Weierstraß-Institut

für Angewandte Analysis und Stochastik

Leibniz-Institut im Forschungsverbund Berlin e. V.

Preprint ISSN 2198-5855

Uniform second order convergence of a complete flux scheme on unstructured 1D grids for a singularly perturbed advection-diffusion equation and some multidimensional extensions

Patricio Farrell, Alexander Linke

submitted: August 10, 2016 (revision: November 30, 2016)

[1] Weierstrass Institute
Mohrenstr. 39
10117 Berlin
Germany
email: patricio.farrell@wias-berlin.de
alexander.linke@wias-berlin.de

No. 2286

Berlin 2016

2010 *Mathematics Subject Classification.* 65L11, 65L20, 65N08, 65N12.

Key words and phrases. singularly perturbed advection-diffusion equation, uniform second-order convergence, finite-volume method, complete flux scheme.

Edited by
Weierstraß-Institut für Angewandte Analysis und Stochastik (WIAS)
Leibniz-Institut im Forschungsverbund Berlin e. V.
Mohrenstraße 39
10117 Berlin
Germany

Fax: +49 30 20372-303
E-Mail: preprint@wias-berlin.de
World Wide Web: http://www.wias-berlin.de/

Abstract

The accurate and efficient discretization of singularly perturbed advection-diffusion equations on arbitrary 2D and 3D domains remains an open problem. An interesting approach to tackle this problem is the complete flux scheme (CFS) proposed by G. D. Thiart and further investigated by J. ten Thije Boonkkamp. For the CFS, uniform second order convergence has been proven on structured grids. We extend a version of the CFS to unstructured grids for a steady singularly perturbed advection-diffusion equation. By construction, the novel finite volume scheme is nodally exact in 1D for piecewise constant source terms. This property allows to use elegant continuous arguments in order to prove uniform second order convergence on unstructured one-dimensional grids. Numerical results verify the predicted bounds and suggest that by aligning the finite volume grid along the velocity field uniform second order convergence can be obtained in higher space dimensions as well.

1 Introduction

Singularly perturbed advection-dominated diffusion problems are extremely challenging to solve numerically [14, 12, 19, 13]. Often stabilization techniques are employed to handle numerical instabilities like upwind or streamline upwind Petrov–Galerkin schemes [14, 2]. Especially useful are those schemes for which one can prove uniform and robust convergence in the discrete maximum norm such that the estimate does not depend on the perturbation parameter. To achieve robustness, only given data of the problem should enter the error estimate, avoiding derivatives of the (unknown) solution that would blow up in boundary layers thus making the error estimate practically worthless for singularly perturbed problems. For example, the famous Il'in–Allen–Southwell (IAS) scheme yields uniform first order convergence in the discrete maximum norm on structured meshes [14]. But there are also robust higher order schemes. The El–Mistikawy–Werle scheme, for instance, converges uniformly with second order on structured grids [14, 21]. Another uniform second order method on structured grids was proposed and analyzed in [20]. It uses a Petrov–Galerkin finite element framework. Such robust numerical methods are highly relevant

for many applications but seem to be analyzed mainly in the framework of finite difference methods on structured grids.

However, from an practical point of view, discretization schemes must yield satisfying results on unstructured and even anisotropic 2D and 3D grids. Therefore, the Voronoï finite volume method in combination with the robust Scharfetter–Gummel scheme — which is just a finite volume variant of the IAS scheme — has become the main tool for solving the van Roosbroeck drift-diffusion equations in semiconductor device simulations [5, 15, 1, 7, 3, 24]. Mesh generators for complicated 2D [16] and 3D [17] domains exist, which allow to construct unstructured Delaunay–Voronoï meshes, though anisotropic meshes in 3D remain a challenge. It is worth pointing out that there are also finite element based simulation tools which take into account the multiscale nature of optoelectronic devices [11].

