Chapter 1
Introduction

Remark 1.1. On the importance of scalar convection-dominated problems and the fundamental problem in their numerical simulation. Scalar convection-dominated problems arise in many models of processes from nature and technical applications, e.g., when modeling mass and energy balances. This situation will be always happen if a species (dissolved or as particle) or a physical quantity, e.g., the temperature, is transported by a flow field. Often, further processes besides the flow field are present in real systems, like chemical reactions.

Solutions of scalar convection-dominated problems possess generally very small structures or scales. These scales are important features of the solution. Numerical methods for the simulation of scalar convection-dominated problems are based in general on a decomposition of the given domain with grids. These grids are often not sufficiently fine to resolve the small scales of the solution, i.e., the small scales cannot be even represented on the given grids. In this situation, it turns out that standard numerical methods fail completely to produce useful approximations of the solutions. Special numerical methods are necessary. In this course, the most important special methods will be introduced and discussed. □

Remark 1.2. Contents of the course, state of the art of the research. This course studies properties of the solution of scalar convection-dominated problems and their numerical approximation. It will start with investigations of problems in one dimension, since

- the main difficulties arise already in one dimension,
- the analysis for one-dimensional problems is comparable simple,
- one can construct for one-dimensional problems numerical methods that produce, in a certain sense, perfect numerical solutions.

Higher-dimensional problems will be studied in the second part of the course. The basic approach consists in trying to transfer the ideas for one-dimensional problems to higher dimensions. It turns out that

- one obtains much better solutions compared with the numerical solutions computed with standard methods,
- however, the solutions are by far not perfect, often they are not even good, e.g., see Augustin *et al.* (2011); John & Schumacher (2014).

Methods for the computation of good numerical solutions for scalar convection-dominated problems in higher dimensions are an active field of research. □

Remark 1.3. Literature. The best monograph on this topic is Roos *et al.* (2008). The course follows in large parts this monograph. A shorter survey paper that can be recommended is Stynes (2005). □

Chapter 2

Analysis of Two-Point Boundary Value Problems

2.1 The Model Problem

Definition 2.1. Linear two-point boundary value problem. A linear two-point boundary value problem has the form

$$-\varepsilon u'' + b(x)u' + c(x)u = f(x), \quad \text{for } x \in (d,e), \tag{2.1}$$

with the boundary conditions

$$\begin{aligned} \alpha_d u(d) - \beta_d u'(d) &= \gamma_d, \\ \alpha_e u(e) - \beta_e u'(e) &= \gamma_e. \end{aligned} \tag{2.2}$$

Here are $b, c, f \in C([e,d])$, $0 < \varepsilon \in \mathbb{R}$, and the constants $\alpha_d, \alpha_e, \beta_d, \beta_e, \gamma_d, \gamma_e in\mathbb{R}$ are given. □

Remark 2.2. Importance of two-point boundary value problems. The boundary value problem (2.1), (2.2) is the simplest model for processes that have diffusion and transport.

An example from Goering (1977) is as follows. Consider a flow reactor with constant temperature in which there is a continuous inflow of a species (reactant) and an outflow of a product. Then, the concentration $c(t,x,y,z) = c(t,\boldsymbol{x})$, [$^{\text{kmol}}/_{\text{m}^3}$], of the species in the reactor is the solution of the following partial differential equation

$$\underbrace{\frac{\partial c}{\partial t} + \operatorname{div}(c\boldsymbol{u})}_{\text{convection}} \underbrace{-\operatorname{div}(D \operatorname{grad} c)}_{\text{diffusion}} = r(c),$$

where $\boldsymbol{u}(t,\boldsymbol{x})$, [$^{\text{m}}/_{\text{s}}$], is the velocity, $r(c(t,\boldsymbol{x}))$, [$^{\text{kmol}}/_{\text{m}^3\text{s}}$], is a function which models the reaction, and $D(t,\boldsymbol{x})$, $^{\text{m}^2}/_{\text{s}}$], is the diffusion coefficient. If the reactor works stationary, i.e., the temporal changes are very slow and they are negligible, if the parameters D and $\boldsymbol{u}$ are constant, and if the concentration

Fig. 2.1 Jean Claude Eugene Péclet (1793 – 1857).

changes only in x-direction, then one obtains an ordinary differential equation for $c(x)$

$$-Dc''(x) + uc'(x) = r(c(x)). \tag{2.3}$$

Numerical simulations are always based on dimensionless equations. Let $x \in [0, L]$, where L, [m], is the length of the reactor. With the dimensionless quantities

$$\xi := \frac{x}{L}, \quad \gamma := \frac{c}{c_0},$$

where c_0, [kmol/m^3], is a constant reference concentration, one gets a dimensionless ordinary differential equation. Using the chain rule gives

$$\frac{d\gamma(\xi)}{d\xi} = \frac{d(c(x)/c_0)}{dx}\frac{dx}{d\xi} = L\frac{c'(x)}{c_0} \quad \text{and} \quad \frac{d^2\gamma(\xi)}{d\xi^2} = L^2\frac{c''(x)}{c_0}.$$

Inserting these expressions in the differential equation (2.3) yields, if $u \neq 0$,

$$-\frac{1}{\mathrm{Pe}}\gamma''(\xi) + \gamma'(\xi) = \rho(\gamma(\xi)),\ \xi \in (0,1), \quad \text{with} \quad \mathrm{Pe} := \frac{uL}{D},\ \rho = \frac{L}{uc_0}r.$$

The dimensionless number Pe is called Péclet number. For completing the problem, appropriate (dimensionless) boundary conditions at $\xi \in \{0, 1\}$ have to be prescribed. □

Remark 2.3. Names for the individual terms in (2.1). Based on the application described in Remark 2.2, the terms in (2.1) are called as follows:

- $-\varepsilon u''$ – diffusive term,
- $b(x)u'$ – convective term, advective term, transport term,
- $c(x)u$ – reactive term.

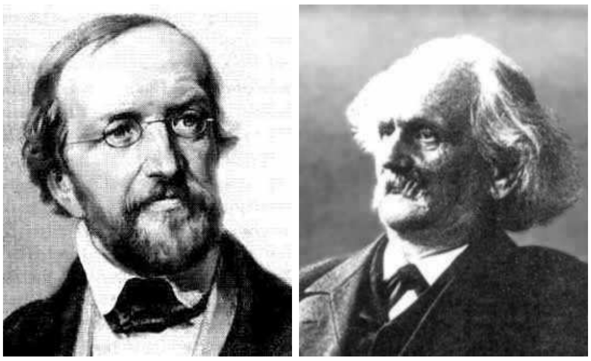

Fig. 2.2 Left: Johann Peter Gustav Lejeune Dirichlet (1805 – 1859), right: Carl Gottfried Neumann (1832 – 1925).

