2}{\beta_\circ - \delta_2} \beta |w|^2 \, dx \\
&\leq ||w||_{\alpha,\beta} + \int_a^b c_1 \alpha |w'|^2 + c_2 |w|^2 \, dx \\
&\leq ||w||_{\alpha,\beta} \left(1 + \max\{c_1, c_2\}\right).
\end{aligned}$$

Taking square root completes the proof. □

Now we have all the tools required to derive bounds for $\mathrm{Diam}(\mathcal{S}(\mathcal{D}))$. We start with the lower bound.

THEOREM 2.6. *The diameter of the solution set $\mathcal{S}(\mathcal{D})$ of the uncertain problem $\mathcal{P}$ with admissible dataset $\mathcal{D}$ has a guaranteed lower bound given by*

$$\textit{Diam}(\mathcal{S}(\mathcal{D})) \geq \underline{K} \frac{\int_a^b \delta_1 |u_\circ'|^2 + \delta_2 |u_\circ|^2 + \delta_3 |u_\circ| \, dx}{\left(||u_\circ||_\circ^2 - \int_a^b \delta_1 |u_\circ'|^2 + \delta_2 |u_\circ|^2 \, dx \right)^{\frac{1}{2}}},$$

where $\underline{K}$ is from Theorem 2.5.

PROOF. Using Theorem 2.5 and applying the error minorant from Lemma 2.1 we have

$$\begin{aligned} r^2 &= \sup_{u\in\mathcal{S}(\mathcal{D})} ||u_\circ - u||_\circ^2 \\ &\geq \underline{K}^2 \sup_{(\alpha,\beta,f)\in\mathcal{D}} \sup_{u\in\mathcal{S}(\mathcal{D})} ||u_\circ - u||_{\alpha,\beta}^2 \\ &\geq \underline{K}^2 \sup_{(\alpha,\beta,f)\in\mathcal{D}} \sup_{w\in H_0^1} \underline{M}_{\alpha,\beta,f}(u_\circ, w) \\ &\geq \underline{K}^2 \sup_{(\alpha,\beta,f)\in\mathcal{D}} \sup_{w\in H_0^1} \left(-||w||_{\alpha,\beta}^2 - 2\int_a^b \alpha u_\circ' w' + \beta u_\circ w + fw \; dx \right). \end{aligned}$$

Substituting α, β and f with equivalent forms from definition 1.5 we have

$$\begin{aligned} \geq \underline{K}^2 \sup_{||g_i||_\infty\leq 1} \sup_{w\in H_0^1} \Bigg(&- \int_a^b (\alpha_\circ + \delta_1 g_1)|w'|^2 + (\beta_\circ + \delta_2 g_2)|w|^2 \; dx \\ &- 2\int_a^b \alpha_\circ u_\circ' w' + \beta_\circ u_\circ w + f_\circ w \; dx \\ &- 2\int_a^b \delta_1 g_1 u_\circ' w' + \delta_2 g_2 u_\circ w + \delta_3 g_3 w \; dx \Bigg). \end{aligned} \tag{2.12}$$

The function $u_\circ$ is the solution of $\mathcal{P}_\circ$ so

$$\int_a^b \alpha_\circ u_\circ' h' + \beta_\circ u_\circ h \; dx = \int_a^b -f_\circ h \; dx \quad \text{for any} \;\; h \in H_0^1$$

Therefore the second term in (2.12) disappears. Estimating the supremum of (2.12) from below by setting $w = tu_\circ$, where $t > 0$ is a constant, we have

$$\begin{aligned} &\geq \underline{K}^2 \sup_{||g_i||_\infty\leq 1} \Bigg(- \int_a^b (\alpha_\circ + \delta_1 g_1)|tu_\circ'|^2 + (\beta_\circ + \delta_2 g_2)|tu_\circ|^2 \; dx \\ &\qquad - 2\int_a^b \delta_1 g_1 u_\circ' t u_\circ' + \delta_2 g_2 u_\circ t u_\circ + \delta_3 g_3 t u_\circ \; dx \Bigg) \\ &= \underline{K}^2 \sup_{||g_i||_\infty\leq 1} \Bigg(- t^2 \left(||u_\circ||_\circ^2 + \int_a^b \delta_1 g_1 |u_\circ'|^2 + \delta_2 g_2 |u_\circ|^2 \; dx \right) \\ &\qquad - 2t\int_a^b \delta_1 g_1 |u_\circ'|^2 + \delta_2 g_2 |u_\circ|^2 + \delta_3 g_3 u_\circ \; dx \Bigg). \end{aligned}$$

It is clear that the supremum with respect to g_i is reached when $g_1 := -1$, $g_2 := -1$ and $g_3(x) = -\mathrm{sgn}(u_\circ(x))$, so

$$= \underline{K}^2 \left(- t^2 \left(||u_\circ||_\circ^2 - \int_a^b \delta_1 |u_\circ'|^2 + \delta_2 |u_\circ|^2 \; dx \right) + 2t \int_a^b \delta_1 |u_\circ'|^2 + \delta_2 |u_\circ|^2 + \delta_3 |u_\circ| \; dx \right). \tag{2.13}$$

The expression (2.13) is a second order polynomial with a negative second order constant with respect to t. Maximum over $t > 0$ is reached at the zero of the derivative.

For abuse of notation we denote

$$T_1 = ||u_\circ||_\circ^2 - \int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2\, dx,$$
$$T_2 = \int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2 + \delta_3|u_\circ|\, dx.$$

The derivative of (2.13) with respect to t is

$$\frac{d}{dt}\underline{K}^2\left(-t^2T_1 + 2tT_2\right) = \underline{K}^2\left(-2tT_1 + 2T_2\right).$$

Zero of the derivative is at the point $t = T_2/T_1$. Substituting $t = T_2/T_1$ in (2.13) we find

$$\underline{K}^2\left(-t^2T_1 + 2tT_2\right) = \underline{K}^2\left(-\left(\frac{T_2}{T_1}\right)^2 T_1 + 2\frac{T_2}{T_1}T_2\right) = \underline{K}^2\frac{T_2^2}{T_1}$$
$$= \underline{K}^2\frac{\left(\int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2 + \delta_3|u_\circ|\, dx\right)^2}{||u_\circ||_\circ^2 - \int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2\, dx}.$$

By taking square roots we arrive at the result

$$\mathrm{Diam}(\mathcal{S}(\mathcal{D})) \geq r \geq \underline{K}\frac{\int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2 + \delta_3|u_\circ|\, dx}{\left(||u_\circ||_\circ^2 - \int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2\, dx\right)^{\frac{1}{2}}}.$$

□

Next we derive upper bounds using first the majorant $\overline{M}_1$ from Lemma 2.2 and then $\overline{M}_2$ from Lemma 2.3.

THEOREM 2.7. *The diameter of the solution set $\mathcal{S}(\mathcal{D})$ of the uncertain problem $\mathcal{P}$ with admissible dataset $\mathcal{D}$ has a guaranteed upper bound given by*

$$\mathit{Diam}(\mathcal{S}(\mathcal{D})) \leq \overline{\mathcal{B}}_1 := 2\overline{K}\left(\left(\int_a^b \frac{(\delta_1 u_\circ')^2}{\alpha_\circ - \delta_1}\, dx\right)^{\frac{1}{2}} + \overline{C}_F||\delta_2|u_\circ| + \delta_3||\right),$$

where $\overline{K}$ is from Theorem 2.5

PROOF. Using Theorem 2.5 and the majorant $\overline{M}_1$ from Lemma 2.2 we have

$$r = \sup_{u\in\mathcal{S}(\mathcal{D})} ||u - u_\circ||_\circ \leq \overline{K} \sup_{u\in\mathcal{S}(\mathcal{D})} ||u - u_\circ||_{\alpha,\beta}$$
$$\leq \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \inf_{y\in H^1} \overline{M}_{1(\alpha,\beta,f)}(u_\circ, y), \quad \text{(Theorem A.6)}$$
$$\leq \overline{K} \inf_{y\in H^1} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \overline{M}_{1(\alpha,\beta,f)}(u_\circ, y),$$

we estimate the infimum from above by setting $y = \alpha_\circ u_\circ'$,

$$\leq \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \overline{M}_{2(\alpha,\beta,f)}(u_\circ, \alpha_\circ u'_\circ)$$
$$= \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \left(\left(\int_a^b \frac{1}{\alpha}(\alpha_\circ u'_\circ - \alpha u'_\circ)^2 dx \right)^{1/2} + \overline{C}_F ||(\alpha_\circ u'_\circ)' - \beta u_\circ - f|| \right).$$

Rewriting α, β and f with equivalent forms from Definition 1.5

$$= \overline{K} \sup_{||g_i||_\infty \leq 1} \left(\left(\int_a^b \frac{(\alpha_\circ u'_\circ - (\alpha_\circ + \delta_1 g_1)u'_\circ)^2}{\alpha_\circ + \delta_1 g_1} dx \right)^{1/2} \right.$$
$$\left. + \ \overline{C}_F ||(\alpha_\circ u'_\circ)' - (\beta_\circ + \delta_2 g_2)u_\circ - (f_\circ + \delta_3 g_3)|| \right).$$

$u_\circ$ satisfies (1.4) $(\alpha_\circ u'_\circ)' - \beta_\circ u_\circ = f_\circ$, so

$$= \overline{K} \sup_{||g_i||_\infty \leq 1} \left(\left(\int_a^b \frac{(-\delta_1 g_1 u'_\circ)^2}{\alpha_\circ + \delta_1 g_1} dx \right)^{1/2} + \ \overline{C}_F || - \delta_2 g_2 u_\circ - \delta_3 g_3 || \right).$$

Maximizing with respect to g_i leads to $g_1 = -1$, $g_2 = -\mathrm{sgn}(u_\circ(x))$ and $g_3 = -1$. Therefore

$$= \overline{K} \left(\left(\int_a^b \frac{(\delta_1 u'_\circ)^2}{\alpha_\circ - \delta_1} dx \right)^{1/2} + \ \overline{C}_F ||\delta_2 |u_\circ| + \delta_3|| \right),$$

The result follows from the fact that $\mathrm{Diam}(\mathcal{S}(\mathcal{D})) \leq 2r$.

□

Next we make another upper bound using instead the majorant $\overline{M}_1$ from Theorem 2.3.

THEOREM 2.8. *The diameter of the solution set $\mathcal{S}(\mathcal{D})$ of the uncertain problem $\mathcal{P}$ with admissible dataset $\mathcal{D}$ has a guaranteed upper bound given by*

$$\mathit{Diam}(\mathcal{S}(\mathcal{D})) \leq \overline{\mathcal{B}}_2 := 2\overline{K} \left(\int_a^b \frac{(\delta_1 u'_\circ)^2}{\alpha_\circ - \delta_1} + \frac{(\delta_3 + \delta_2 |u_\circ|)^2}{\beta_\circ - \delta_2} \, dx \right)^{\frac{1}{2}},$$

where $\overline{K}$ is from Theorem 2.5

PROOF. Using Theorem 2.5 and the majorant $\overline{M}_2$ from Lemma 2.3 we have

$$r = \sup_{u\in\mathcal{S}(\mathcal{D})} ||u - u_\circ||_\circ \leq \overline{K} \sup_{u\in\mathcal{S}(\mathcal{D})} ||u - u_\circ||_{\alpha,\beta}$$
$$\leq \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \inf_{y\in H^1} \overline{M}_{2(\alpha,\beta,f)}(u_\circ, y)$$
$$\leq \overline{K} \inf_{y\in H^1} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \overline{M}_{1(\alpha,\beta,f)}(u_\circ, y),$$

again we estimate the infimum from above by setting $y = \alpha_\circ u'_\circ$,

$$\begin{aligned} (2.14) \qquad &\leq \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \overline{M}_{1(\alpha,\beta,f)}(u_\circ, \alpha_\circ u'_\circ) \\ &= \overline{K} \sup_{(\alpha,\beta,f)\in\mathcal{D}} \left(\int_a^b \frac{1}{\alpha}(\alpha_\circ u'_\circ - \alpha u'_\circ)^2 + \frac{1}{\beta}((\alpha_\circ u'_\circ)' - f - \beta u_\circ)^2 \, dx \right)^{\frac{1}{2}}, \end{aligned}$$

now replace α, β, f with equivalent forms from Definition 1.5,

$$= \overline{K} \sup_{||g_i||_\infty \leq 1} \left(\int_a^b \frac{(\alpha_\circ u'_\circ - (\alpha_\circ + \delta_1 g_1)u'_\circ)^2}{\alpha_\circ + \delta_1 g_1} + \frac{((\alpha_\circ u'_\circ)' - (f_\circ + \delta_3 g_3) - (\beta_\circ + \delta_2 g_2)u_\circ)^2}{\beta_\circ + \delta_2 g_2} \, dx \right)^{\frac{1}{2}},$$

since $u_\circ$ satisfies (1.4) we have

$$= \overline{K} \sup_{||g_i||_\infty \leq 1} \left(\int_a^b \frac{(-\delta_1 g_1 u'_\circ)^2}{\alpha_\circ + \delta_1 g_1} + \frac{(-\delta_3 g_3 - \delta_2 g_2 u_\circ)^2}{\beta_\circ + \delta_2 g_2} \, dx \right)^{\frac{1}{2}},$$

we estimate g_i as the maximum in the denominator and minimum in the numerator,

$$\leq \overline{K} \left(\int_a^b \frac{(\delta_1 u'_\circ)^2}{\alpha_\circ - \delta_1} + \frac{(\delta_3 + \delta_2 |u_\circ|)^2}{\beta_\circ - \delta_2} \, dx \right)^{\frac{1}{2}}.$$

Now the proof follows from the fact that $\mathrm{Diam}(\mathcal{S}(\mathcal{D})) \leq 2r$. □

Next we present an alternate approach which does not require the mean solution $u_\circ$. This is based on the energy-estimate. Two different bounds can be made using the different majorants $\overline{M}_1$ and $\overline{M}_2$.