An interesting attempt to construct uniformly convergent second order finite volume schemes for applications to semiconductor devices and plasma physics was undertaken by the group of J. ten Thije Boonkkamp. In a series of papers [8, 25, 23, 9], he and his coworkers have considerably extended a uniformly convergent second order finite difference scheme originally used by G. D. Thiart [22]. In [25], for the first time this approach was called *complete flux scheme* (CFS). This name is due to the fact that their finite volume flux approximation adds a potential source term contribution to the well-known Scharfetter–Gummel flux of the differential operator.

However, to the best of our knowledge J. ten Thije Boonkkamp and coworkers have not extended their scheme to unstructured meshes. Therefore, the main goal of this paper is to study the convergence of a practical version of the CFS (interpreted as a Voronoï finite volume method) to unstructured meshes. In order to simplify our arguments, we will restrict our contribution to the case of a constant velocity field. In this case, the stiffness matrix of the scheme is just the well-known Scharfetter–Gummel matrix, which is an M-Matrix and invertible [6]. For more general velocity fields, a CFS on unstructured grids will lead to stiffness matrices whose invertibility still has to be investigated.

We make an adjustment to the CFS which is very useful from a practical point of view. The source term is usually only known discretely at each grid point (which is the case for systems of reaction-advection-diffusion equations). Hence, it is reasonable (and simple!) to assume that the source term is constant on each control volume. This implies that the source term jumps at the interface of two cells. For such a piecewise constant source term, we will derive the total numerical flux in the sense of J. ten Thije Boonkkamp. In particular, this implies that for piecewise constant source terms, the derived finite volume scheme is nodally exact in 1D.

In the theoretical part of our contribution, we will analyze the complete flux scheme on (completely) unstructured one-dimensional grids and prove uniform second order convergence. The proof is not merely an extension of existing work on uniform grids based on standard finite difference techniques [9] but rather involves the Green's function of the one-dimensional advection-diffusion operator as well as uniform second order convergence of a similar scheme where the source term is not piecewise constant but piecewise linear.

Even though the proofs are carried out in 1D, we show how this scheme can be extended to higher dimensions. Numerical studies suggest that in order to keep the second order uniform convergence in 2D, one has to use locally orthogonal grids where one direction is aligned along the advection vector. This observation

seems to be reasonable since for very small perturbation parameters, information in the advection-diffusion equations is exchanged only along given *characteristics*. Aligning the grid along these characteristics leads also to a discretization, where the discrete information cannot spoil the discrete solution at neighboring grid points. We point out that the locally orthogonal grids need not be completely structured. Therefore, there is good hope that automatic grid generation for such grids is feasible. We also conjecture that similar grid restrictions hold for finite element discretizations of advection-diffusion problems.

Our paper is structured as follows. After providing a brief introduction to the CFS in the following section, we prove that the CFS with constant source term on each cell converges quadratically on unstructured meshes in the third and present numerical examples which corroborate the theory in the fourth section. In the following, we denote the standard Lebesgue spaces with $L_p(\Omega)$. The Sobolev spaces of integrability order p and smoothness k shall be given by $W_p^k(\Omega)$. As often done, we set $H^k(\Omega) := W_2^k(\Omega)$.

2 Complete Flux Scheme

In this section, we give an introduction to the complete flux scheme. Consider on the unit interval $\Omega := [0,1] \subseteq \mathbb{R}$ the advection-diffusion problem

$$-Du_{xx} + vu_x = s, \qquad u(0) = u(1) = 0. \tag{1}$$

Here $D > 0$ represents the diffusion constant and v represents the constant velocity which we assume to be positive without loss of generality. Even though this is only a one-dimensional problem it can already become quite challenging when the diffusion constant is small compared to the velocity. The source term s is a function which we assume to be in $W_\infty^2(\Omega)$, i. e. in the Sobolev space where weak derivatives up to second order lie in $L_\infty(\Omega)$. This specific choice will become apparent in Section 3. Also note that the whole discussion can be generalised to arbitrary intervals and it is only for notational convenience that we restrict ourselves to the unit interval.

The advection-diffusion problem (1) can be rewritten to

$$f_x = s, \qquad u(0) = u(1) = 0, \tag{2}$$

where the flux function f is given by

$$f := -Du_x + vu. \tag{3}$$

Hence, a finite volume method is adequate for the numerical solution of the original advection-diffusion problem (1) as it allows to mirror numerically the continuous flux conservation property. The complete flux scheme takes into account that the above problem can be inhomogeneous.