The model problem (2.1), (2.2) is called convection–diffusion(–reaction) problem, if $b(x) \not\equiv 0$.

The Péclet number describes the ratio of convection and diffusion. If this ratio is high, then the numerical solution of (2.1), (2.2) might be difficult. □

Definition 2.4. Boundary conditions. Let $\gamma_d, \gamma_e \in \mathbb{R}$, $\alpha_d, \alpha_e \in \mathbb{R} \setminus \{0\}$. Boundary conditions of the form:

1. $$u(d) = \gamma_d, \quad u(e) = \gamma_e$$
   are called boundary conditions of first kind or Dirichlet boundary conditions,
2. $$u'(d) = \gamma_d, \quad u'(e) = \gamma_e$$
   are called boundary conditions of second kind or Neumann boundary conditions,
3. $$\alpha_d u(d) + u'(d) = \gamma_d, \quad \alpha_e u(e) + u'(e) = \gamma_e$$
   are called boundary conditions of third kind or Robin[1] boundary conditions.

Dirichlet boundary conditions are

- most important in applications,
- best understood from the point of view of the analysis.

In this course, it will be concentrated on Dirichlet boundary conditions. □

[1] Victor Gustave Robin (1855 – 1897)

Remark 2.5. Normalization of a linear two-point boundary value problem.

- One can assume, without loss of generality, that $x \in (0,1)$. The original interval (d,e) is mapped to $(0,1)$ with the transform
$$x \mapsto \frac{x-d}{e-d}.$$
- One can also assume, without loss of generality, that homogeneous boundary conditions $\gamma_d = \gamma_e = 0$ are prescribed. To this end, one subtracts from $u(x)$ a smooth function $\psi(x)$ which satisfies the original boundary conditions. If, e.g., the original Dirichlet boundary conditions are given by
$$u(d) = \gamma_d, \quad u(e) = \gamma_e,$$
then one can set
$$\psi(x) = \gamma_d \frac{x-e}{d-e} + \gamma_e \frac{x-d}{e-d}$$
and
$$u^*(x) = u(x) - \psi(x).$$
It follows that $u^*(x)$ is the solution of a two-point boundary value problem with homogeneous Dirichlet boundary conditions.

□

Definition 2.6. Model problem. The model problem has the form

$$Lu := -\varepsilon u'' + b(x)u' + c(x)u = f(x) \quad \text{for } x \in (0,1), \tag{2.4}$$

with the boundary conditions

$$u(0) = u(1) = 0. \tag{2.5}$$

For the data of the problem it is assumed that $b, c, f \in C([0,1])$ and $0 < \varepsilon \in \mathbb{R}$. □

Remark 2.7. Differential operator. In (2.4), L denotes a differential operator. An operator is a map between two function spaces. A linear operator is a linear map A, defined on a linear space X, such that

$$A(\alpha u + \beta v) = \alpha A u + \beta A v$$

for all scalars α, β and all $u, v \in X$. A differential operator is an operator, which, if it is applied to appropriate functions, contains derivatives of these function. For the complete definition of an operator, its domain has to be given. □

Example 2.8. Standard example. The boundary value problem

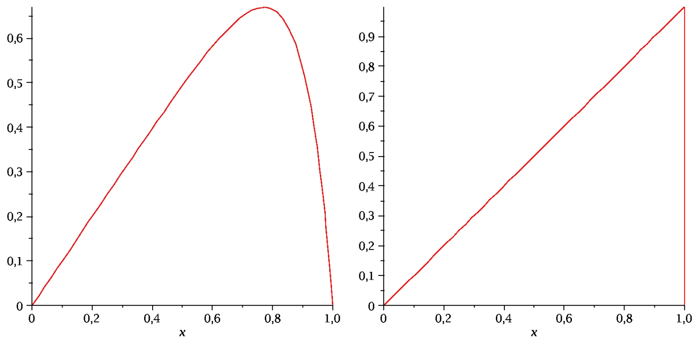

Fig. 2.3 Solution of Example 2.8 for $\varepsilon = 0.1$ (left) and $\varepsilon = 0.0001$ (right).

$$-\varepsilon u'' + u' = 1 \quad \text{on } (0,1), \quad u(0) = u(1) = 0$$

has the solution

$$u(x) = x - \frac{\exp\left(-\frac{1-x}{\varepsilon}\right) - \exp\left(-\frac{1}{\varepsilon}\right)}{1 - \exp\left(-\frac{1}{\varepsilon}\right)}.$$

The smaller the parameter ε, the steeper becomes the solution at the right boundary, see Figure 2.3. This part of the solution is called (boundary) layer. Such strong changes of the solutions in a very small part of the domain lead to difficulties in the numerical approximation of the solution. □

Remark 2.9. Transform of the model problem to a symmetric problem. Let $b(x)$ be sufficiently smooth. If one defines

$$\tilde{u}(x) := u(x) \exp\left(-\frac{1}{2\varepsilon}\int_0^x b(\xi)\, d\xi\right), \quad x \in [0,1], \tag{2.6}$$

then one can transform (2.4), (2.5) to the symmetric problem

$$-\varepsilon \tilde{u}''(x) + \tilde{c}(x)\tilde{u}(x) = \tilde{f}(x), \; x \in (0,1), \quad \tilde{u}(0) = \tilde{u}(1) = 0,$$

where

$$\tilde{c}(x) := \frac{1}{4\varepsilon} b^2(x) - \frac{1}{2} b'(x) + c(x), \quad \tilde{f}(x) := f(x) \exp\left(-\frac{1}{2\varepsilon}\int_0^x b(\xi)\, d\xi\right),$$

exercise. The weak or variational formulation of this problem, see Section 4.1, contains only symmetric bilinear forms. □

Definition 2.10. Reduced problem, reduced solution. The reduced problem is obtain by setting formally $\varepsilon = 0$, yielding

$$L_0 u_0 := b(x)u_0' + c(x)u_0 = f(x), \quad \text{for } x \in (0,1).$$

The Dirichlet boundary condition has to be set at the boundary where the convection comes from, i.e., where the inflow is situated. In the case $b(x) > 0$ for all $x \in [0,1]$, the problem has the boundary condition

$$u_0(0) = 0,$$

and for $b(x) < 0$ for all $x \in [0,1]$, the boundary condition is

$$u_0(1) = 0.$$

The solution of the reduced problem is called reduced solution. □

Example 2.11. Reduced problem for Example 2.8. The reduced problem for Example 2.8 has the form

$$u_0' = 1 \quad \text{in } (0,1), \quad u_0(0) = 0.$$

Its solution is $u_0(x) = x$.