THEOREM 2.9. *Assume the boundary conditions of the uncertain problem $\mathcal{P}$ are zeros. Then the diameter of the solution set $\mathcal{S}(\mathcal{D})$ has a guaranteed upper bound given by*

$$Diam(\mathcal{S}(\mathcal{D})) \leq \overline{\mathcal{B}}_3 := 2\overline{C}_\oplus \left|\left| |f_\circ| + \delta_3 \right|\right| \left(H + \overline{\delta_2} \, \overline{C}_\oplus^2 + ||\delta_3|| \right),$$

where $\overline{C}_\oplus$ is from Theorem 2.4,

$$H := \sup_{x\in(a,b)} \left\{ \frac{\delta_1(x)}{(\alpha_\circ(x)(\alpha_\circ - \delta_1)(x))^{\frac{1}{2}}} \right\} \quad \textit{and} \quad \overline{\delta_2} := \sup_{x\in(a,b)} \{\delta_2(x)\}.$$

PROOF. We use the majorant $\overline{M}_1$ such that the exact solution is $u_\circ$. We denote $\overline{M}_{1(\alpha_\circ,\beta_\circ,f_\circ)} =: \overline{M}_{1\circ}$.

$$
\begin{aligned}
r &= \sup_{u\in\mathcal{S}(\mathcal{D})} ||u-u_\circ||_\circ = \sup_{u\in\mathcal{S}(\mathcal{D})} \inf_{y\in H^1} \overline{M}_{1\circ}(u,y) \\
&\le \inf_{y\in H^1} \sup_{u\in\mathcal{S}(\mathcal{D})} \overline{M}_{1\circ}(u,y) \le \sup_{\substack{(\alpha,\beta,f)\in\mathcal{D}\\ u=\mathcal{S}(\alpha,\beta,f)}} \overline{M}_{1\circ}(u,\alpha u') \\
&= \sup_{\substack{(\alpha,\beta,f)\in\mathcal{D}\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\left(\int_a^b \frac{1}{\alpha_\circ}(\alpha u' - \alpha_\circ u')^2 \, dx \right)^{\frac{1}{2}} + \overline{C}_F ||(\alpha u')' - \beta_\circ u - f_\circ|| \right) \\
&\le \sup_{\substack{(\alpha,\beta,f)\in\mathcal{D}\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\left(\int_a^b \frac{1}{\alpha_\circ}(\alpha u' - \alpha_\circ u')^2 \, dx \right)^{\frac{1}{2}} + \overline{C}_\oplus ||\beta u + f - \beta_\circ u - f_\circ|| \right).
\end{aligned}
$$

Replacing α, β, f with equivalent forms from (1.5)

$$
\begin{aligned}
&= \sup_{\substack{||g_i||_\infty\le 1\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\left(\int_a^b \frac{1}{\alpha_\circ}((\alpha_\circ + \delta_1 g_1)u' - \alpha_\circ u')^2 \, dx \right)^{\frac{1}{2}} \right. \\
&\qquad \left. + \ \overline{C}_\oplus ||(\beta_\circ + \delta_2 g_2)u + (f_\circ + \delta_3 g_3) - \beta_\circ u - f_\circ|| \right)
\end{aligned}
$$

$$
\begin{aligned}
&= \sup_{\substack{||g_i||_\infty\le 1\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\left(\int_a^b \frac{(\delta_1 g_1 u')^2}{\alpha_\circ} \, dx \right)^{\frac{1}{2}} + \overline{C}_\oplus ||\delta_2 g_2 u + \delta_3 g_3|| \right) \\
&\le \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\left|\left| \frac{\delta_1 u'}{(\alpha_\circ)^{\frac{1}{2}}} \right|\right| + \overline{C}_\oplus ||\delta_2 u|| + \overline{C}_\oplus ||\delta_3|| \right)
\end{aligned}
$$

(2.15)

$$
\le \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\sup_{x\in(a,b)} \left\{ \frac{\delta_1(x)}{(\alpha_\circ(x)(\alpha_\circ-\delta_1)(x))^{\frac{1}{2}}} \right\} ||u||_{\alpha,\beta} + \sup_{x\in(a,b)} \{\delta_2(x)\} \overline{C}_\oplus^2 ||u||_{\alpha,\beta} + \overline{C}_\oplus ||\delta_3|| \right),
$$

Denote

$$
H := \sup_{x\in(a,b)} \left\{ \frac{\delta_1(x)}{(\alpha_\circ(x)(\alpha_\circ-\delta_1)(x))^{\frac{1}{2}}} \right\} \quad \text{and} \quad \overline{\delta_2} := \sup_{x\in(a,b)} \{\delta_2(x)\}.
$$

We use the energy estimate from Theorem 2.4 to get rid of the unknown u in (2.15)

$$
\begin{aligned}
&\sup_{u\in\mathcal{S}(\mathcal{D})} \left(H||u||_{\alpha,\beta} + \overline{\delta_2}\, \overline{C}_\oplus^2 ||u||_{\alpha,\beta} + \overline{C}_\oplus ||\delta_3|| \right) \\
&\le H\overline{C}_\oplus \left|\left| |f_\circ| + \delta_3 \right|\right| + \overline{\delta_2}\, \overline{C}_\oplus^3 \left|\left| |f_\circ| + \delta_3 \right|\right| + \overline{C}_\oplus ||\delta_3|| \\
&= \overline{C}_\oplus \left|\left| |f_\circ| + \delta_3 \right|\right| \left(H + \overline{\delta_2}\, \overline{C}_\oplus^2 + ||\delta_3|| \right).
\end{aligned}
$$

Since $\mathrm{Diam}(\mathcal{S}(\mathcal{D})) \le 2r$ the proof follows. □

Next we do the same but we use the majorant $\overline{M}_2$.

THEOREM 2.10. *Assume the boundary conditions of the uncertain problem $\mathcal{P}$ are zeros. Then the diameter of the solution set $Diam(\mathcal{S}(\mathcal{D}))$ has a guaranteed upper bound given by*

$$Diam(\mathcal{S}(\mathcal{D})) \leq \overline{\mathcal{B}}_4 := 2\overline{C}_{\oplus}\left|\left||f_\circ| + \delta_3\right|\right|\left(H_1 + H_2\overline{C}_{\oplus}\right) + 2\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|,$$

where $\overline{C}_{\oplus}$ is from Theorem 2.4 and

$$H_1 := \operatorname*{ess\,sup}_{x\in(a,b)}\left\{\frac{\delta_1(x)}{(\alpha_\circ(x)(\alpha_\circ-\delta_1)(x))^{\frac{1}{2}}}\right\} \quad \textit{and} \quad H_2 := \operatorname*{ess\,sup}_{x\in(a,b)}\left\{\frac{\delta_2(x)}{(\beta_\circ(x))^{\frac{1}{2}}}\right\}.$$

PROOF. Following the proof of Theorem 2.9 we have

$$\begin{aligned}
r &\leq \sup_{\substack{(\alpha,\beta,f)\in\mathcal{D}\\ u=\mathcal{S}(\alpha,\beta,f)}} \overline{M}_{2\circ}(u,\alpha u') \\
&= \sup_{\substack{(\alpha,\beta,f)\in\mathcal{D}\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\int_a^b \frac{1}{\alpha_\circ}(\alpha u' - \alpha_\circ u')^2 + \frac{1}{\beta_\circ}((\alpha u')' - f_\circ - \beta_\circ u)^2\, dx\right)^{\frac{1}{2}} \\
&= \sup_{\substack{||g_i||_\infty\leq 1\\ u=\mathcal{S}(\alpha,\beta,f)}} \left(\int_a^b \frac{((\alpha_\circ+\delta_1 g_1)u' - \alpha_\circ u')^2}{\alpha_\circ} + \frac{((\beta_\circ+\delta_2 g_2)u + f_\circ + \delta_3 g_3 - \beta_\circ u - f_\circ)^2}{\beta_\circ}\, dx\right) \\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\int_a^b \frac{(\delta_1|u'|)^2}{\alpha_\circ} + \frac{(\delta_2|u|+\delta_3)^2}{\beta_\circ}\, dx\right)^{\frac{1}{2}} \\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\int_a^b \left(\frac{\delta_1|u'|}{(\alpha_\circ)^{\frac{1}{2}}} + \frac{\delta_2|u|+\delta_3}{(\beta_\circ)^{\frac{1}{2}}}\right)^2 dx\right)^{\frac{1}{2}} \\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\left(\int_a^b \frac{\delta_1^2}{\alpha_\circ}|u'|^2\, dx\right)^{\frac{1}{2}} + \left(\int_a^b \frac{\delta_2^2}{\beta_\circ}|u|^2\, dx\right)^{\frac{1}{2}} + \left(\int_a^b \frac{\delta_3^2}{\beta_\circ}\, dx\right)^{\frac{1}{2}}\right) \\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})} \left(\left(\int_a^b \frac{\delta_1^2}{\alpha_\circ(\alpha_\circ-\delta_1)}\alpha|u'|^2\, dx\right)^{\frac{1}{2}} + \left(\int_a^b \frac{\delta_2^2}{\beta_\circ}|u|^2\, dx\right)^{\frac{1}{2}} + \left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|\right).
\end{aligned}$$

Denote

$$H_1 := \operatorname*{ess\,sup}_{x\in(a,b)}\left\{\frac{\delta_1(x)}{(\alpha_\circ(x)(\alpha_\circ-\delta_1)(x))^{\frac{1}{2}}}\right\} \quad \text{and} \quad H_2 := \operatorname*{ess\,sup}_{x\in(a,b)}\left\{\frac{\delta_2(x)}{(\beta_\circ(x))^{\frac{1}{2}}}\right\}.$$

Now we have

$$
\begin{aligned}
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})}\left(\left(\int_a^b H_1^2\alpha|u'|^2\,dx\right)^{\frac{1}{2}}+\left(\int_a^b H_2^2|u|^2\,dx\right)^{\frac{1}{2}}+\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|\right)\\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})}\left(H_1\left(\int_a^b \alpha|u'|^2+\beta|u|^2\,dx\right)^{\frac{1}{2}}+H_2\left(\int_a^b |u|^2\,dx\right)^{\frac{1}{2}}+\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|\right)\\
&\leq \sup_{u\in\mathcal{S}(\mathcal{D})}\left(H_1||u||_{\alpha,\beta}+H_2\overline{C}_\oplus||u||_{\alpha,\beta}+\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|\right)\\
&\leq H_1\overline{C}_\oplus\left|\left||f_\circ|+\delta_3\right|\right|+H_2\overline{C}_\oplus^2\left|\left||f_\circ|+\delta_3\right|\right|+\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|\\
&= \overline{C}_\oplus\left|\left||f_\circ|+\delta_3\right|\right|\left(H_1+H_2\overline{C}_\oplus\right)+\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|.
\end{aligned}
$$

Since $\mathrm{Diam}(\mathcal{S}(\mathcal{D}))\leq 2r$ the proof follows. □

CHAPTER 3

Sensitivity Analysis and Numerical Approximation of Diam$(\mathcal{S}(\mathcal{D}))$

The contents of the following chapters are numerical results and the tests were implemented in Matlab. All of the codes used can be found at `https://github.com/Yzivv/Accuracy-Analysis-Thesis-Codes`.

In this chapter we investigate the sensitivity of the bounds $\underline{\mathcal{B}}$ and $\overline{\mathcal{B}}$ from Chapter 2 with respect to changes in the admissible data $\mathcal{D}$. We present numerical tests which illustrate that there is no obvious way to determine which bound will perform the best. Additionally, we show a method of numerically approximating Diam$(\mathcal{S}(\mathcal{D}))$ and include this approximation in all of the tests.

The following things are tested:

- Increasing the value of one of the indeterminacy parameters δ_i
- Increasing the difference between the coefficients $\alpha_\circ$, $\beta_\circ$ and $f_\circ$
- Adding oscillation to $\alpha_\circ$, $\beta_\circ$ or $f_\circ$

Recall that the formulas for the upper bounds are (Theorems 2.7, 2.8, 2.9, 2.10)

$$\begin{aligned}
\overline{B}_1 &= 2\overline{K}\left(\left(\int_a^b \frac{(\delta_1 u_\circ')^2}{\alpha_\circ - \delta_1}\,dx\right)^{\frac{1}{2}} + \overline{C}_F||\delta_2|u_\circ| + \delta_3||\right),\\
\overline{B}_2 &= 2\overline{K}\left(\int_a^b \frac{(\delta_1 u_\circ')^2}{\alpha_\circ - \delta_1} + \frac{(\delta_3 + \delta_2|u_\circ|)^2}{\beta_\circ - \delta_2}\,dx\right)^{\frac{1}{2}},\\
\overline{B}_3 &= 2\overline{C}_\oplus\left|\left||f_\circ| + \delta_3\right|\right|\left(H + \overline{\delta_2}\,\overline{C}_\oplus^2 + ||\delta_3||\right),\\
\overline{B}_4 &= 2\overline{C}_\oplus\left|\left||f_\circ| + \delta_3\right|\right|\left(H_1 + H_2\overline{C}_\oplus\right) + 2\left|\left|\frac{\delta_3}{\beta_\circ^{\frac{1}{2}}}\right|\right|
\end{aligned}$$

and the lower bound is (Theorem 2.6)

$$\underline{B} = \frac{\int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2 + \delta_3|u_\circ|\,dx}{\left(||u_\circ||_\circ^2 - \int_a^b \delta_1|u_\circ'|^2 + \delta_2|u_\circ|^2\,dx\right)^{\frac{1}{2}}}.$$

It is clear that there is little hope of understanding these formulas intuitively. In our numerical tests we find that $\overline{B}_1$ may be the most commonly reliable but the recommended tactic is to simply calculate all of the bounds to find the best one.

3.1. Brute-force Method for Approximating the Diameter

The exact value of Diam($\mathcal{S}(\mathcal{D})$) is not a computable quantity. It can be approximated by taking a finite subset $U \in \mathcal{S}(\mathcal{D})$ and computing the quantity

$$\text{Approx} = \sup_{v,w \in U} ||v - w||_\circ \approx \text{Diam}(\mathcal{S}(\mathcal{D}))$$

In order to quickly generate a subset $U \in \mathcal{S}(\mathcal{D})$ we choose the coefficients α, β, f from $\mathcal{D}$ such that they are piecewise constant. This reduces the differential equation (1.4) to a linear system of equations which can be rapidly solved. Note that functions of this type are dense in the set $\mathcal{D}$ if we allow arbitrary amount of pieces in the domain. In practice the mesh needs to be fine enough to reasonably account for the oscillations of $\mathcal{D}$.

After the subset U is created it remains to compute the quantity $\sup_{v,w \in U} ||v-w||_\circ$. This takes the bulk of the computation time. We use the most obvious algorithm to compute it (see Algorithm 1). The algorithm calculates the distance between each possible pair in the set and outputs the largest distance found. This algorithm has time complexity $O(\frac{n^2}{2})$ when duplicate comparisons are removed.