Suppose there are $N + 2$ nodes

$$0 = x_0 < x_1 < \ldots < x_N < x_{N+1} = 1. \tag{4}$$

We then define the cell interfaces to be in the middle of two neighboring nodes, that is we set

$$x_{j+1/2} := \frac{x_j + x_{j+1}}{2}$$

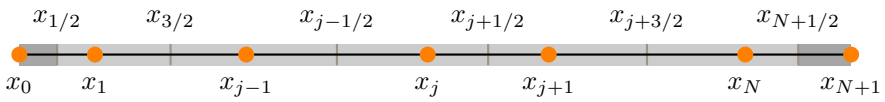

Figure 1: Nodes (orange), Voronoï boxes (grey) and half boxes (dark grey).

for $j = 0, \dots, N$. These points form a new grid which is commonly referred to as Voronoï mesh. The Voronoï boxes or cells are given by

$$K_j := [x_{j-1/2}, x_{j+1/2}]$$

for $j = 1, \dots, N$. Corresponding to both boundary nodes, we introduce two half boxes

$$K_0 := [x_0, x_{1/2}] \quad \text{and} \quad K_{N+1} := [x_{N+1/2}, x_{N+1}].$$

Figure 1 gives an example for such a mesh. Furthermore, it will be useful to define for $j = 0, \dots, N+1$ the mesh parameters

$$h_j := \mathrm{vol}(K_j) \quad \text{and} \quad h := \max_{0 \le j \le N+1} \{h_j\}$$

as well as for $j = 0, \dots, N$

$$\Delta x_{j+1/2} := x_{j+1} - x_j \quad \text{and} \quad \Delta x := \max_{0 \le j \le N} \Delta x_{j+1/2}.$$

Before we go into the details of the complete flux scheme, we state now two special cases of the advection-diffusion problem (1). Since in practice one usually only knows constant samples of the source at each node, there are two natural choices for its approximation to the whole unit interval. Firstly, one can extend the sample value to the entire cell and secondly connect the samples via a linear spline interpolant.

Hence, we define two modified advection-diffusion problems. The first one is given by

$$-D\bar{u}_{xx} + v\bar{u}_x = \bar{s}, \qquad \bar{u}(0) = \bar{u}(1) = 0. \tag{5}$$

Here $\bar{s}$ is defined to be the following $\mathcal{O}(h)$ approximation of the original source term

$$\bar{s}(x) := \begin{cases} s(x_0), & x \in [x_0, x_{1/2}), \\ s(x_i), & x \in [x_{i-1/2}, x_{i+1/2}), \\ s(x_{N+1}), & x \in [x_{N+1/2}, x_{N+1}], \end{cases}$$

for $i = 1, \dots, N$, see Figure 2. In Section 3 we will present an alternative proof showing that the solution to the modified problem (5) yields an $\mathcal{O}(\Delta x^2)$ approximation to the solution of the original problem (1).

The second modified advection-diffusion problem is given by

$$-D\bar{\bar{u}}_{xx} + v\bar{\bar{u}}_x = \bar{\bar{s}}, \qquad \bar{\bar{u}}(0) = \bar{\bar{u}}(1) = 0. \tag{6}$$

Here $\bar{\bar{s}}$ is defined to be the following piecewise linear $\mathcal{O}(\Delta x^2)$ approximation of the original source term

$$\bar{\bar{s}}(x) := s_{i+1/2}(x) := \frac{s(x_{i+1}) - s(x_i)}{\Delta x_{i+1/2}}(x - x_i) + s(x_i), \qquad x \in [x_i, x_{i+1}]$$

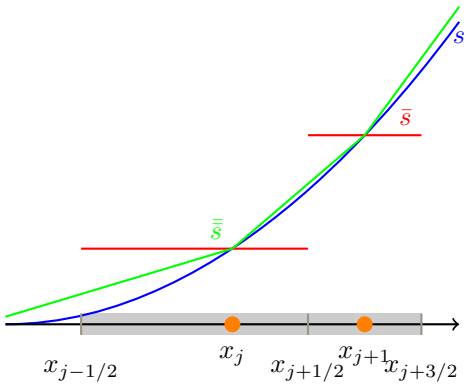

Figure 2: The source term s and its approximations $\bar{s}$ and $\bar{\bar{s}}$.

for $i = 0, \ldots, N$.