It follows that the solution of the non-reduced problem from Example 2.8 is the sum of the solution of the reduced problem and another term, which is responsible for the fulfillment of the second boundary condition. □

2.2 Existence and Uniqueness of the Solution of the Model Problem

Remark 2.12. The model problem. For the investigation of the unique solvability of the two-point boundary value problem (2.4), (2.5) is the value of $\varepsilon > 0$ not important. After having divided the equation by ε and having renamed the data, one can consider the problem

$$Lu := -u''(x) + b(x)u'(x) + c(x)u(x) = f(x), \quad \text{for } x \in (0,1), \tag{2.7}$$

with the boundary conditions

$$u(0) = u(1) = 0. \tag{2.8}$$

□

Definition 2.13. Classical solution. A function $u(x)$ is called classical solution of (2.7), (2.8), if

- $u \in C^2(0,1) \cap C([0,1])$,
- $u(x)$ satisfies (2.7) identically,
- $u(x)$ satisfies the boundary conditions (2.8).

□

2.2.1 Investigation of the Differential Equation (2.7)

Remark 2.14. Contents of this section. A classical solution of (2.7) has to satisfy the first two properties of Definition 2.13. This topic is studied in this section. The necessary tools are known already from the theory of ordinary differential equations, e.g., see Numerical Mathematics I. □

Definition 2.15. Linearly independent functions. Two functions $u_1(x)$ and $u_2(x)$, both defined on the interval (a,b), are called linearly independent if from

$$c_1u_1(x) + c_2u_2(x) = 0 \quad \text{for all } x \in (a,b),$$

it follows that $c_1 = c_2 = 0$. They are called linearly dependent, if they are not linearly independent. □

Remark 2.16. Wronski determinant. If two linearly dependent functions, which are defined on (a,b), are continuously differentiable, then it follows from the condition for the linear dependence also that

$$c_1u_1'(x) + c_2u_2'(x) = 0 \quad \text{for all } x \in (a,b).$$

Hence, with $u_1(x), u_2(x)$ there are also $u_1'(x), u_2'(x)$ linearly dependent. It follows that the homogeneous linear system of equations

$$\begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

has a nontrivial solution for all $x \in (a,b)$. Thus, the so-called Wronski determinant

$$W(x) := \det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{pmatrix}$$

vanishes for all $x \in (a,b)$. Note that the reverse statement is not true: the Wronski determinant might vanish for all $x \in (a,b)$ but the two functions are linearly independent. One can find an example for this case in Emmrich (2004). □

Lemma 2.17. Linear independence of two solutions of the homogeneous differential equation. *Let* $x_0 \in (0,1)$ *be arbitrary. Two classical solutions, defined on* $(0,1)$*, of the homogeneous second order linear differential equation with continuous coefficients* (2.7) *are linearly independent if and only if the corresponding Wronski determinant does not vanish for* x_0*.*

Proof. The proof should be known from the course on the theory of ordinary differential equations or from Numerical Mathematics I. It is given just for completeness of presentation.

Fig. 2.4 Left: Joseph Marie Wronski (1758 – 1853), right: Joseph Liouville (1809 – 1882).

i) $W(x_0) \neq 0 \implies$ *linear independence.* Let $u_1(x)$ and $u_2(x)$ be classical solutions of

$$-u''(x) + b(x)u'(x) + c(x)u(x) = 0, \quad x \in (0,1),$$

where $b, c \in C([0,1])$. One obtains for the Wronski determinant, applying the product rule,

$$\begin{aligned} W'(x) &= \big(u_1(x)u_2'(x) - u_1'(x)u_2(x)\big)' \\ &= u_1'(x)u_2'(x) + u_1(x)u_2''(x) - u_1''(x)u_2(x) - u_1'(x)u_2'(x) \\ &= u_1(x)u_2''(x) - u_1''(x)u_2(x) \\ &= u_1(x)\,\big(b(x)u_2'x(x) + c(x)u_2(x)\big) - u_2(x)\,\big(b(x)u_1'x(x) + c(x)u_1(x)\big) \\ &= b(x)\,\big(u_1(x)u_2'x(x) - u_1'x(x)u_2(x)\big) + c(x)\,(u_1(x)u_2(x) - u_1(x)u_2(x)) \\ &= b(x)W(x). \end{aligned}$$

Hence, the Wronski determinant solves the homogeneous first order linear differential equation

$$y'(x) = b(x)y(x).$$

The general solution of this differential equation is given by the Liouville formula. Applying this formula to the differential equation for $W(x)$, then one gets for every $x_0 \in (0,1)$ that

$$W(x) = W(x_0)\exp\left(\int_{x_0}^{x} b(\xi)\ d\xi\right), \quad x \in (0,1).$$

Since the exponential takes only positive values, it follows that the Wronski determinant is (not) equal to zero if and only if it is (not) equal to zero at an arbitrary point $x_0 \in (0,1)$. In particular, one has with Remark 2.16 that in the case $W(x_0) \neq 0$ the functions $u_1(x)$ and $u_2(x)$ are linearly independent.

ii) Linear independence $\implies W(x_0) \neq 0$. This statement is proved by contradiction. Assume that $u_1(x)$ and $u_2(x)$ are two linearly independent solutions and the Wronski determinant vanishes in $x_0 \in (0,1)$. Then, it follows from part i) that it vanishes even in the whole interval $(0,1)$. Hence, there is a nontrivial solution of the linear system of equations

$$\begin{pmatrix} u_1(x) & u_2(x) \\ u_1'(x) & u_2'(x) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

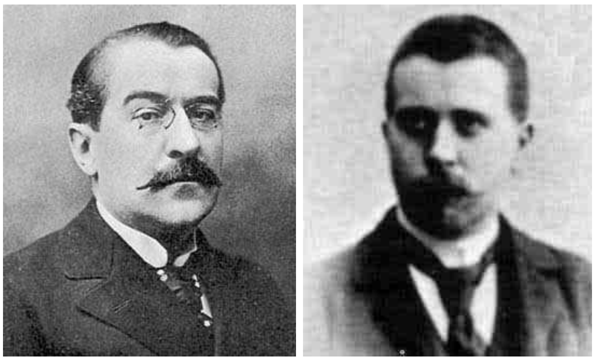

Fig. 2.5 Left: Charles Emile Picard (1856 – 1941), right: Ernst Leonard Lindelöf (1870 – 1946).