The formulas for how the problem $\mathcal{P}$ can be solved for piecewise constant coefficients are as follows. Let $[x_0, \cdots, x_n]$ be a mesh with n points in the domain $[a, b]$, where $x_0 = a$ and $x_n = b$ and let the coefficient functions be piecewise constants such that

$$\begin{aligned} \alpha(x) &= \alpha_k, \text{ when } x \in [x_{k-1}, x_k], \\ \beta(x) &= \beta_k, \text{ when } x \in [x_{k-1}, x_k], \\ f(x) &= f_k, \text{ when } x \in [x_{k-1}, x_k]. \end{aligned}$$

The differential equation (1.4) can now be solved separately on each subinterval with boundary conditions imposed by the adjacent equations. A single equation of the type

$$\alpha u'' - \beta u = f$$

where α, β and f are constant has a general solution of the form

$$u = c_1 e^{\left(\sqrt{\beta/\alpha}\right)x} + c_2 e^{-\left(\sqrt{\beta/\alpha}\right)x} - \frac{f}{\beta}. \tag{3.1}$$

With n subintervals we will have n equations of the form (3.1). We denote the solutions $u_1, u_2, \cdots, u_n$ such that

$$u_k(x) = c_{k1} e^{\left(\sqrt{\beta_k/\alpha_k}\right)x} + c_{k2} e^{-\left(\sqrt{\beta_k/\alpha_k}\right)x} - \frac{f_k}{\beta_k}.$$

This leads to $2n$ unknown parameters c_{ij}. On the boundary of each subinterval the function values and derivative values must be equal. The boundary conditions (1.4) also need to be satisfied. Now there are $2n$ equations which is enough to solve

the system. The system has the form

$$\begin{cases} u_1(x_0) & = A, \\ u_i(x_i) - u_{i+1}(x_i) & = 0 \quad \text{for } i = 1, 2, \ldots, n-1, \\ u_i'(x_i) - u_{i+1}'(x_i) & = 0 \quad \text{for } i = 1, 2, \ldots, n-1, \\ u_n(x_n) & = B. \end{cases} \tag{3.2}$$

To help with notation denote

$$h_k = \sqrt{\frac{\beta_k}{\alpha_k}}.$$

Writing out (3.2) we have

$$\begin{cases} c_{11}e^{h_1x_0} + c_{12}e^{-h_1x_0} - \frac{f_1}{\beta_1} & = A, \\ c_{i1}e^{h_ix_i} + c_{i2}e^{-h_ix_i} - \frac{f_i}{\beta_i} - c_{(i+1)1}e^{h_{i+1}x_i} - c_{i2}e^{-h_{i+1}x_i} + \frac{f_{i+1}}{\beta_{i+1}} & = 0, \text{ for } i = 1, \ \ldots \ , n-1, \\ h_i\left(c_{i1}e^{h_ix_i} - c_{i2}e^{-h_ix_i}\right) - h_{i+1}\left(c_{(i+1)1}e^{h_{i+1}x_i} - c_{i2}e^{-h_{i+1}x_i}\right) & = 0, \text{ for } i = 1, \ \ldots \ , n-1, \\ c_{n1}e^{h_nx_n} + c_{n2}e^{-h_nx_n} - \frac{f_n}{\beta_n} & = B, \end{cases}$$

In matrix form this becomes

$$\boldsymbol{A}\boldsymbol{c} = \boldsymbol{F},$$

where

$$\boldsymbol{A} = \begin{bmatrix} e^{h_1x_0} & e^{-h_1x_0} & 0 & 0 & 0 & 0 & \cdots & 0 & 0 \\ e^{h_1x_1} & e^{-h_1x_1} & -e^{h_2x_1} & -e^{-h_2x_1} & 0 & 0 & \cdots & 0 & 0 \\ h_1e^{h_1x_1} & -h_1e^{-h_1x_1} & -h_2e^{h_2x_1} & h_2e^{-h_2x_1} & 0 & 0 & \cdots & 0 & 0 \\ 0 & 0 & e^{h_2x_2} & e^{-h_2x_2} & -e^{h_3x_2} & -e^{-h_3x_2} & \cdots & 0 & 0 \\ 0 & 0 & h_2e^{h_2x_2} & -h_2e^{-h_2x_2} & -h_3e^{h_3x_2} & h_3e^{-h_3x_2} & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \cdots & e^{h_nx_n} & e^{-h_nx_n} \end{bmatrix},$$

$$\boldsymbol{c} = \begin{bmatrix} c_{11} \\ c_{12} \\ c_{21} \\ c_{22} \\ c_{31} \\ c_{32} \\ \vdots \\ c_{n1} \\ c_{n2} \end{bmatrix} \quad \text{and} \quad \boldsymbol{F} = \begin{bmatrix} \frac{f_1}{\beta_1} + A \\ \frac{f_1}{\beta_1} - \frac{f_2}{\beta_2} \\ 0 \\ \frac{f_2}{\beta_2} - \frac{f_3}{\beta_3} \\ 0 \\ \vdots \\ \frac{f_n}{\beta_n} + B \end{bmatrix}.$$

A linear system of equations of this kind is solved by numerical programs very quickly. In our case, we used matlab. Algorithm 1 is used to find the quantity $\sup_{v,w\in U} ||v - w||$.

Algorithm 1: Brute Force Algorithm

```
Compute the Subset U ∈ S(D).
N = Amount of elements in U
Denote the nth element of U as U(n).
result = 0
for k=1:N
  for h=k+1:N
    distance = ||U(k) − U(h)||∘
    if distance > result
      result = distance
Approximated Diameter = result
```

Result: Approximated Diameter

3.2. Numerical Tests for Sensitivity

We start by setting each of the coefficient functions α, β and f as the constant value 1. Later tests add modifications to the first test. The boundary conditions are zeros and the domain is $(0, 1)$ in all tests.

Test 1. The mean coefficients are

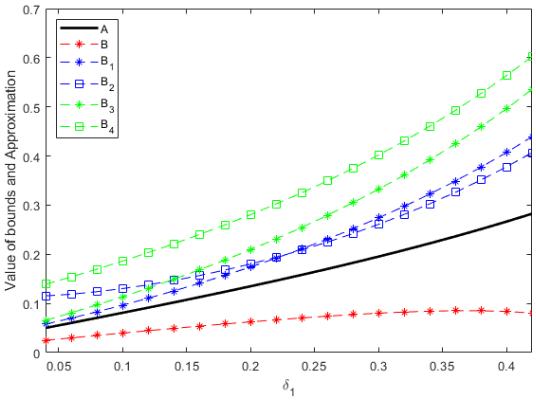
$$\alpha_\circ(x) = \beta_\circ(x) = f_\circ(x) = 1,$$

and the indeterminacy parameters are

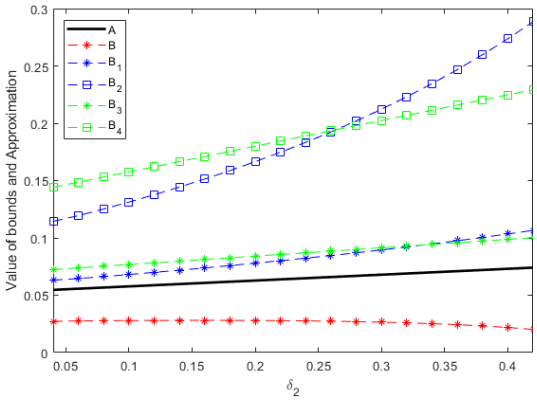
$$\delta_1(x) = \delta_2(x) = \delta_3(x) = 0.05.$$

We test how changing each of the indeterminacy parameters affects the bounds. In Figure 3.1 we can see that the magnitude of the bounds and approximations is most sensitive to changes in δ_1 and δ_3. Indeterminacy of β has a very small effect on the approximation. Changes with respect to δ_1 and δ_2 are nonlinear and with respect to changes of δ_3 they are linear for both the approximation and the bounds.

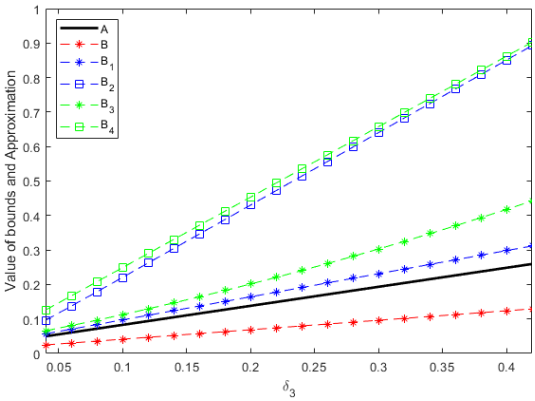
The bounds $\overline{\mathcal{B}}_1$ and $\overline{\mathcal{B}}_3$ which use the majorant $\overline{M}_1$ are sharper then their counterparts $\overline{\mathcal{B}}_2$ and $\overline{\mathcal{B}}_4$ which use the majorant $\overline{M}_2$. The best bound to use in this case would be $\overline{\mathcal{B}}_1$.



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

FIGURE 3.1. Test 1 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

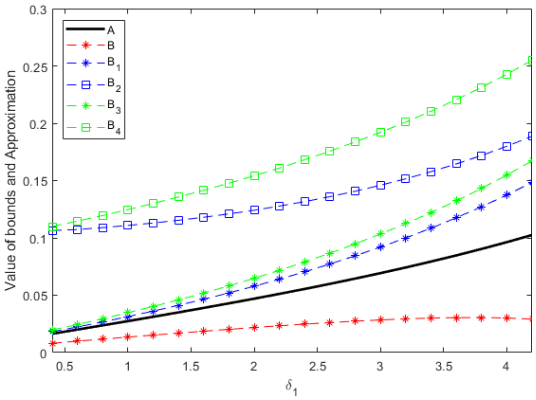
Test 2. Results are shown in Figure 3.2. The mean coefficients are

$$\alpha_\circ(x) = 10, \quad \beta_\circ(x) = f_\circ(x) = 1,$$

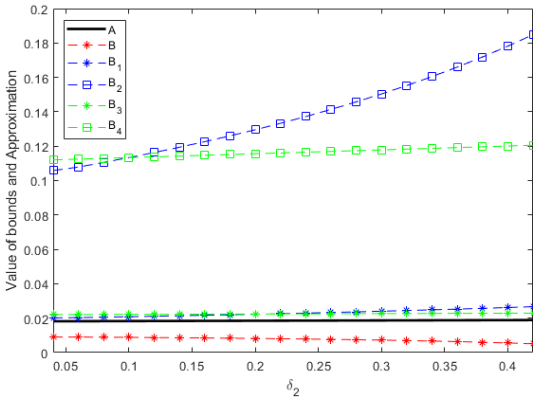
and the indeterminacy parameters are

$$\delta_1(x) = 0.5, \quad \delta_2(x) = \delta_3(x) = 0.05.$$

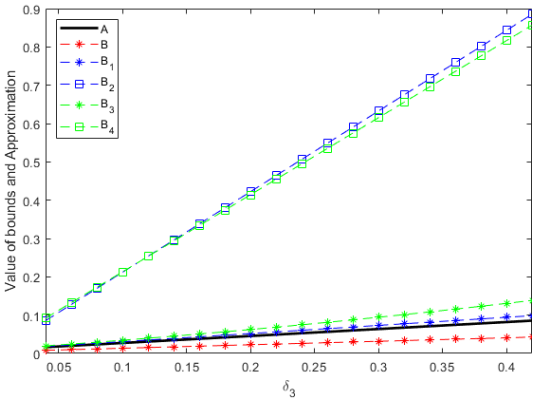
The magnitude of indeterminacy when any of the indeterminacy parameters are increased is now lower then in test 1 but the best bound is still $\overline{\mathcal{B}}_1$. Differences between using the majorant $\overline{M}_1$ and $\overline{M}_2$ is even higher. $\overline{\mathcal{B}}_1$ and $\overline{\mathcal{B}}_3$ are much sharper then the bounds that use $\overline{M}_2$. This effect is most prominent when δ_3 is increased since $\overline{\mathcal{B}}_2$ and $\overline{\mathcal{B}}_4$ rapidly increase and even start off rather far from the approximation.



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

FIGURE 3.2. Test 2 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

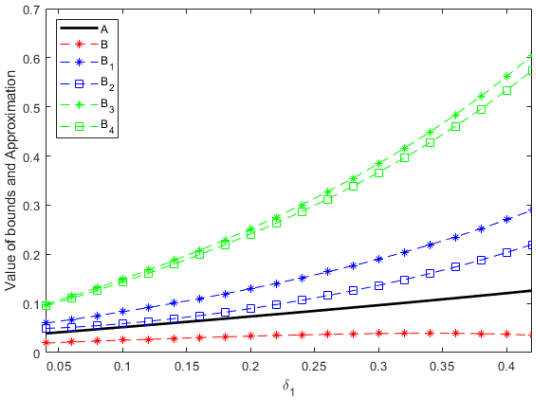
Test 3. Results are shown in Figure 3.3. The mean coefficients are

$$\alpha_\circ(x) = 1, \quad \beta_\circ(x) = 10, \quad f_\circ(x) = 1,$$

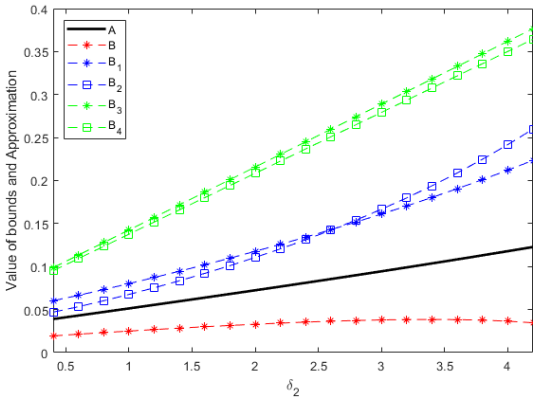
and the indeterminacy parameters are

$$\delta_1(x) = 0.05, \quad \delta_2(x) = 0.5, \quad \delta_3(x) = 0.05.$$

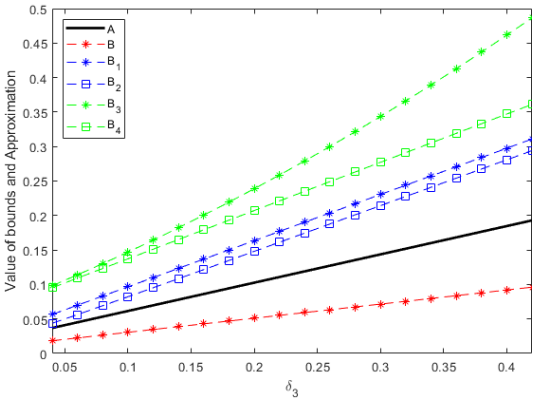
In this case, the bound $\overline{\mathcal{B}}_2$ is outperforming $\overline{\mathcal{B}}_1$. The main difference is not with which majorant is used. Bounds $\overline{\mathcal{B}}_1$ and $\overline{\mathcal{B}}_2$ use the mean solution and outperform the bounds $\overline{\mathcal{B}}_3$ and $\overline{\mathcal{B}}_4$ which use the energy-estimate. As we stated in Chapter 2, when the value of β is high the majorant $\overline{M}_2$ is usually better which we see since $\overline{\mathcal{B}}_2$ is sharper then $\overline{\mathcal{B}}_1$ and $\overline{\mathcal{B}}_4$ is sharper then $\overline{\mathcal{B}}_3$.