Now we turn our attention to the complete flux scheme. Following [23], the complete flux scheme is derived from a two-point boundary value problem on each interval limited by two neighboring nodes. Note, that it is possible to consider nonconstant D and v. However, we do not pursue this here. For some still to be determined boundary values u_j and u_{j+1} we wish to solve on the interval $[x_j, x_{j+1}]$ the problem

$$f_x = (-Du_x + vu)_x = s, \qquad u(x_j) = u_j \quad \text{and} \quad u(x_{j+1}) = u_{j+1} \tag{7}$$

for all interior nodes, i. e. $j = 1, \ldots, N$. The goal is to derive an exact expression for the flux at the cell interface $f_{j+1/2} := f(x_{j+1/2})$. This implies that the interface flux will not only depend on the diffusion, advection and boundary values but also on the source term. Integrating the above ODE from $x_{j+1/2}$ to $x \in [x_j, x_{j+1}]$ yields

$$f(x) - f_{j+1/2} = \int_{x_{j+1/2}}^{x} f_y(y)dy = \int_{x_{j+1/2}}^{x} s(y)dy =: S(x) \tag{8}$$

Defining the Péclet number P as well as the integrating factor $M(x)$

$$P := \frac{v}{D}, \qquad M(x) := e^{-P(x - x_{j+1/2})}$$

and noting that $M_x = -PM$ we compute

$$Mf = -D(Mu_x - MPu) = -D(Mu_x + M_x u) = -D(Mu)_x.$$

Now substituting the expression for f obtained via (8) in the above formula gives after rearrangement

$$M(x)f_{j+1/2} = -D(M(x)u(x))_x - M(x)S(x).$$

From this we deduce via integration from x_j to x_{j+1} that the (exact) flux at the cell interface is given by

$$f_{j+1/2} = -D\frac{M(x_{j+1})u_{j+1} - M(x_j)u_j}{\int_{x_j}^{x_{j+1}} M(x)dx} - \frac{\int_{x_j}^{x_{j+1}} M(x)S(x)dx}{\int_{x_j}^{x_{j+1}} M(x)dx}. \tag{9}$$

We define the fluxes

$$f^h_{j+1/2} := -D\frac{M(x_{j+1})u_{j+1} - M(x_j)u_j}{\int_{x_j}^{x_{j+1}} M(x)dx},$$
$$f^i_{j+1/2} := -\frac{\int_{x_j}^{x_{j+1}} M(x)S(x)dx}{\int_{x_j}^{x_{j+1}} M(x)dx}.$$

The first corresponds to the homogeneous flux, that is to problem (7) with $s = 0$. Analogously, the second term reflects the inhomogeneous flux. Using the Bernoulli function

$$B(x) := \frac{x}{e^x - 1}$$

we can rewrite the homogeneous flux to

$$f^h_{j+1/2} = -\frac{D}{\Delta x_{j+1/2}}\Big\{B(P\Delta x_{j+1/2})u_{j+1} - B(-P\Delta x_{j+1/2})u_j\Big\}. \tag{10}$$

This is the well-known Scharfetter–Gummel scheme [15]. We can now set up a linear system to determine the unknown u_j for $j = 1, \ldots, N$. Note that as long as s does not depend on the solution u, the inhomogeneous flux will only enter the right-hand side of this system.