Consider the function

$$v(x) := c_1u_1(x) + c_2u_2(x).$$

By the assumption, it is $v(x_0) = v'(x_0) = 0$. The function $v(x)$ is a solution of the differential equation, because this equation is linear. Hence, $v(x)$ solves the following initial value problem

$$-v''(x) + b(x)v'(x) + c(x)v(x) = 0, \quad x \in (x_0, 1), \quad v(x_0) = v'(x_0) = 0$$

for every $x_0 \in (0, 1)$. The application of the Theorem of Picard–Lindelöf shows that this initial value problem has only the trivial solution. Hence, one gets that $v(x) = 0$ for all $x \in (0, 1)$. This result contradicts the linear independence of $u_1(x)$ and $u_2(x)$. ■

Theorem 2.18. Super position principle. *Consider the homogeneous linear differential equation*

$$-u'' + b(x)u' + c(x)u = 0, \quad x \in (0, 1),$$

with coefficients $b, c \in C([0, 1])$. Then, there are two linearly independent solutions in $C^2([0, 1])$ and every classical solution can be represented as a linear combination of them.

Proof. The proof should be known from the course on ordinary differential equations or Numerical Mathematics I. It is presented here just for completeness.

i) Existence and uniqueness of the solution of the initial value problem. One can rewrite the second order differential equation (2.7) as an equivalent system of first order differential equations

$$\frac{d}{dx}\begin{pmatrix} u(x) \\ u'(x) \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ c(x) & b(x) \end{pmatrix} \begin{pmatrix} u(x) \\ u'(x) \end{pmatrix}.$$

From the Theorem of Picard–Lindelöf, it follows that for given initial conditions $u(x_0)$, $u'(x_0)$, $x_0 \in (0, 1)$, there is a uniquely determined solution $(u(x), u'(x))$ of the initial value problem in $[0, 1]$. Both components of the solution are in $C^1([0, 1])$, from which $u \in C^2([0, 1])$ follows.

ii) Existence of two linearly independent solutions. Let $u_1(x)$ be the solution with the initial conditions $u(x_0) = 1, u'(x_0) = 0$ and $u_2(x)$ be the solution with the initial conditions $u(x_0) = 0, u'(x_0) = 1$. One obtains for the Wronski determinant that

$$W(x_0) = u_1(x_0)u_2'(x_0) - u_1'(x_0)u_2(x_0) = 1.$$

With Lemma 2.17, it follows that $u_1(x)$ and $u_2(x)$ are linearly independent.

iii) Representation of each classical solution as a linear combination. Each linear combination

$$u(x) = c_1u_1(x) + c_2u_2(x), \quad c_1, c_2 \in \mathbb{R},$$

is a solution of the differential equation. Consider a function of the form

$$u(x) = au_1(x) + bu_2(x), \quad x \in [0,1],\ a, b \in \mathbb{R}.$$

This function satisfies the initial value problem with the initial conditions a, b, where a, b are two arbitrary real numbers. In this way, all possible combinations of initial data are covered. Hence, it follows that every solution of the differential equation can be represented in this form. ■

Theorem 2.19. Classical solution of the inhomogeneous differential equation. *Consider the inhomogeneous linear differential equation*

$$-u'' + b(x)u' + c(x)u = f(x), \quad x \in (0,1),$$

with $b, c, f \in C([0,1])$. Then, there exists a classical solution $u_p(x)$, the so-called particular solution, and every classical solution can be represented in the form

$$u(x) = c_1u_1(x) + c_2u_2(x) + u_p(x), \quad c_1, c_2 \in \mathbb{R},$$

where $\{u_1(x), u_2(x)\}$ is a system of two linearly independent solutions (fundamental system) of the corresponding homogeneous differential equation. It is $u \in C^2([0,1])$.

Proof. The proof of this theorem follows from the global (existence and uniqueness) Theorem of Picard–Lindelöf, *exercise.* ■

2.2.2 Investigation of the Two-Point Boundary Value Problem (2.7)*,* (2.8)

Example 2.20. Non-uniqueness of the solution of the two-point Dirichlet boundary value problem. Consider the differential equation

$$-u''(x) - u(x) = 0.$$

The general solution of this linear homogeneous differential equation has the form

$$u(x) = c_1 \cos x + c_2 \sin x, \quad c_1, c_2 \in \mathbb{R}.$$

- Let the boundary conditions be given by

$$u(0) = u(\pi/2) = 1,$$

then there is the unique solution $u(x) = \cos x + \sin x$.

- If the boundary conditions are prescribed by

$$u(0) = u(\pi) = 1,$$

then there is no solution of the boundary value problem, since one gets that at the same time $c_1 = 1$ and $c_1 = -1$ have to hold.

- With the boundary conditions

$$u(0) = 1, \quad u(\pi) = -1,$$

there are infinitely many solution, because from the boundary conditions one gets only that $c_1 = 1$. The value for c_2 can be chosen arbitrarily.

This example shows that even in simple cases there might be not a unique solution of the two-point boundary value problem (2.7), (2.8). The problem is not well-posed. It will be shown in the following that the coefficients of this problem have to satisfy certain conditions such that a unique solution is guaranteed. □

Theorem 2.21. Existence and uniqueness of a solution of the model problem with homogeneous right-hand side. *Consider the two-point boundary value problem* (2.7), (2.8) *with* $b \in C^1([0,1])$, $c \in C([0,1])$, *and* $f(x) \equiv 0$. *If for all* $x \in (0,1)$

$$\tilde{c}(x) := \frac{1}{4}b^2(x) - \frac{1}{2}b'(x) + c(x) \geq 0, \tag{2.9}$$

then (2.7), (2.8) *has only the trivial solution.*

Proof. It is obvious that $u(x) \equiv 0$ is a solution of the considered problem.