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

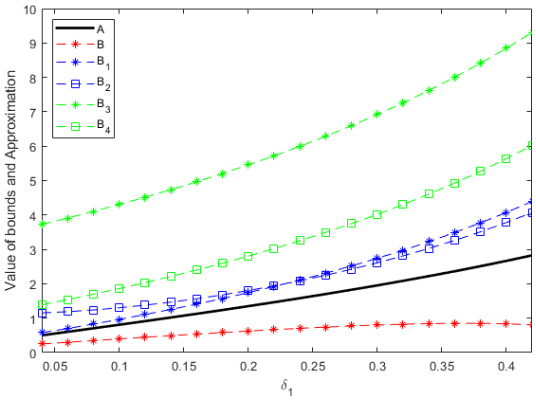
FIGURE 3.3. Test 3 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

Test 4. Results are shown in Figure 3.4. The mean coefficients are

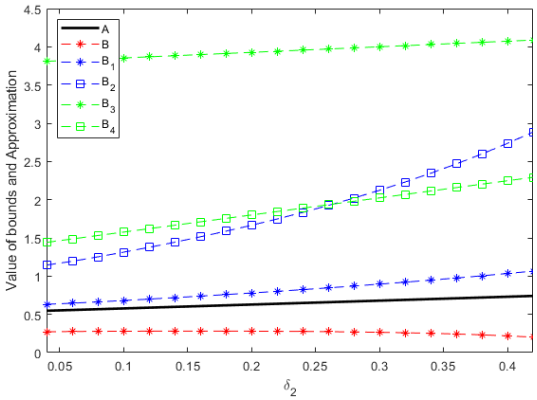
$$\alpha_\circ(x) = 1, \quad \beta_\circ(x) = 1, \quad f_\circ(x) = 10,$$

and the indeterminacy parameters are

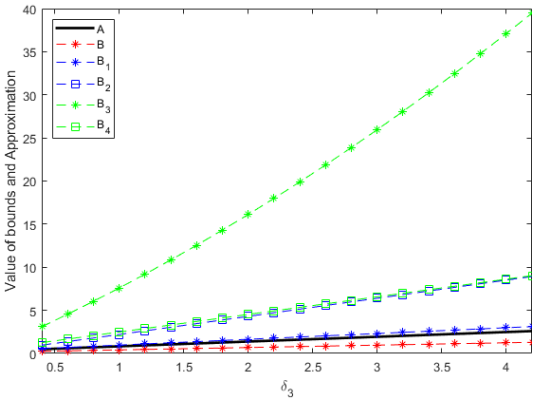
$$\delta_1(x) = 0.05, \quad \delta_2(x) = 0.05, \quad \delta_3(x) = 0.5.$$



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

FIGURE 3.4. Test 4 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

Note that the magnitudes of approximations and bounds are very high now. When δ_3 or δ_2 is changed the only bound that remains quite sharp is $\overline{\mathcal{B}}_1$. The energy-estimate bound $\overline{\mathcal{B}}_3$ is extremely coarse in all cases even though it uses the same majorant as $\overline{\mathcal{B}}_1$.

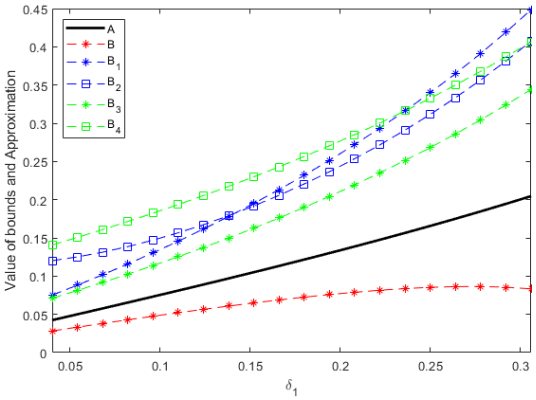
Test 5. Results are shown in Figure 3.5. In the remaining tests 5-7 we add oscillation to one of the coefficients. The mean coefficients are

$$\alpha_\circ(x) = \frac{\sin(12x)}{4} + 1, \quad \beta_\circ(x) = 1, \quad f_\circ(x) = 1,$$

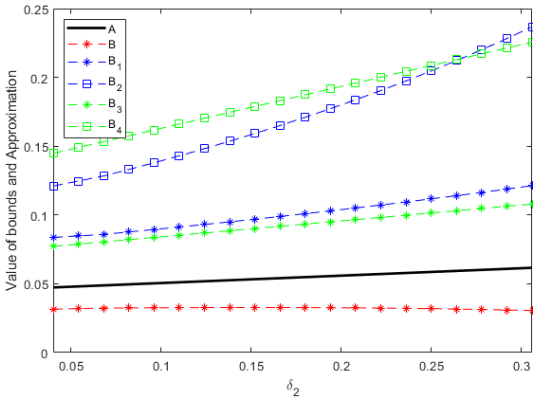
and the indeterminacy parameters are

$$\delta_1(x) = 0.05, \quad \delta_2(x) = 0.05, \quad \delta_3(x) = 0.05.$$

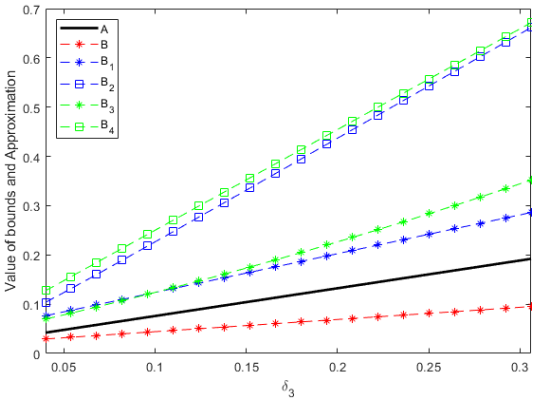
Adding oscillation to the coefficient α interestingly causes the energy estimate bound $\overline{\mathcal{B}}_3$ to perform the best. The bounds which use the majorant $\overline{M}_1$ are still the best since β is rather small. Compared to test 1, the magnitudes are quite similar.



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

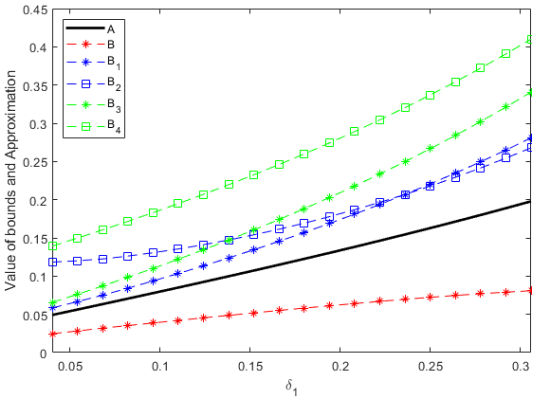
FIGURE 3.5. Test 5 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

Test 6. Results are shown in Figure 3.6. The mean coefficients are

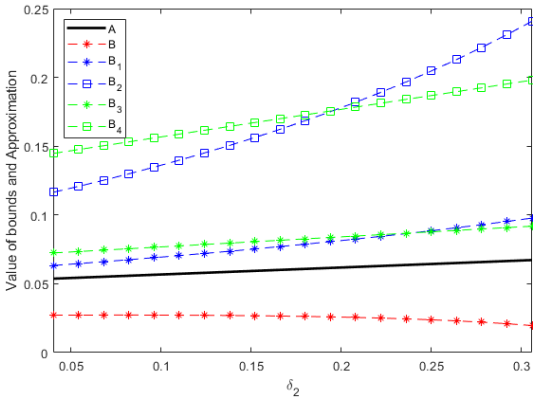
$$\alpha_\circ(x) = 1, \quad \beta_\circ(x) = \frac{\sin(12x)}{4} + 1, \quad f_\circ(x) = 1,$$

and the indeterminacy parameters are

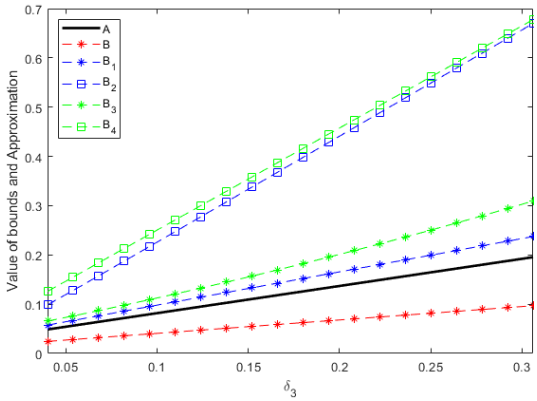
$$\delta_1(x) = 0.05, \quad \delta_2(x) = 0.05, \quad \delta_3(x) = 0.05.$$



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

FIGURE 3.6. Test 6 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

Adding a bit of oscillation to β has almost no effect on the bounds or approximation. Comparing with test 1 the figures are nearly identical.

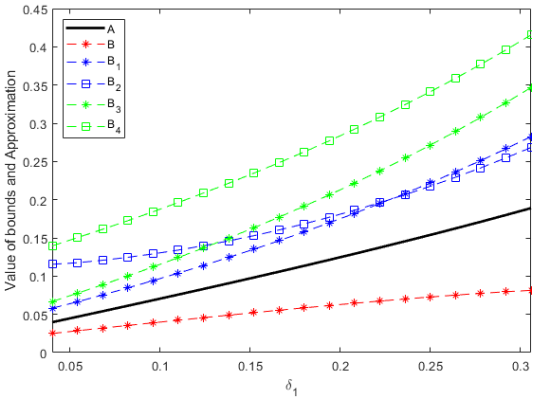
Test 7. Results are shown in Figure 3.7. The mean coefficients are

$$\alpha_\circ(x) = 1, \quad \beta_\circ(x) = 1, \quad f_\circ(x) = \frac{\sin(12x)}{4} + 1,$$

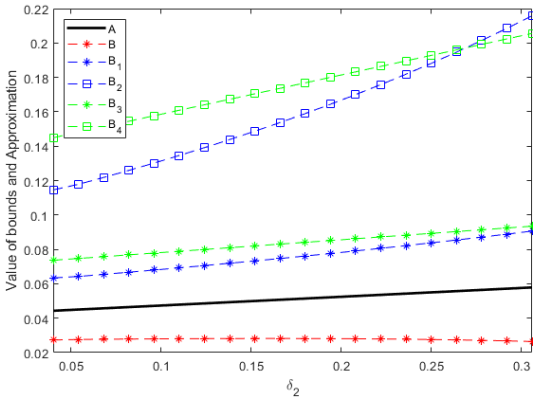
and the indeterminacy parameters are

$$\delta_1(x) = 0.05, \quad \delta_2(x) = 0.05, \quad \delta_3(x) = 0.05.$$

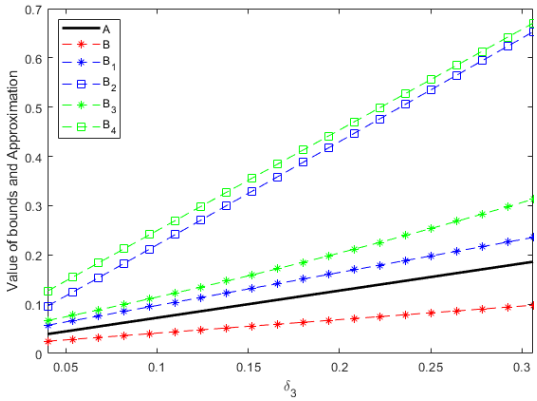
As with test 6, a small oscillation in f is having very little effect on the results. Magnitude is more or less the same as in test 1 and the bounds are similarly spaced from one another.



(A) Changing δ_1.



(B) Changing δ_2.



(C) Changing δ_3.

FIGURE 3.7. Test 7 Results. A is the brute-force approximation, B is the lower bound and B_1-B_4 are the upper bounds.

CHAPTER 4

Machine Learning Model for Quantifying Size of the Solution Set

In this chapter we create neural networks which approximate the brute-force method discussed in section (3.1) (see Figure 4.1). In our case, since we have analytical methods to control $\mathrm{Diam}(\mathcal{S}(\mathcal{D}))$, it is not clear why one would want to use such a model. The main advantages compared to the brute-force method and analytical bounds are:

- Using a model like this requires no knowledge of analysis
- It is inexpensive to use unlike the brute-force method
- Analytical bounds can be coarse (see Chapter 3)

Even in such a simple case, it is reasonable to use this type of model. It is even more interesting if this method can be extended to problems where analytical methods are not an option and the brute-force method is even more expensive. We only focus on problem $\mathcal{P}$ but in future research this will be tried for more complicated problems. The main goal of this chapter is to give a proof of concept that training machine learning models to analyze uncertainty errors is a feasible task.

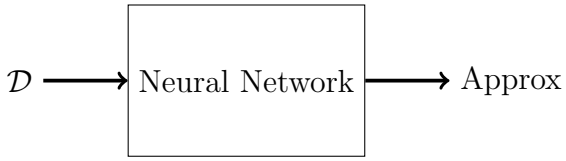


FIGURE 4.1. Purpose of the model.

We do not go into technical details about how neural networks work and for those who want to review basic facts about machine learning and neural networks we recommend the freely available books [3] or for finnish readers [7].

4.1. Model Description

The size and geometry of our model depends on how general the inputs $\mathcal{D}$ can be. Simple dense neural networks are suitable for our analysis but in more complicated problems it may be required to use some more elaborate models.

First we transform the set $\mathcal{D}$ into something that can be input into a neural network which has a finite input layer. This is done by scanning the "strip" of admissible data of each function α, β and f. We pick a finite number of equidistant points $[x_0, \ldots, x_n]$ in the domain (a, b) and look at the maximum and minimum values of each function at these points.