Proof by contradiction. Assume that $u(x) \not\equiv 0$ is another classical solution. From Theorem 2.19, it follows that $u \in C^2([0,1])$. Applying the transform from Remark 2.9, one gets the problem

$$-\tilde{u}''(x) + \tilde{c}(x)\tilde{u}(x) = 0, \quad x \in (0,1),\ \tilde{u}(0) = \tilde{u}(1) = 0. \tag{2.10}$$

One solution of this problem is $\tilde{u}(x) \equiv 0$. Let $\tilde{u}(x)$ be another solution. Multiplying (2.10) with the second solution, integrating in $(0,1)$, and applying integration by parts yields

$$\begin{aligned} 0 &= \int_0^1 \left(-\tilde{u}''(x)\tilde{u}(x) + \tilde{c}(x)\tilde{u}^2(x)\right)\, dx \\ &= -\tilde{u}''(1)\tilde{u}(1) + \tilde{u}''(0)\tilde{u}(0) + \int_0^1 \Big(\left(\tilde{u}'(x)\right)^2 + \tilde{c}(x)\tilde{u}^2(x)\Big)\, dx \\ &= \int_0^1 \Big(\left(\tilde{u}'(x)\right)^2 + \tilde{c}(x)\tilde{u}^2(x)\Big)\, dx, \end{aligned}$$

since $\tilde{u}(x)$ vanishes at the boundary. Since $\tilde{c}(x) \geq 0$, the term in the integral is non-negative. Hence, this term must vanish. It follows that $(\tilde{u}'(x))^2 \equiv 0$, i.e., $\tilde{u}'(x) \equiv 0$, from what one gets that $\tilde{u}(x)$ has to be a constant. The continuity of $\tilde{u}(x)$ gives in combination with the boundary conditions that $\tilde{u}(x) \equiv 0$. Hence, one gets from (2.6)

$$u(x) = \tilde{u}(x) \exp\left(\frac{1}{2}\int_0^x b(\xi)\ d\xi\right) \equiv 0,$$

in contradiction to the assumption. ■

Remark 2.22. Constant coefficients. In the special case of constant coefficients, condition (2.9) reduces to

$$D := \frac{b^2}{4} + c \geq 0.$$

It is possible to prove statements concerning the solvability of the two-point boundary value problem also for the case $D < 0$, see (Emmrich, 2004, Satz 2.2.2). □

Remark 2.23. Different criterion for the uniqueness of the solution of the fully homogeneous problem. Consider the two-point boundary value problem (2.7), (2.8) with homogeneous right-hand side. Let $u_1(x), u_2(x)$ be two linearly independent solutions of the equation and denote

$$R := \det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix}.$$

The general solution of the homogeneous differential equation has the form

$$u(x) = c_1 u_1(x) + c_2 u_2(x).$$

The parameters have to be determined from the boundary conditions

$$0 = c_1 u_1(0) + c_2 u_2(0), \quad 0 = c_1 u_1(1) + c_2 u_2(1),$$

which is equivalent to the solution of the linear system of equations

$$\begin{pmatrix} u_1(0) & u_2(0) \\ u_1(1) & u_2(1) \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The solution of this system is unique ($c_1 = c_2 = 0$) if and only if $R \neq 0$. Exactly in this case there is only the trivial solution of the fully homogeneous two-point boundary value problem. □

Remark 2.24. The inhomogeneous two-point boundary value problem. Consider the two-point boundary value problem (2.7), (2.8) with inhomogeneous right-hand side. Let $u_1(x), u_2(x)$ be two linearly independent solutions of the corresponding homogeneous differential equation and let

$$A(x) := \det \begin{pmatrix} u_1(0) & u_2(0) \\ u_1(x) & u_2(x) \end{pmatrix}, \quad B(x) := \det \begin{pmatrix} u_1(x) & u_2(x) \\ u_1(1) & u_2(1) \end{pmatrix}.$$

For the investigation of the boundary value problem with inhomogeneous right-hand side, Green's function will be used. □

Definition 2.25. Green[2]'s function. The function $\varGamma(x,\xi)$ is called Green's function for the homogeneous two-point boundary value problem $Lu = 0$, $u(0) = u(1) = 0$, if:

1. $\varGamma(x,\xi)$ is continuous on the square $Q := \{(x,\xi) \ : \ x,\xi \in [0,1]\}$.
2. In both of the triangles
$$Q_1 := \{(x,\xi) \ : \ 0 < \xi < x < 1\}, \quad Q_2 := \{(x,\xi) \ : \ 0 < x < \xi < 1\}$$
there exist continuous partial derivatives $\partial_x \varGamma(x,\xi)$ and $\partial_{xx}\varGamma(x,\xi)$.
3. For fixed $\xi \in I = (0,1)$ is $\varGamma(x,\xi)$, as a function of x, a solution of $L\varGamma = 0$ for $x \neq \xi$, $x \in I$.
4. On the diagonal $x = \xi$, the first derivative has a jump of the form
$$\partial_x \varGamma(x+0,x) - \partial_x \varGamma(x-0,x) = -1, \quad 0 < x < 1.$$
5. $\varGamma(0,\xi) = \varGamma(1,\xi) = 0$ for all $\xi \in (0,1)$.

□

Theorem 2.26. Existence and uniqueness of the solution of the model problem with inhomogeneous right-hand side. *Consider the model problem* (2.7), (2.8) *with* $b, c, f \in C([0,1])$. *If the corresponding fully homogeneous two-point boundary value problem has only the trivial solution, then there is exactly one classical solution of the two-point boundary value problem* (2.7), (2.8). *This solution has the form*
$$u(x) = \int_0^1 \varGamma(x,\xi) f(\xi)\, d\xi$$
with Green's function
$$\varGamma(x,\xi) = \frac{1}{R\ W(\xi)} \begin{cases} A(\xi)B(x) \ \textit{for}\ 0 \le \xi \le x \le 1, \\ A(x)B(\xi) \ \textit{for}\ 0 \le x \le \xi \le 1. \end{cases}$$

Proof. *Sketch.* Direct calculations show that $\varGamma(x,\xi)$ is a Green's function. Similarly, a direct calculation shows that $u(x)$ is a solution of the two-point boundary value problem, *exercise*. Hence, there exists a solution. The uniqueness follows by assuming that there is a second solution. The difference of both solutions satisfies the fully homogeneous problem. But by the assumption, this problem has only the trivial solution. Hence, the second solution is the same solution as the solution given in the theorem. ■

Theorem 2.27. Existence and uniqueness of the solution of the model problem with homogeneous right-hand side, reverse statement of Theorem 2.26. *If the inhomogeneous two-point boundary value problem* (2.7), (2.8) *has for a right-hand side* $f(x) \in C([0,1])$ *exactly one classical solution, then there is only the trivial solution for the corresponding fully homogeneous two-point boundary value problem.*

[2] Georg Green (1793 – 1841)

Proof. Let $u(x)$ be the unique classical solution of the inhomogeneous two-point boundary value problem for $f(x)$ and let $u_{\mathrm{hom}}(x)$ be a nontrivial solution of the fully homogeneous two-point boundary value problem. From the linearity of the problem, it follows that $u(x) + u_{\mathrm{hom}}(x)$ is a classical solution of the two-point boundary value problem for the same right-hand side $f(x)$, which is a contradiction to the assumed uniqueness of this solution. ■

Corollary 2.28. Existence and uniqueness of the solution of the model problem with arbitrary Dirichlet boundary conditions. *Consider the model problem* (2.7) *with* $b \in C^1([0,1])$, $c, f \in C([0,1])$, *and with the Dirichlet boundary conditions* $u(0) = a$, $u(1) = b$ *with* $a, b \in \mathbb{R}$. *If for all* $x \in (0,1)$ *condition* (2.9) *is satisfied, then there is exactly one classical solution of the two-point boundary value problem.*

Proof. Inhomogeneous Dirichlet boundary conditions can be transferred to the right-hand side, see Remark 2.5. This transform is two times continuously differentiable and it can be performed in such a way that the new right-hand side is continuous in $[0,1]$. For this transformed two-point boundary value problem with homogeneous Dirichlet boundary conditions, one can apply the previous statements. It follows from Theorem 2.21 that the fully homogeneous problem has only the trivial solution. With Theorem 2.26, one gets that there is exactly one classical solution. Since the back transform is also two times continuously differentiable, one obtains that there is exactly one solution of the two-point boundary value problem with inhomogeneous Dirichlet boundary conditions. ■

Remark 2.29. Another sufficient condition for unique solvability of the fully homogeneous problem. There are also other sufficient conditions than (2.9) for the fully homogeneous problem to possess only the trivial solution, e.g., see Corollary 2.37 below. □

2.3 Maximum Principle and Stability

Remark 2.30. Differential operator. In this section, L defined by

$$(Lu)(x) := -u''(x) + b(x)u'(x) + c(x)u(x), \quad x \in (0,1),$$

denotes a linear differential operator. For $b, c \in C([0,1])$, this operator maps obviously from $C^2(0,1)$ into $C(0,1)$.

This section proves a maximum principle for this operator. With this principle, the stability of the solution can be shown. Finally, a strong maximum principle will be proved. □

Lemma 2.31. First form of the maximum principle. *Let* $b \in C([0,1])$ *and* $c(x) = 0$ *for all* $x \in [0,1]$. *Then it holds for each* $u \in C^2(0,1) \cap C([0,1])$*:*

i) from $(Lu)(x) \le 0$ *for all* $x \in (0,1)$, *it follows that* $u(x) \le \max\{u(0), u(1)\}$ *for* $x \in [0,1]$,

ii) from $(Lu)(x) \ge 0$ *for all* $x \in (0,1)$, *it follows that* $u(x) \ge \min\{u(0), u(1)\}$ *for* $x \in [0,1]$.

Proof. It is sufficient to prove statement i). Statement ii) follows by replacing $u(x)$ with $-u(x)$.

In the first step, it will be proved that the statement follows with the stronger assumption $(Lu)(x) < 0$ on $(0,1)$. Assume that the function $u(x)$ takes its maximum not on the boundary but in the interior of the interval. Then, there exists an argument $x_0 \in (0,1)$ with $u'(x_0) = 0$ (local extremum) and $u''(x_0) \leq 0$ (local maximum). It follows that

$$(Lu)(x_0) = -u''(x_0) + b(x_0)u'(x_0) = -u''(x_0) \geq 0,$$

which contradicts the assumption.

In the next step, statement i) will be proved. Denote for $\delta, \lambda > 0$

$$w(x) = \delta e^{\lambda x}, \quad x \in [0,1].$$

If λ is sufficiently large, more precisely if $\lambda > \max_{x\in[0,1]} b(x)$, then it follows for all $x \in (0,1)$ that

$$(Lw)(x) = -\lambda^2 w(x) + b(x)\lambda w(x) = -\lambda\left(\lambda - b(x)\right) w(x) < 0.$$

Using the linearity of the differential operator, one gets

$$\left(L(u+w)\right)(x) = (Lu)(x) + (Lw)(x) < 0.$$

Applying the first part of the proof gives

$$u(x) + w(x) \leq \max\{u(0) + w(0), u(1) + w(1)\}.$$

Statement i) follows now for $\delta \to 0$. ■

Theorem 2.32. Maximum principle. *Let $b, c \in C([0,1])$ and $c(x)$ in $[0,1]$ non-negative. Then, it holds for each $u \in C^2(0,1) \cap C([0,1])$ that:*

i) from $(Lu)(x) \leq 0$ for all $x \in (0,1)$, it follows that $u(x) \leq \max\{0, u(0), u(1)\}$ for $x \in [0,1]$,

ii) from $(Lu)(x) \geq 0$ for all $x \in (0,1)$, it follows that $u(x) \geq \min\{0, u(0), u(1)\}$ for $x \in [0,1]$.

Proof. Again, statement ii) follows from statement i) by replacing $u(x)$ with $-u(x)$.

Since $u(x)$ is continuous in $[0,1]$, the set

$$\mathcal{M}^+ := \{x \in (0,1) \ : \ u(x) > 0\}$$

is either empty or a union of open subintervals from $(0,1)$, see the course Calculus I. Let $\mathcal{M}^+ = \emptyset$, i.e., $u(x)$ is in $(0,1)$ nonpositive. Then, statement i) is trivially satisfied.

Let $\mathcal{M}^+ = (0,1)$. Then it is for $x \in (0,1)$

$$-u''(x) + b(x)u'(x) \leq -u''(x) + b(x)u'(x) + c(x)u(x) = (Lu)(x) \leq 0.$$

From Lemma 2.31, it follows that

$$u(x) \leq \max\{u(0), u(1)\},$$

which proves statement i) also in this case.