In Figure 4.2 the function admissible dataset $\mathcal{D}_\alpha$ is scanned at ten points. This is done also for $\mathcal{D}_\beta$ and $\mathcal{D}_f$ so in total this produces 60 inputs for our neural network.

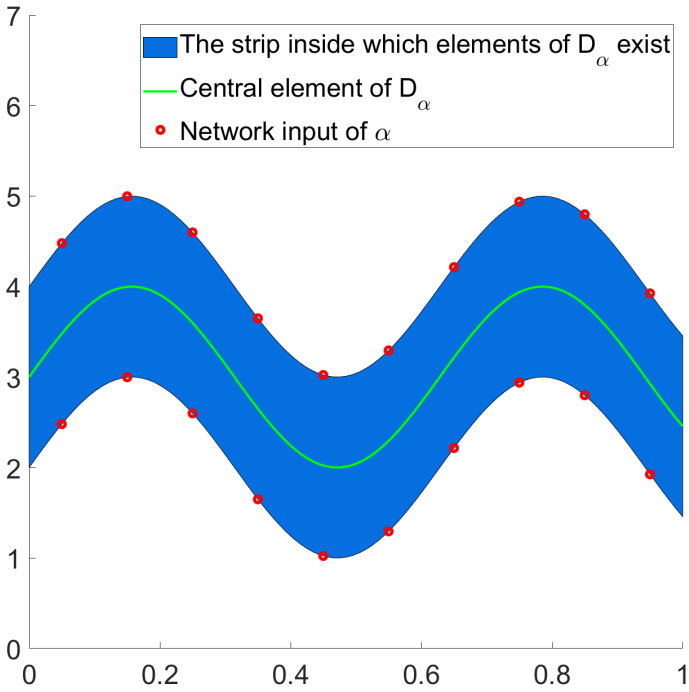


FIGURE 4.2. Admissible dataset $\mathcal{D}_\alpha$ and network input.

The amount of looking points is at our disposal and a good choice depends on how much the value of δ_i changes in the domain and how much the central element $\alpha_\circ$ is allowed to oscillate.

The training output of our model is generated with the brute-force method discussed in section (3.1). This means that our network output is not the true diameter but an approximation of an approximation. In our case, the brute-force method works well since the problem is simple but how one makes sure that approximations are accurate is non-trivial in general.

4.2. Constraining and Generating Training Data

For all of the example models we present in the next section we will have constraints on the possible inputs $\mathcal{D}$. Usually it would not make sense to include an admissible dataset where

$$(\alpha(x) = 10^{50}, \beta(x) = 10^{50}, f(x) = 10^{50}) \in \mathcal{D}.$$

In real world problems we have some clue as to what range our coefficients can reasonably be in and also how pathologically they can oscillate. For this reason we impose that the mean coefficients and indeterminacy parameters of each example in our training data are in some constraint sets

$$(\alpha_\circ, \beta_\circ, f_\circ) \in \mathcal{C}_\circ,$$
$$(\delta_1, \delta_2, \delta_3) \in \mathcal{C}_u.$$

The choice of these sets $\mathcal{C}_\circ$ and $\mathcal{C}_u$ is up to us and we will have a different choice for each of our example models.

Once we have chosen the constraints for our training examples we generate a sufficiently large set of input-output pairs to use as training data for our neural network. The generation of this training set is described in algorithm 2.

Algorithm 2: Generating Training Data

Pick the constraint sets for training data $\mathcal{C}_\circ$, $\mathcal{C}_u$
Pick the size of the training set and amount of looking points
  size of training set = N;
  amount of looking points = K;
Initialize the network input and target output matrices
  NetworkInputs = zeros(6K,N);
  TargetOutputs = zeros(1,N);
while $i \leq N$ **do**
  Pick random mean coefficients and indeterminacies such that
    $(\alpha_\circ, \beta_\circ, f_\circ) \in \mathcal{C}_\circ$;
    $(\delta_1, \delta_2, \delta_3) \in \mathcal{C}_u$;
  Scan the indeterminacy set formed by the above choices
    X = scanned input vector;
    NetworkInputs(:,i) = X;
  Compute the brute-force approximation of Diam($\mathcal{S}(\mathcal{D})$)
    output = Brute Force Approximation;
    TargetOutputs(1,i) = output;
end
Result: NetworkInputs and TargetOutputs

4.3. Example Models

We present two different example models. The models differ on four things: the amount of looking-points where the admissible data is scanned (see Figure 4.2), the constraints on the inputs that our training data has, the size of the training data and the structure of the neural network.

In real-world problems we always have some idea of the kind of admissible datasets our problem setting may have. This would guide us in how we construct our training data. In our case, we do not have a real-world example but we still constrain the training data to a reasonable generality.

For creating and training our neural networks we used the Deep Learning Toolbox of Matlab. We will present present the results of training for each model. The algorithm used for training the networks is the Levenberg-Marquardt algorithm.

4.3.1. Model 1. The first model has the following structure:

(1) The number of looking-points is 4.
(2) Neural Network has an input layer with 24 neurons and one hidden layer with 10 neurons using sigmoid activation (see Figure 4.3).
(3) The possible training inputs have the constraints (4.1)-(4.3).
(4) We use 2000 training examples which are split into 70% training, 15% validation and 15% testing.

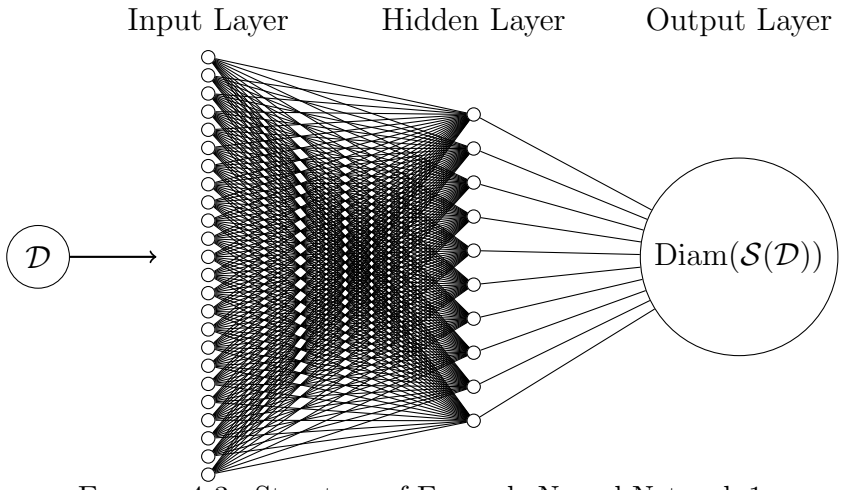


FIGURE 4.3. Structure of Example Neural Network 1.

We use a special case of possible training inputs $\mathcal{D}$. The central functions and indeterminacy parameters are piecewise constant such that

$$\begin{aligned} \alpha_\circ(x) &= \alpha_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ \beta_\circ(x) &= \beta_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ f_\circ(x) &= f_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ \delta_j(x) &= \delta_{ji}, \quad \text{when} \quad x \in [x_{i-1}, x_i]. \end{aligned} \tag{4.1}$$

where $[x_0, x_1, x_2, x_3, x_4] = [0, 0.25, 0.5, 0.75, 1]$. The mean elements and indeterminacies are given range constraints

$$\alpha_\circ(x) \in [1, 3], \quad \beta_\circ(x) \in [1, 3], \quad f_\circ(x) \in [3, 10] \tag{4.2}$$

$$\delta_1(x) \leq 0.3\alpha_\circ(x), \quad \delta_2(x) \leq 0.3\beta_\circ(x), \quad \delta_3(x) \leq 0.3f_\circ(x). \tag{4.3}$$

For illustration of a possible input $\mathcal{D}_1$ and the corresponding solution set see Figures (4.4) and (4.5). In Figure (4.4) the input $\mathcal{D}_1$ is scanned at four points on the domain. This produces the input vector

$$\begin{aligned} \text{Network Input: } \hat{\mathcal{D}_1} = {} & 1.1, 1.5, 1.9, 1.5, 1.7, 2.5, 2.5, 2.4, 1.8, 1.4, 2.0, 1.2, \\ & 3.2, 2.5, 3.4, 1.7, 5.1, 5.9, 3.8, 5.2, 7.0, 7.2, 5.7, 9.3. \end{aligned}$$

Calculating the brute-force approximation of the diameter we find in this case the value

$$0.7556 \approx \text{Diam}(\mathcal{S}(\mathcal{D}_1)).$$

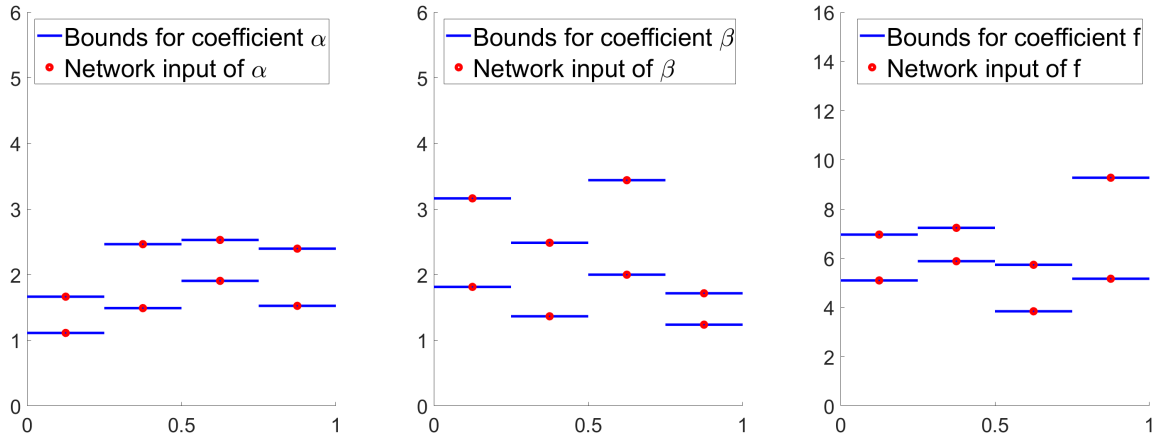


FIGURE 4.4. Example input $\mathcal{D}_1$.

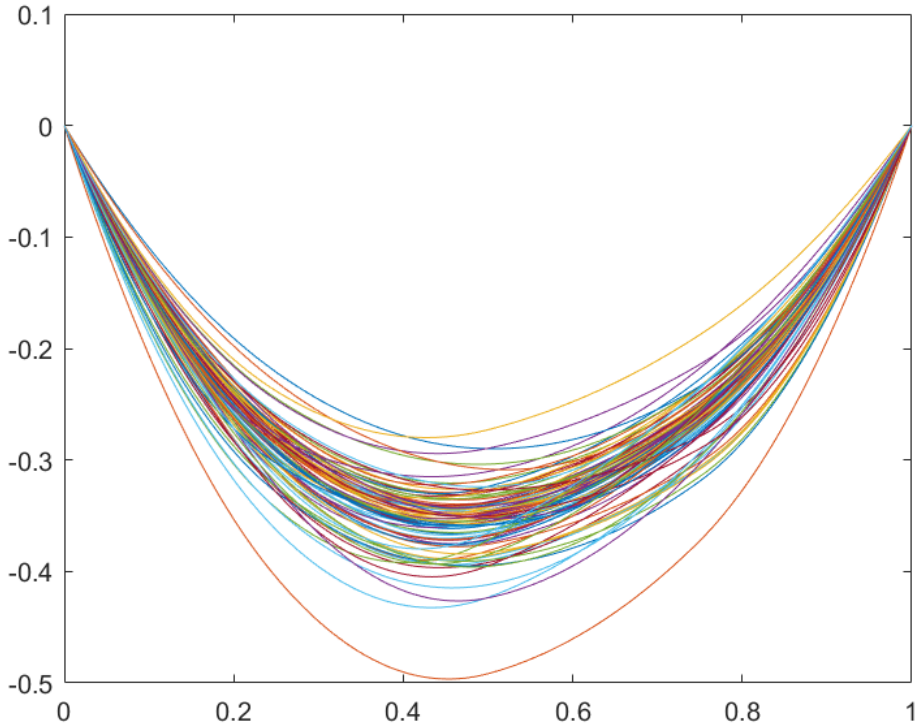


FIGURE 4.5. Piece of the solution set $\mathcal{S}(\mathcal{D}_1)$.

For the training we used the default settings of the Levenberg-Marquardt algorithm in the matlab nntool (see Figure)

Training Info | Training Parameters

showWindow	true	mu	0.001
showCommandLine	false	mu_dec	0.1
show	25	mu_inc	10
epochs	1000	mu_max	10000000000
time	Inf		
goal	0		
min_grad	1e-07		
max_fail	6		

FIGURE 4.6. Training parameters of the Levenberg-Marquardt algorithm.

In Figure 4.7 on the x-axis are the target output values which are the brute-force approximations and on the y-axis are the output values of the neural network. Each circle "Data" represents one example data point. We can see that the model produces results which are very close to the correct output on the validation and test sets too, so no overfitting is present. Figure 4.8 shows the training progression of our model. In this case the model trains very quickly.

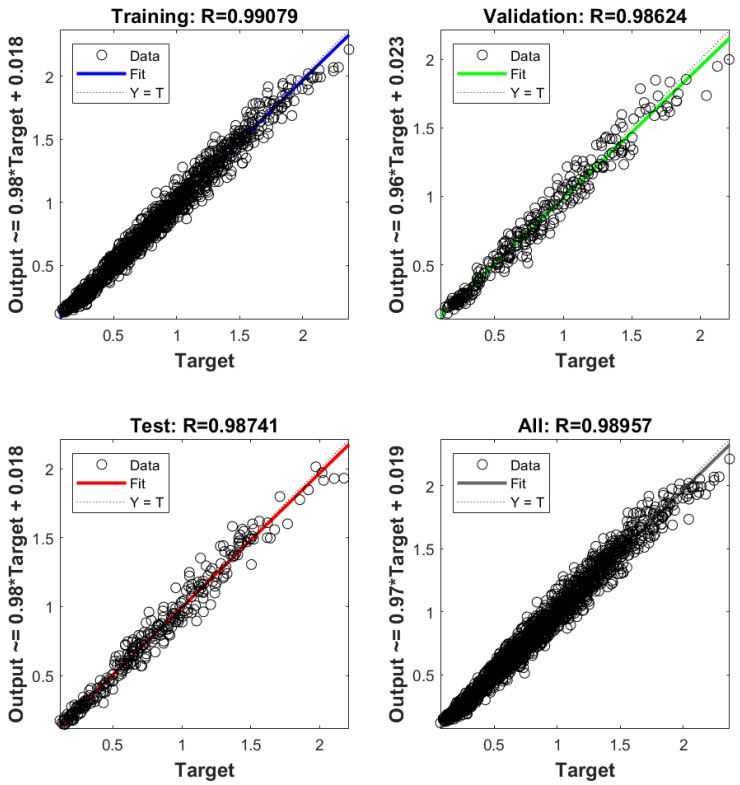


FIGURE 4.7. Model 1 Regression.