Consider now $\emptyset \neq \mathcal{M}^+ \neq (0,1)$. It will be shown that $\mathcal{M}^+$ has to be arbitrarily close to 0 or 1. Let $(a_0, b_0) \subseteq \mathcal{M}^+$. If $a_0 \neq 0$ and $u(a_0) > 0$, then it is, because of the continuity of $u(x)$ that either $u(0) > 0$ or there exists a value $0 \leq a_1 < a_0$ with $u(a_1) = 0$. An analog statement holds true for b_0. Thus, one can assume that $a_0 = 0$ or $u(a_0) = 0$ as well as

$b_0 = 1$ or $u(b_0) = 0$. That means, (a_0, b_0) is chosen as large as possible. Altogether, there are four cases to consider. From the assumption, it follows that for all $x \in (a_0, b_0)$

$$(Lu)(x) \le 0 \quad \Longrightarrow \quad -u''(x) + b(x)u'(x) \le -c(x)u(x) \le 0.$$

Now, one can apply again Lemma 2.31. One obtains for all $x \in (a_0, b_0)$

$$0 < u(x) \le \max\{u(a_0), u(b_0)\}. \tag{2.11}$$

Obviously, it is not possible that $u(a_0) = u(b_0) = 0$, because there would be a contradiction to (2.11). The case $a_0 = 0, b_0 = 1$ was already considered. Hence, there remain the cases $a_0 = 0$ and $u(b_0) = 0$ as well as $u(a_0) = 0$ and $b_0 = 1$.

So far, the following is proved: If the set $\mathcal{M}^+$ is not empty, then there are numbers $\hat{a}, \hat{b} \in [0, 1]$ with $\hat{a} \le \hat{b}$, such that

$$\mathcal{M}^+ = (0, \hat{a}) \cup (\hat{b}, 1),$$

where $u(\hat{a}) = 0$ if $\hat{a} \ne 0$, and $u(\hat{b}) = 0$, if $\hat{b} \ne 1$. Using in the next step that $u(x)$ is bounded in $[\hat{a}, \hat{b}]$ by zero from above and applying in the following step (2.11), one gets for $x \in (0, 1)$

$$\begin{aligned} u(x) &\le \max\left\{\max_{x\in(0,\hat{a})} u(x), \max_{x\in(\hat{b},1)} u(x), 0\right\} \\ &\le \max\left\{\max\{u(0), u(\hat{a})\}, \max\{u(\hat{b}), u(1)\}, 0\right\} \\ &= \max\{0, u(0), u(1)\}. \end{aligned}$$

■

Remark 2.33. Physical interpretation. The model problem (2.7), (2.8) can be written in the form

$$(Lu)(x) = f(x), \quad u(0) = u_0, u(1) = u_1.$$

If $(Lu)(x) \le 0$ for all $x \in (0, 1)$, i.e., $f(x) \le 0$ for all $x \in (0, 1)$, then there are no sources of $u(x)$ in $(0, 1)$. The maximum principle says that if at least one of the boundary values is positive, $u(x)$ takes its largest value at the boundary. For instance, if $u(x)$ is a concentration and there are no sources of the concentration in the domain, then there does not exist a local maximum of the concentration in the domain that is larger than the concentration at the boundary. □

Corollary 2.34. Inverse monotonicity, comparison principle. *Let the assumption of Theorem 2.32 be satisfied and let $u, v \in C^2(0, 1) \cap C([0, 1])$ with $u(0) \le v(0)$ and $u(1) \le v(1)$. If $(Lu)(x) \le (Lv)(x)$ for all $x \in (0, 1)$, then it follows that $u(x) \le v(x)$ for $x \in [0, 1]$.*

Proof. This statement follows by applying Theorem 2.32 to the difference $(u - v)(x)$. ■

Theorem 2.35. Stability of the solution, continuous dependence on the data. *Consider the two-point boundary value problem* (2.7), (2.8), *with*

$b, c, f \in C([0,1])$. *If* $c(x)$ *is non-negative in* $[0,1]$, *then it holds for each classical solution* $u(x)$ *the following estimate*

$$\|u\|_{C([0,1])} \leq \Lambda \|f\|_{C([0,1])},$$

where the constant $\Lambda > 0$ *depends on* $b(x), c(x)$, *but not on* $f(x)$.

Proof. Set for $\lambda > 0$

$$w(x) := Be^{\lambda x} - A, \quad x \in [0,1],$$

with

$$A := \Lambda B, \quad B := \|f\|_{C([0,1])}, \quad \Lambda := e^{\lambda} - 1 > 0,$$

where λ will be specified later. Then, using $c(x) \geq 0$, it is for $x \in (0,1)$

$$\begin{aligned} (Lw)(x) &= -\left(\lambda^2 - \lambda b(x) - c(x)\right) Be^{\lambda x} - Ac(x) \\ &\leq -\left(\lambda^2 - \lambda b(x) - c(x)\right) Be^{\lambda x}. \end{aligned}$$

Now, λ is chosen such that $\left(\lambda^2 - \lambda b(x) - c(x)\right) e^{\lambda x} \geq 1$. This relation is satisfied if λ is sufficiently large, e.g., if

$$\lambda \geq \max_{x \in [0,1]} \left(\frac{b(x)}{2} + \sqrt{\frac{b^2(x)}{4} + c(x) + e^{-\lambda x}} \right).$$

One gets for all $x \in (0,1)$ that

$$(Lw)(x) \leq -B = -\|f\|_{C([0,1])}.$$

It follows for all $x \in (0,1)$, using the definition of the norm in $C([0,1])$, that

$$(L(\pm u + w))(x) = \pm f(x) + (Lw)(x) \leq |f(x)| - \|f\|_{C([0,1])} \leq 0.$$

The application of the maximum principle gives

$$\pm u(x) + w(x) \leq \max\{0, \pm u(0) + w(0), \pm u(1) + w(1)\} = \max\{0, w(0), w(1)\}.$$

Hence, it is for all $x \in (0,1)$

$$\pm u(x) \leq \max\{0, w(0), w(1)\} - w(x).$$

From $e^{\lambda x} \geq 1$, it follows that

$$w(x) \geq B - A = w(0), \quad w(1) = Be^{\lambda} - A,$$

i.e.,

$$\begin{aligned} |u(x)| &\leq \max\{0, w(0), w(1)\} - w(x) \leq \max\{0, B - A, Be^{\lambda} - A\} + A - B \\ &= \max\{A - B, 0, B(e^{\lambda} - 1)\} = \max\{A - B, 0, \Lambda B\} = \max\{A - B, 0, A\} = A. \end{aligned}$$

This inequality is the statement of the theorem. ■

Remark 2.36. Non-normalized problem. For the non-normalized two-point boundary value problem (2.1), (2.2) with Dirichlet boundary conditions $u(d) = \alpha$, $u(e) = \beta$, one obtains analogously

$$\|u\|_{C([e,d])} \leq \Lambda \|f\|_{C([e,d])} + \max\{|\alpha|, |\beta|\},$$

where Λ depends also on $e-d$, but not on α, β, (Emmrich, 2004, Satz 2.5.4), *exercise.*

One can see that this estimate is in fact a stability estimate, if one applies it to the difference $u(x) - \tilde{u}(x)$. Here, $u(x)$ is the solution of the problem with exact data and $\tilde{u}(x)$ is the solution of a problem with perturbed right-hand side $\tilde{f}$ or perturbed boundary conditions $\tilde{\alpha}, \tilde{\beta}$. It follows from the linearity of the problem that

$$\|u - \tilde{u}\|_{C([e,d])} \le \Lambda \left\| f - \tilde{f} \right\|_{C([e,d])} + \max\left\{ |\alpha - \tilde{\alpha}|, \left| \beta - \tilde{\beta} \right| \right\}.$$

Hence, changes in the solution depend continuously, in the norm of $C([e,d])$, on changes of the data of the problem. □

Corollary 2.37. Uniqueness of the solution of the fully homogeneous problem. *Consider the two-point boundary value problem* (2.7), (2.8) *with* $b, c \in C([0,1])$ *and* $f(x) \equiv 0$. *If* $c(x)$ *is non-negative in* $[0,1]$, *then the problem has only the trivial solution* $u(x) \equiv 0$.

Proof. This statement follows directly from the estimate from Theorem 2.35, since $\|f\|_{C([0,1])} = 0$. ■

Remark 2.38. Different proof of Corollary 2.37. The statement of Corollary 2.37 follows also from the maximum principle, Theorem 2.32, because for a fully homogeneous problem, both statements i) and ii) of this theorem can be applied and it is $u(0) = u(1) = 0$. □

Corollary 2.39. Uniqueness of the solution of the inhomogeneous problem. *Consider the two-point boundary value problem* (2.7), (2.8) *with* $b, c, f \in C([0,1])$ *and let* $c(x)$ *be non-negative in* $[0,1]$. *Then there is exactly one classical solution of the boundary value problem.*

Proof. The statement of this corollary follows directly from Corollary 2.37 and Theorem 2.26. ■

Lemma 2.40. Another maximum principle. *Let* $b, c \in C([0,1])$, *let* $c(x)$ *be non-negative in* $[0,1]$, *and let* $u \in C^2(0,1) \cap C([0,1])$. *If* $(Lu)(x) < 0$ *for all* $x \in (0,1)$, *then there is no local non-negative maximum of* $u(x)$ *in* $(0,1)$.

Proof. Assume that there is a non-negative local maximum $x_0 \in (0,1)$. From standard calculus, one obtains that $u(x_0) \ge 0$, $u'(x_0) = 0$, and $u''(x_0) \le 0$. It follows that

$$(Lu)(x_0) = -u''(x_0) + b(x_0)u'(x_0) + c(x_0)u(x_0) \ge 0,$$

which is a contradiction to the assumption. ■

Theorem 2.41. Strong maximum principle. *Let* $b, c \in C([0,1])$ *and let* $c(x)$ *non-negative in* $[0,1]$. *If* $u \in C^2(0,1) \cap C([0,1])$ *has in* $(0,1)$ *a non-negative local maximum and it is* $(Lu)(x) \le 0$ *for all* $x \in (0,1)$, *then* $u(x)$ *is constant.*

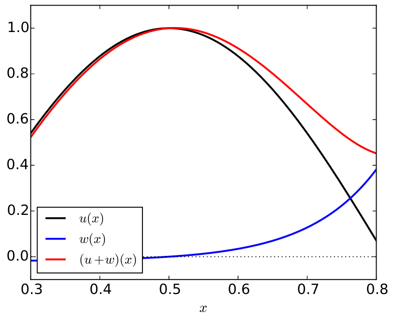

Fig. 2.6 Illustration to the proof of Theorem 2.41, $x_0 = 0.5$, $x_1 = 0.3$, $x_2 = 0.8$.

Proof. Assume that the function $u(x)$ takes in x_0 a non-negative local maximum, i.e., it is in particular $u(x_0) \geq 0$.

Assume that $u(x)$ is not constant. Then, there is a point $x_2 \in (0,1)$ with $u(x_2) < u(x_0)$. Without loss of generality, let $x_2 > x_0$. The case $x_2 < x_0$ can be proved analogously. Now, one chooses x_2 and $x_1 \in [0, x_0)$ such that $u(x)$ takes in x_0 the largest value with respect to the interval $[x_1, x_2]$. This maximum is taken in a closed interval.

Define for $\delta, \lambda > 0$

$$w(x) := \delta \left(e^{\lambda(x-x_0)} - 1 \right), \quad x \in [x_1, x_2].$$

Then, one has obviously

$$w(x) \begin{cases} < 0 \text{ for } x < x_0, \\ = 0 \text{ for } x = x_0, \\ > 0 \text{ for } x > x_0. \end{cases}$$

Now, one chooses λ sufficiently large, e.g., satisfying the inequality

$$\lambda > \max_{x \in [x_1, x_2]} \left(\frac{b(x)}{2} + \sqrt{\frac{b^2(x)}{4} + c(x)} \right).$$

Then, it follows with $c(x) \geq 0$ for all $x \in (x_1, x_2)$ that

$$(Lw)(x) = -\left(\lambda^2 - \lambda b(x) - c(x) \right) \delta e^{\lambda(x-x_0)} - c(x)\delta < 0.$$

One has from the assumption also that

$$(L(u+w))(x) = (Lu)(x) + (Lw)(x) < 0, \quad x \in (x_1, x_2).$$

Now, δ is chosen to be sufficiently small, such that

$$u(x_2) + w(x_2) < u(x_0).$$

Hence, compare Figure 2.6,

$$
\begin{aligned}
u(x) + w(x) &< u(x) \le u(x_0), \quad \text{for } x \in [x_1, x_0),\\
u(x_0) + w(x_0) &= u(x_0) \ge 0,\\
u(x_2) + w(x_2) &< u(x_0),
\end{aligned}
$$

from what follows that the function $(u + w)(x)$ has a non-negative maximum in (x_1, x_2). This property contradicts the statement of Lemma 2.40, since $(L(u+w))(x) < 0$. It follows that the assumption, $u(x)$ being not constant, is wrong. ■

Remark 2.42. Minimum principles. One obtains corresponding minimum principles by replacing $u(x)$ with $-u(x)$. □