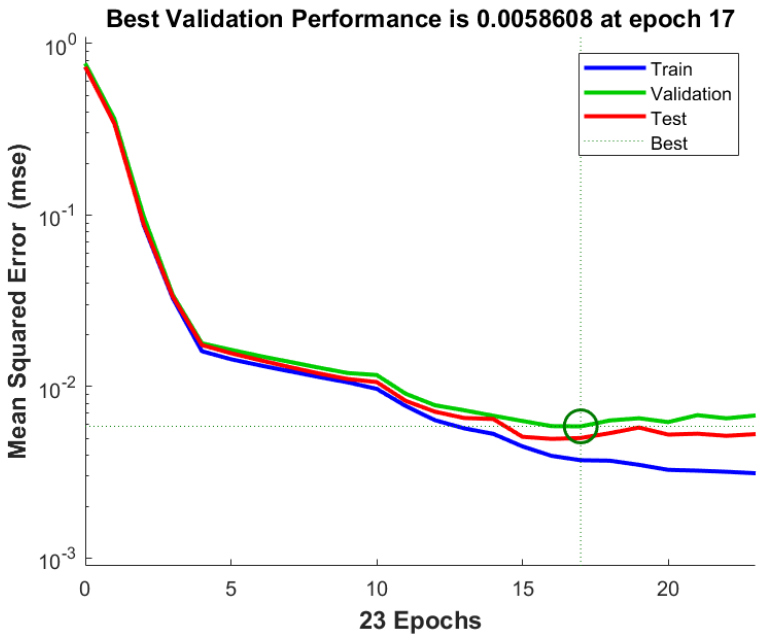


FIGURE 4.8. Model 1 Performance Curve.

4.3.2. Model 2. Model 2 has the following structure:

(1) The number of looking-points is 10.
(2) Neural Network has an input layer with 60 neurons and one hidden layer with N neurons using sigmoid activation (see Figure).
(3) The possible training inputs have the constraints (4.4)-(4.6).
(4) We use training examples which are split into 70% training, 15% validation and 15% testing.

Model 2 has slightly looser constraints on the possible inputs then model 1. The central functions and indeterminacy constants are still piecewise constant such that

$$\begin{aligned} \alpha_\circ(x) &= \alpha_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ \beta_\circ(x) &= \beta_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ f_\circ(x) &= f_{\circ i}, \quad \text{when} \quad x \in [x_{i-1}, x_i], \\ \delta_j(x) &= \delta_{ji}, \quad \text{when} \quad x \in [x_{i-1}, x_i]. \end{aligned} \tag{4.4}$$

where $[x_0, \ldots, x_{10}] = [0, \ldots, 1]$. The mean elements and indeterminacies have range constraints

$$\alpha_\circ(x) \in [1, 5], \quad \beta_\circ(x) \in [1, 5], \quad f_\circ(x) \in [1, 10] \tag{4.5}$$

$$\delta_1(x) \leq 0.3\alpha_\circ(x), \quad \delta_2(x) \leq 0.3\beta_\circ(x), \quad \delta_3(x) \leq 0.3f_\circ(x). \tag{4.6}$$

For the training we used the same settings as for model 1 which are the default settings of the training algorithm in matlab (see Figure 4.6).

In Figure on the x-axis are the target output values which are the brute-force approximations and on the y-axis are the output values of the neural network. Each circle "Data" represents one example data point. We can see that the model produces results which are very close to the correct output on the validation and test sets too,

so no overfitting is present. Figure shows the training progression of our model. In this case the model trains very quickly.

4.4. Comparing Analytical, Brute-force and ML-Model Methods

In this section we analyse how models 1 and 2 perform compared to the analytical bounds derived in Chapter 2 and the brute force approximations. First we test on the class of admissible datasets that was used for training and then we try other more general types of inputs to see if our models are able to generalize to them.

We find that the models are quite competitive with the other methods. However there are also problems. The most clear problem is that the training data made for model 2 has not been sufficiently accurate. The numerical brute-force approximations for this model were not sharp enough which causes a systematic error where the network model 2 outputs approximations that are substantially smaller then we would hope.

Adding some more general oscillations to the admissible dataset that is used as input has a surprisingly small effect on how well the models perform. Even in the final test where the admissible dataset can oscillate quite heavily, model 1 which was trained only on piecewise datasets made from 4 pieces performs well.

4.4.1. Test 1: Data similar to training data. Here we present tests where model 1 is evaluated on a test set made the same way as the training data for model 1. In the same way model 2 is tested on a test set similar to the training dataset of model 2.

First using the constraints (4.1)-(4.3) we create a test set of 500 examples. For this test set we compute the lower bound from Theorem 2.6 and the upper bound from Theorem 2.7. We compare the outputs of model 1, the bounds and the brute-force approximations in Figure 4.9.

Model 1 performs very well in this test. In every example the network output is between the analytical bounds and very close to the brute-force approximation. For some datapoints it would be preferable to use model 1 since the analytical bounds are very coarse.

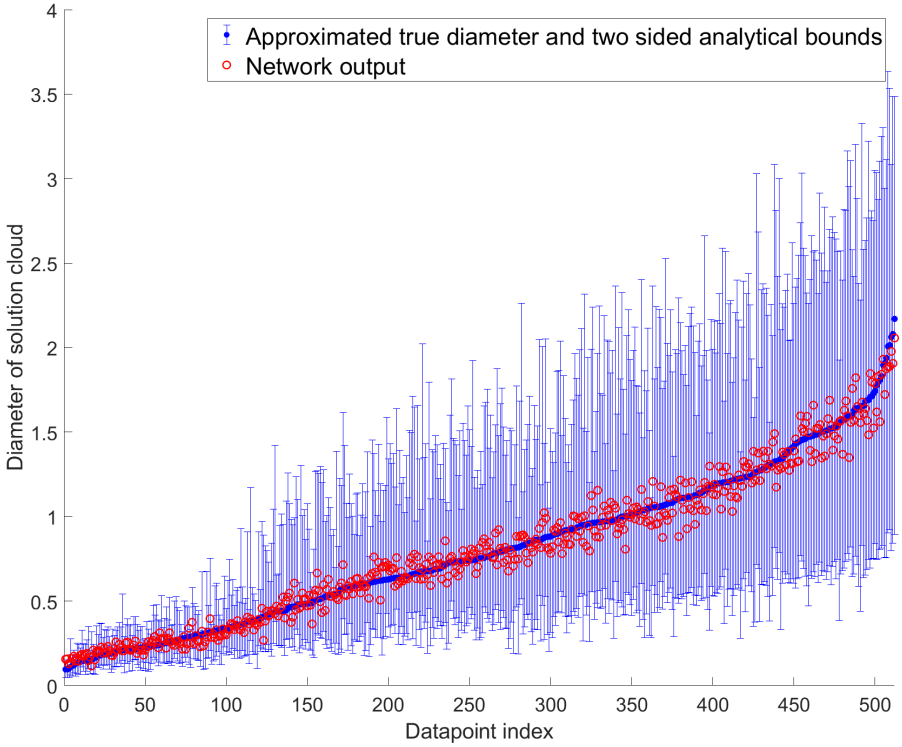


FIGURE 4.9. Comparison of the results obtained analytically and computed by numerical methods and model 1 for 500 test examples similar to training examples.

Similarily we test model 2 with a test set of 300 examples with the constraints (4.4)-(4.6). The results are shown in Figure 4.10.

Model 2 outputs approximations close to the brute-force approximation and mostly between the analytical bounds. However, the result does not look exactly how we would like. Numerical approximations seem lower then expected since they are very close to the lower bounds. This issue is most likely caused by the set approximating the solution cloud being too small.

We will see that this is indeed the case in later tests 2 and 3. Model 2 systematically outputs an approximation of the brute-force method which is lower then sharply calculated brute-force approximations.

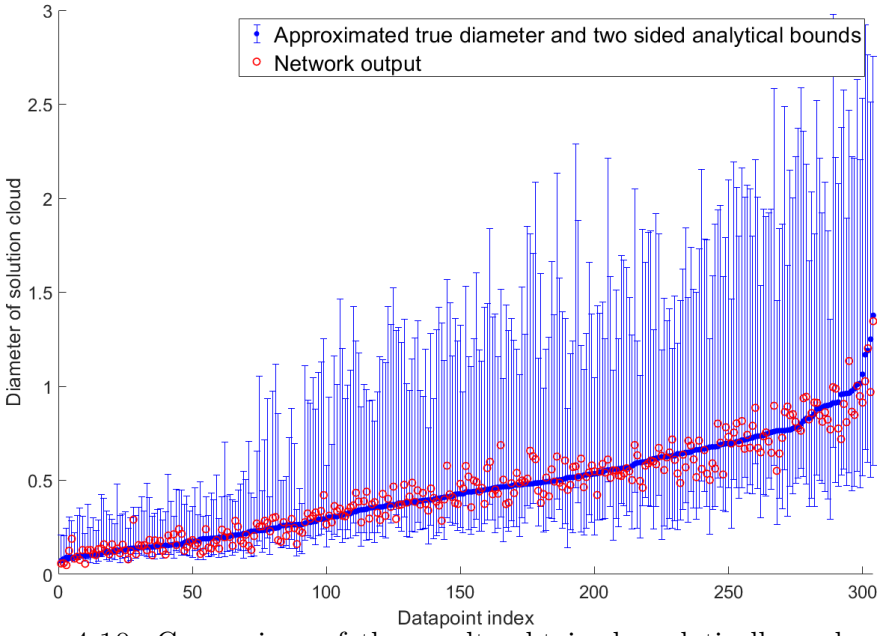


FIGURE 4.10. Comparison of the results obtained analytically and computed by numerical methods and model 2 for 300 test examples similar to training examples.

4.4.2. Test 2: More general data. Next we create more complicated test sets. These sets do not have the piecewise constant constraint (4.1) on the mean coefficients and indeterminacies. In practice, we would like our network to be able to deal with general admissible datasets that may have shapes like the one in Figure 4.2. We still impose the range constraints (4.2) and (4.3).

The way this more general data is created is by picking some random points in the allowed range and fitting a spline curve to these points. Details are explained in Algorithm 3. In this particular test we made a dataset of 150 examples and the important 'spline_point_amount' parameter in Algorithm 3 was set to 3 (see Figure 4.11). The other input parameters are taken from the constraints of Model 1 which are also inside the constraints of Model 2. The 'mean_function_range' parameter is taken from (4.2) and 'max_indeterminacy' from (4.3).

Algorithm 3: Spline admissible data creation

Input:
mean_function_range (float array [a,b]),
max_indeterminacy (float),
spline_point_amount (int),
domain (float array [a,b])

N = spline_point_amount
D = max_indeterminacy
R = mean_function_range
`/* Make the mean function spline first */`
spline_fit_points = rand(1,N)*(R(2)-R(1)) + R(1)
`/* the above variable is the y-axis fit points of the spline */`
domain_points = vector with N equidistant points in the domain
`/* if domain = [0,1] and N = 3 then domain_points = [0.25,0.5,0.75] */`
mean_spline = splinefit(domain_points, spline_fit_points)
`/* Next we make the indeterminacy spline function */`
indeterminacy_spline_points = rand(1,N)*D
indeterminacy_spline = splinefit(domain_points, indeterminacy_spline_points)
`/* Now we can define the lower and upper bound functions which define the admissible strip */`
upperbound_function(x) = mean_spline(x)*(1+indeterminacy_spline(x))
lowerbound_function(x) = mean_spline(x)*(1-indeterminacy_spline(x))

Result: upperbound_function, lowerbound_function

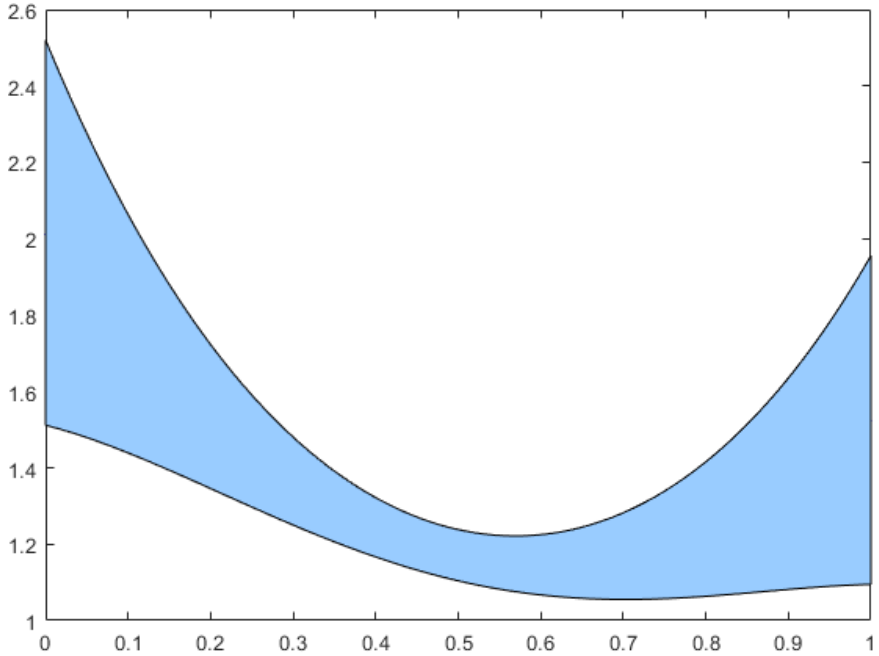


FIGURE 4.11. Example of an output by Algorithm 3 with input: mean_function_range = [1,3], max_indeterminacy = 0.3, spline_point_amount = 3, domain = [0,1]. The area between the lower and upper bound functions is colored.

On this test set the results are for Model 1 are shown in Figure 4.12 and for Model 2 in Figure 4.15. Model 1 performs very well for every datapoint. It is clear that the network outputs outperform the analytical bounds and would be preferable to use in practice.

Model 2 gives a lower approximation of the brute-force method. This showcases the fact that accurate and careful creation of the training data is important. The model has learned to do what it is told to do very well but the training data has not been sufficiently sharp. Regardless of this systematic error the model still outputs fairly reasonable results.

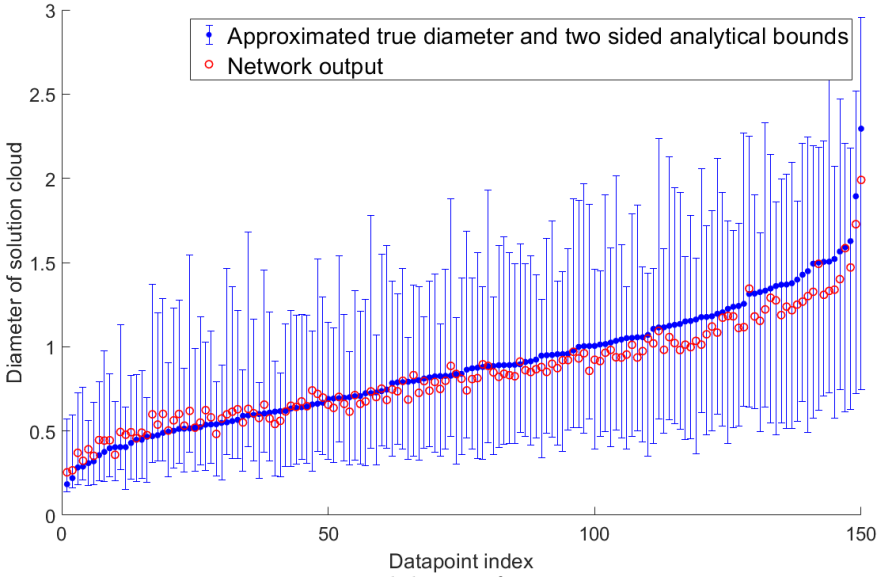


FIGURE 4.12. Model 1 performance in test 2.

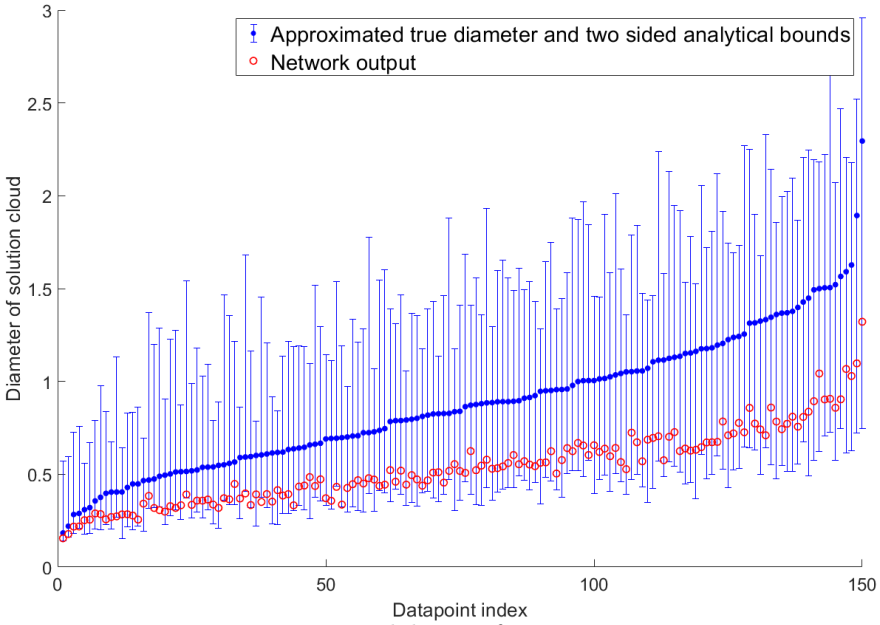


FIGURE 4.13. Model 2 performance in test 2.

4.4.3. Test 3: More general and more oscillating data. Now we use even more oscillating admissible datasets generated by using Algorithm 3 with the same inputs as in test 2 except now 'spline_point_amount' is set to 6. This allows the possible admissible dataset to oscillate much more (see Figure 4.14). The total size of the test set for test 3 is 250 datapoints.

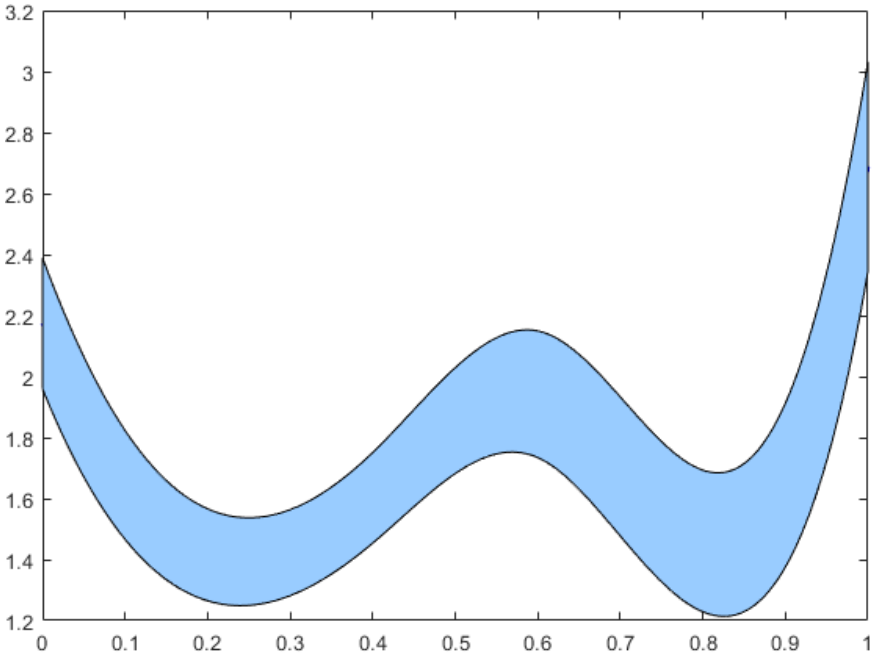


FIGURE 4.14. Example of an output by Algorithm 3 with input: mean_function_range = [1,3], max_indeterminacy = 0.3, spline_point_amount = 6, domain = [0,1]. The area between the lower and upper bound functions is colored.

The results of test 3 are shown for Model 1 in Figure 4.15 and for Model 2 in Figure 4.16. Model 1 still performs surprisingly well even though it was trained on a much less oscillating dataset then the one used in this test. There is clearly more variation from the brute-force approximation then in test 2 so the oscillation is having an effect on performance. Every example still falls between the analytical bounds and it would be quite reasonable to use Model 1 in practice.

Model 2 has the same issue as in the other tests of giving lower approximations. Similarily to Model 1 the variation of distance from the brute-force approximation is higher then in test 2.

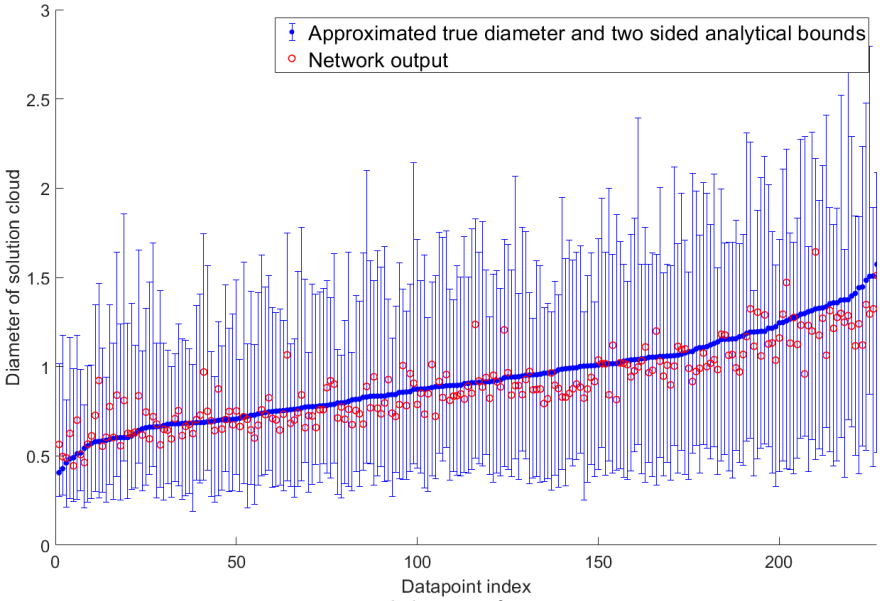


FIGURE 4.15. Model 1 performance in test 3.

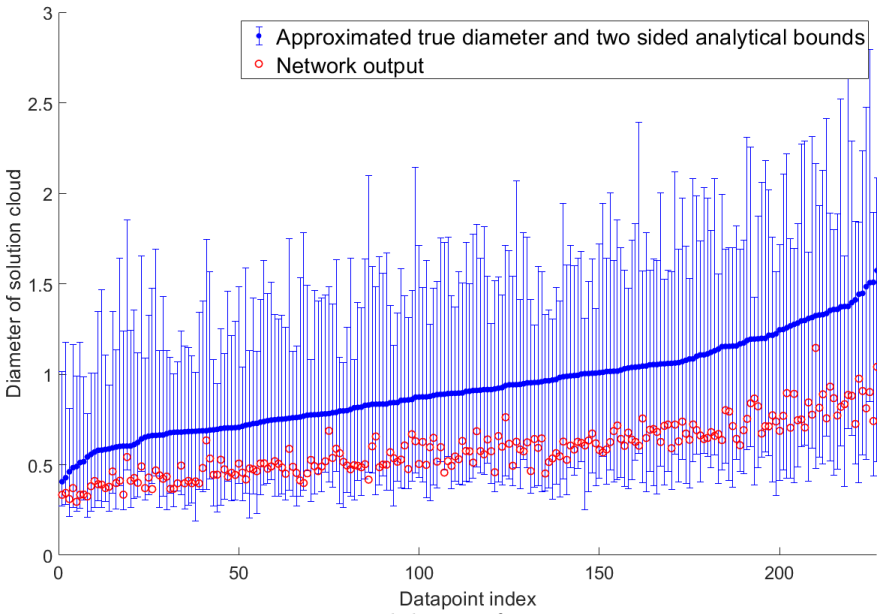


FIGURE 4.16. Model 2 performance in test 3.

Conclusions

Estimation of errors generated by uncertain data is an important problem of mathematical modeling. These errors generate limits of quantitative analysis and must be accounted in real life computations. There are well developed analytical methods based on a posteriori error estimates of the functional type. They provide reliable and computable bounds of uncertainty errors. We have shown new methods for analysis of uncertainty errors based on supervised machine learning methods. Comparisons between ML-methods and analytical methods discussed in section 4.4 show that ML-methods can be quite competitive.

The models shown were not trained to perfection and there is certainly room for improvement if this task were given to an experienced machine learning engineer. Compared to a typical machine learning project this task has the advantage that the data is fairly easy to generate. Various types of training examples with more and more general forms could also be used as training data.

Our results concern only the simplest problems and there is no guarantee that the same approach will work in more complicated scenarios. However, even models working in simple problems may have a use. Checking uncertainty errors even in simple simulations and other computational tasks may be rather difficult or expensive. Error analysis is often ignored in such cases. Machine learning models like the ones we presented are very cheap and easy to use and allow error quantification to be done even with very limited mathematical knowledge and computational power.

In a similar fashion as we made models for quantifying $\mathrm{Diam}(\mathcal{S}(\mathcal{D}))$ one could try to make models for different tasks like quantifying local errors. Current methods for such tasks have the same issues as the ones we have discussed. Analytical methods are coarse and difficult to implement. Numerical methods are computationally expensive.

Efficient and accurate quantification of uncertainty errors will be more and more important in the future when computer simulations become a larger part of performing scientific experiments.

APPENDIX A

Mathematical Background

In this chapter we briefly introduce theories of mathematical analysis that are used in Chapter 2. We will give the definition of Sobolev Spaces in one dimension and look at some fundamental results from calculus of variations. For a detailed look at this theory we recommend [4], [5], [6] or for finnish readers [12].

1.1. Function Spaces

We recall basic function spaces and give the definition of one dimensional Sobolev Spaces. Sobolev Spaces occur naturally when trying to choose a function space for solutions of variational problems. The notation we use for the basic function spaces is as follows:

$L^p(a,b) :=$ Space of measurable functions $f : (a,b) \to \overline{\mathbb{R}}$ which are integrable with power p.

$C^k(a,b) :=$ Space of k times differentiable functions $f : (a,b) \to \mathbb{R}$.

$C_0^k(a,b) :=$ Subspace of $C^k(a,b)$ that contains functions which are zero on the boundary.

For detailed exposition of the above spaces see e.g. [6].

The important space of functions for our analysis in chapter 2 is the one-dimensional Sobolev Space $H^1(a,b)$. Sobolev spaces are motivated by the idea of weakening the definition of derivatives. Classically, the solutions of differential equations of degree k are differentiable up to the degree k. However, there are examples where solutions found from C^k are clearly not satisfactory (see [4], Chapter 0). To define better spaces in which solutions are found, the notion of *weak* or *generalized* derivative is required.

DEFINITION A.1. Let $u \in L^1_{\text{loc}}(a,b)$. We say that $v \in L^1_{\text{loc}}(a,b)$ is the *weak derivative* of u if

$$\int_a^b v(x)\varphi(x)\,dx = -\int_a^b u(x)\varphi'(x)\,dx, \quad \forall\, \varphi \in C_0^\infty(a,b).$$

We use the normal notation of derivatives for the weak derivative so $v = u'$. If such a function v exists we call u a *weakly differentiable function*.

If a function is classically differentiable, the weak and classical derivatives coincide.

Sobolev spaces consist of L^p functions for which weak derivatives exist and are also in L^p. In our case, we only need the case $p = 2$ but these spaces can be defined for any p.

DEFINITION A.2. Let $W^{1,2}(a,b)$ be the set of functions $u : (a,b) \to \mathbb{R},\ u \in L^2(a,b)$ which have weak derivatives such that $u' \in L^2(a,b)$. We endow this space with the norm

$$||u||_{1,2} = (||u||_2^2 + ||u'||_2^2)^{\frac{1}{2}}.$$

We denote $W^{1,2}(a,b)$ also as $H^1(a,b)$ which is a Hilbert space.

The space H^1 contains C^1 but also more general functions such as piecewise differentiable functions.

1.2. Calculus of variations

There are three main things we want to know about a variational problem:

- Does a unique solution exist?
- How regular is the solution if it exists?
- Does the solution satisfy the Euler-Lagrange equation?

An existence theorem is presented and we prove that problem $\mathcal{P}$ (1.1) solves the weak Euler-Lagrange equation. The regularity question is disregarded.

THEOREM A.3. *Existence. Let $F : [a,b] \times \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ be a function such that*

- *F is a Carathéodory function so:*
  $x \to F(x,y,z)$ is measurable for all $(y,z) \in \mathbb{R} \times \mathbb{R}$,
  $(y,z) \to F(x,y,z)$ is continuous for almost every $x \in [a,b]$.
- *$z \to F(x,y,z)$ is convex for almost every $x \in [a,b]$ and every $y \in \mathbb{R}$.*
- *$F(x,y,z) \geq k_1|z|^p + k_2$ for some $k_1 > 0$, $k_2 \in \mathbb{R}$ and some $p > 1$.*

Then there exists a minimizer of the functional $J : X \to \mathbb{R}$

$$J(u) = \int_a^b F(x,u,u')\,dx,$$

where

$$X := \{u \in H^1(a,b) : u(a) = A, u(b) = B\}$$

For a proof of Theorem A.3 see [12] section 3.4.1.

REMARK A.4. Problem $\mathcal{P}$ satisfies the assumptions of Theorem A.3. The function inside the integral in this case is

$$F(x,y,z) = \frac{1}{2}\alpha(x)|z|^2 + \frac{1}{2}\beta(x)|y|^2 + f(x)y,$$

where $\alpha(x)$ and $\beta(x)$ are positive and bounded from above and $f \in L^2(a,b)$. F is measurable function and continuous with respect to y and z so the Carathéodory condition holds.

For the convexity condition we must prove that for any $z_1, z_2 \in \mathbb{R}$ and $t \in [0,1]$

$$F(x,y,tz_1 + (1-t)z_2) \leq tF(x,y,z_1) + (1-t)F(x,y,z_2).$$

It is sufficient to prove that

$$|tz_1 + (1-t)z_2|^2 \leq t|z_1|^2 + (1-t)|z_2|^2. \tag{A.1}$$

Since $t \in [0,1]$ we have

$$\begin{aligned}
& (t^2-t)(z_1-z_2)^2 \leq 0 \\
\Leftrightarrow \quad & (t^2-t)z_1^2 - 2(t^2-t)z_1z_2 + (t^2-t)z_2^2 \leq 0 \\
\Leftrightarrow \quad & (t^2-t)z_1^2 + 2(1-t)tz_1z_2 + (t^2-t)z_2^2 \leq 0 \\
\Leftrightarrow \quad & t^2z_1^2 + z_2^2 - 2tz_2^2 + t^2z_2^2 + 2(1-t)tz_1z_2 \leq tz_1^2 + z_2^2 - tz_2^2 \\
\Leftrightarrow \quad & t^2z_1^2 + (1-t)^2z_2^2 + 2(1-t)tz_1z_2 \leq tz_1^2 + (1-t)z_2^2 \\
\Leftrightarrow \quad & (tz_1 + (1-t)z_2)^2 \leq tz_1^2 + (1-t)z_2^2.
\end{aligned}$$

Finally for the growth condition we can compute

$$\begin{aligned}
F(x,y,z) &= \frac{1}{2}\alpha(x)|z|^2 + \frac{1}{2}\beta(x)|y|^2 + f(x)y \\
&\geq \frac{1}{2}\inf_{x\in[a,b]}\{\alpha(x)\}|z|^2 + \frac{1}{2}\beta(x)|y|^2 + f(x)y.
\end{aligned}$$

We can now choose $k_1 := \frac{1}{2}\inf_{x\in[a,b]}\{\alpha(x)\}$. The terms not dependent on z form a second degree polynomial with a positive constant for the second degree term. This is bounded from below and we can compute the minimum over $y \in \mathbb{R}$ by finding the zero of the derivative

$$\frac{d}{dy}(\frac{1}{2}\beta(x)|y|^2 + f(x)y) = \beta(x)y + f(x).$$

The above derivative has a zero at the point $y = \frac{-f(x)}{\beta(x)}$. Now we can choose the constant k_2 as the infimum over $x \in [a,b]$ so

$$k_2 := \inf_{x\in[a,b]}\left\{\frac{-f(x)}{\beta(x)}\right\}.$$

Now we find

$$F(x,y,z) \geq k_1|z|^2 + k_2,$$

so the growth condition holds (with $p=2$).

THEOREM A.5. *Euler-Lagrange Equation. A solution of problem $\mathcal{P}$ (1.1) solves the Euler-Lagrange equation in the weak sense.*

$$\int_a^b \alpha u'\varphi' + \beta u\varphi + f\varphi \, dx = 0 \quad \forall \varphi \in H_0^1(a,b).$$

PROOF. Assume that u solves problem $\mathcal{P}$ so $\inf_{v\in \mathrm{H}_0^1(a,b)} J(v) = J(u)$. Let $t \neq 0$ and $\varphi \in \mathrm{H}_0^1(a,b)$. Now define

$$\begin{aligned}
G(t) &:= J(u+t\varphi) - J(u) \\
&= \int_a^b \frac{1}{2}\alpha|u'+t\varphi'|^2 + \frac{1}{2}\beta|u+t\varphi|^2 + f(x)(u+t\varphi)\, dx - \int_a^b \frac{1}{2}\alpha|u'|^2 + \frac{1}{2}|u|^2 + fu\, dx \\
&= \int_a^b \frac{1}{2}\alpha(2t\varphi'u' + t^2\varphi'^2) + \frac{1}{2}\beta(2t\varphi u + t^2\varphi^2) + ft\varphi\, dx.
\end{aligned}$$

u is the minimizer of J so $0 \leq G(t)$ for any t. Assume now that $t > 0$. Then

$$\begin{aligned} 0 &\leq \frac{G(t)}{t} \\ &= \int_a^b \frac{1}{2}\alpha(2\varphi' u' + t\varphi'^2) + \frac{1}{2}\beta(2\varphi u + t\varphi) + f\varphi \, dx \\ &= \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi + \frac{t}{2}(\varphi'^2 + \varphi^2) \, dx. \end{aligned}$$

Letting $t \to 0$ since $\varphi \in H_0^1$ dominated convergence gives us

$$\begin{aligned} 0 &\leq \lim_{t\to 0} \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi + \frac{t}{2}(\varphi'^2 + \varphi^2) \, dx \\ &= \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi + \lim_{t\to 0} \frac{t}{2}(\varphi'^2 + \varphi^2) \, dx \\ &= \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi \, dx. \end{aligned} \tag{A.2}$$

Similarily setting $t < 0$ we have

$$\begin{aligned} 0 &\geq \frac{G(t)}{t} \\ &= \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi + \frac{t}{2}(\varphi'^2 + \varphi^2) \, dx. \end{aligned}$$

And letting $t \to 0$ we have

$$0 \geq \int_a^b \alpha\varphi' u' + \beta\varphi u + f\varphi \, dx. \tag{A.3}$$

Combining (A.2) and (A.3) the proof is completed. □

1.3. Inequalities

THEOREM A.6. *Let $L(x, y)$ be a functional defined on two nonempty sets of elements X and Y. Then*

$$\sup_{y\in Y} \inf_{x\in X} L(x, y) \leq \inf_{x\in X} \sup_{y\in Y} L(x, y).$$

PROOF. Clearly

$$L(x, y) \geq \inf_{k\in X} L(k, y), \quad \forall x \in X.$$

Taking the supremum of $y \in Y$ we have

$$\sup_{y\in Y} L(x, y) \geq \sup_{y\in Y} \inf_{k\in X} L(k, y), \quad \forall x \in X.$$

Since this holds for any x we take the infimum over $x \in X$ and arrive at

$$\inf_{x\in X} \sup_{y\in Y} L(x, y) \geq \sup_{y\in Y} \inf_{k\in X} L(k, y).$$

□

THEOREM A.7. *Let $v \in H_0^1(0,1)$. Then*

$$||v||_2 \leq C||v'||_2,$$

where $C \geq \frac{1}{\pi}$.

PROOF. Rearranging the inequality we are looking for a constant C, such that

$$\frac{1}{C^2} = \inf_{v \in H_0^1(0,1)} \left\{ \frac{||v'||_2^2}{||v||_2^2} \right\}.$$

Assume that $u \in H_0^1(0,1)$ minimizes the functional $\mathcal{R} : H_0^1(0,1) \to \mathbb{R}$,

$$m := \mathcal{R}(u) = \inf_{v \in H_0^1(0,1)} \mathcal{R}(v), \quad \mathcal{R}(v) := \frac{||v'||_2^2}{||v||_2^2}.$$

The functional $\mathcal{R}$ is known as the *Rayleigh quotient*. Now for any $\varepsilon > 0$ and $h \in H_0^1(0,1)$

$$\mathcal{R}(u) \leq \mathcal{R}(u + \varepsilon h) =: G(\varepsilon).$$

Since u is the minimizer of $\mathcal{R}$ we must have $G'(0) = 0$ (since it is a critical point), so

$$\begin{aligned}
&\lim_{\varepsilon \to 0} \frac{\mathcal{R}(u+\varepsilon h) - \mathcal{R}(u)}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{(\int_0^1 |u' + \varepsilon h'|^2)/(\int_0^1 |u + \varepsilon h|^2) - (\int_0^1 |u'|^2)/(\int_0^1 |u|^2)}{\varepsilon} \\
&= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\frac{\int_0^1 (u')^2 + 2u'h'\varepsilon + (\varepsilon h')^2}{\int_0^1 u^2 + 2\varepsilon uh + (\varepsilon h)^2} - \frac{\int_0^1 |u'|^2}{\int_0^1 |u|^2} \right) \\
&= \lim_{\varepsilon \to 0} \frac{1}{\varepsilon} \left(\frac{\left(\int_0^1 (u')^2 + 2u'h'\varepsilon + (\varepsilon h')^2\right)\left(\int_0^1 |u|^2\right) - \left(\int_0^1 u^2 + 2\varepsilon uh + (\varepsilon h)^2\right)\left(\int_0^1 |u'|^2\right)}{\left(\int_0^1 u^2 + 2\varepsilon uh + (\varepsilon h)^2\right)\left(\int_0^1 |u|^2\right)} \right) \\
&= \lim_{\varepsilon \to 0} \left(\frac{\left(\int_0^1 2u'h' + \varepsilon (h')^2\right)\left(\int_0^1 |u|^2\right) - \left(\int_0^1 2uh + \varepsilon (h)^2\right)\left(\int_0^1 |u'|^2\right)}{\left(\int_0^1 u^2 + 2\varepsilon uh + (\varepsilon h)^2\right)\left(\int_0^1 |u|^2\right)} \right) \\
&= \frac{\left(\int_0^1 2u'h'\right)\left(\int_0^1 |u|^2\right) - \left(\int_0^1 2uh\right)\left(\int_0^1 |u'|^2\right)}{\left(\int_0^1 |u|^2\right)^2} = 0.
\end{aligned}$$

This holds if, and only if

$$\left(\int_0^1 u'h'\right)\left(\int_0^1 |u|^2\right) - \left(\int_0^1 uh\right)\left(\int_0^1 |u'|^2\right) = 0.$$

Rearranging this we have

$$\frac{\int_0^1 |u'|^2}{\int_0^1 |u|^2} = \frac{\int_0^1 u'h'}{\int_0^1 uh}.$$

The left hand side is equal to the minimum m. This gives us

$$m\int_0^1 uh = \int_0^1 u'h'.$$

Rearranging we find

$$\int_0^1 u'h' = -\int_0^1 -muh.$$

The above equality defines the weak derivative $u'' = -mu$. Since $-mu$ is in fact continuous (since it is in H^1) u'' can be considered as a classical derivative. This means that the minimum m can only be the minimizer if it is an eigenvalue of the boundary value problem

$$\begin{aligned} u'' &= -mu, \\ u(0) &= 0, \\ u(1) &= 0. \end{aligned} \tag{A.4}$$

The solution of (A.4) has the form

$$u = \sin(\sqrt{m}x).$$

In order for the boundary conditions to hold we must have $m = (k\pi)^2$, $k \in \mathbb{N} \setminus \{0\}$. Since we want the minimizer we can choose $k = 1$ and we find

$$\pi^2 \geq \frac{1}{C^2} \quad \Rightarrow \quad C \geq \frac{1}{\pi}.$$

□

From Theorem A.7 by a simple change of variables we obtain the next result.

THEOREM A.8. *Let $v \in H_0^1(a,b)$. Then*

$$||v||_2 \leq \overline{C}_F ||v||_{\alpha,\beta},$$

where $||\cdot||_{\alpha,\beta}$ is the energy norm from (1.5) and

$$\overline{C}_F := \frac{b-a}{\pi}(\operatorname{ess\,sup}\{\alpha(x)^{-1}\})^{\frac{1}{2}}.$$

This inequality is called the Friedrich's inequality.

Bibliography

[1] Mali, Neittaanmäki, Repin: *Accuracy Verification Methods*, Springer, 2014.

[2] Repin: *A Posteriori Estimates for Partial Differential Equations*, De Gruyter, 2008.

[3] Goodfellow-et-al-2016: *Deep Learning*, MIT Press, 2016, `http://www.deeplearningbook.org`.

[4] Dacorogna: *Introduction to the Calculus of Variations*, Imperial College Press, 2004.

[5] Buttazzo, Giaquinta, Hildebrandt: *One-dimensional Variational Problems : An Introduction*, Clarendon Press, 1998.

[6] Evans: *Partial Differential Equations*, American Mathematical Society, 2010.

[7] Tuominen, Neittaanmäki: *Tekoälyn perusteita ja sovelluksia*, Jyväskylän yliopisto, 2019, `http://urn.fi/URN:ISBN:978-951-39-7796-2`

[8] G.I. Schueller: *A state-of-the-art report on computational stochastic mechanics*, Probab. Eng. Mech. 1997.

[9] I. Hlavaecek, J. Chleboun, and I. Babueska: *Uncertain input data problems and the worst scenario method*, Elsevier, Amsterdam, 2004.

[10] Neittaanmäki, Repin: *Reliable methods for computer simulation*, Elsevier, 2004.

[11] P. G. Ciarlet: *The Finite Element Method for Elliptic Problems*, North Holland, 1978.

[12] Dacorogna: *Direct Methods in the Calculus of Variations* Springer, 1989.