By integrating $f_x = s$ from (7) over each interior Voronoï box, we obtain

$$f_{j+1/2} - f_{j-1/2} = \int_{K_j} s\,dx.$$

Substituting equations (9) and (10), we obtain the $(N + 2) \times (N + 2)$ linear system

$$\begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ & & \vdots & & \\ & & a_j^T & & \\ & & \vdots & & \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix} \begin{pmatrix} u_0 \\ \vdots \\ u_j \\ \vdots \\ u_{N+1} \end{pmatrix} = \begin{pmatrix} u(x_0) \\ \vdots \\ \int_{K_j} s\,dx - \left(f^i_{j+1/2} - f^i_{j-1/2}\right) \\ \vdots \\ u(x_{N+1}) \end{pmatrix}, \tag{11}$$

where a_j is nonzero only for the indices $i = j - 1, j, j + 1$ and given by

$$a_j = \begin{pmatrix} 0 \\ \vdots \\ -\frac{D}{\Delta x_{j-1/2}}B(-P_{j-1/2}) \\ D\left\{\frac{B(-P_{j+1/2})}{\Delta x_{j+1/2}} + \frac{B(P_{j-1/2})}{\Delta x_{j-1/2}}\right\} \\ -\frac{D}{\Delta x_{j+1/2}}B(P_{j+1/2}) \\ \vdots \\ 0 \end{pmatrix} \quad \text{for} \quad P_{j+1/2} = \frac{v}{D}\Delta x_{j+1/2}.$$

We abbreviate the linear system with

$$A_{\mathrm{SG}}\mathbf{u} = \mathbf{b}(s).$$

Note that thus far the linear system (11) yields a numerical solution which exactly reproduces the flux function at each cell interface. However, it is not always feasible to obtain an analytic expression for the right-hand side since the integral of the source term is involved. Hence, we compute the inhomogeneous flux for the two previously introduced special cases.

For the advection-diffusion problem (5), we need to replace the source with a piecewise constant approximation. Setting $s_j := s(x_j)$, we compute for the inhomogeneous flux

$$\bar{f}^i_{j+1/2} = -\Delta x_{j+1/2}\big\{V(P\Delta x_{j+1/2})s_{j+1} - V(-P\Delta x_{j+1/2})s_j\big) \ ,$$

where the function V is defined as

$$V(x) := \frac{e^{x/2} - 1 - \frac{1}{2}x}{x(e^x - 1)}.$$

It is depicted in Figure 3 and has the following properties:

$$V(0) = 1/8, \quad \lim_{x\to\infty} V(x) = 0 \quad \text{and} \quad \lim_{x\to-\infty} V(x) = 1/2.$$

This results in a right-hand side of the form

$$\mathbf{b}(\bar{s}) = \begin{pmatrix} u(x_0) \\ \vdots \\ h_j s_j - \left(\bar{f}^i_{j+1/2} - \bar{f}^i_{j-1/2}\right) \\ \vdots \\ u(x_{N+1}) \end{pmatrix}. \tag{12}$$

On the other hand for problem (6) we obtain for the inhomogeneous flux

$$\bar{\bar{f}}^i_{j+1/2} = -\Delta x_{j+1/2}\big\{W(P\Delta x_{j+1/2})s_{j+1} - W(-P\Delta x_{j+1/2})s_j\big) \ ,$$

with

$$W(x) := \frac{e^x - 1 - x - \frac{1}{2}x^2}{x^2(e^x - 1)} - \frac{1}{8}.$$

The function is shown in Figure 3 and has the following properties:

$$W(0) = 1/24, \quad \lim_{x\to\infty} W(x) = -1/8 \quad \text{and} \quad \lim_{x\to-\infty} W(x) = 3/8.$$

In this case, we deduce

$$\mathbf{b}(\bar{\bar{s}}) = \begin{pmatrix} u(x_0) \\ \vdots \\ \frac{1}{8}\Delta x_{j-1/2}\, s_{j-1} + \frac{3}{8}(\Delta x_{j-1/2} + \Delta x_{j+1/2})s_j + \frac{1}{8}\Delta x_{j+1/2}\, s_{j+1} - \left(\bar{\bar{f}}^i_{j+1/2} - \bar{\bar{f}}^i_{j-1/2}\right) \\ \vdots \\ u(x_{N+1}) \end{pmatrix}. \tag{13}$$

We point out that in both cases the matrix is the same; only the right-hand sides differ. Using the right-hand sides (12) or (13), we can now define two schemes that numerically solve the boundary value problem (1).

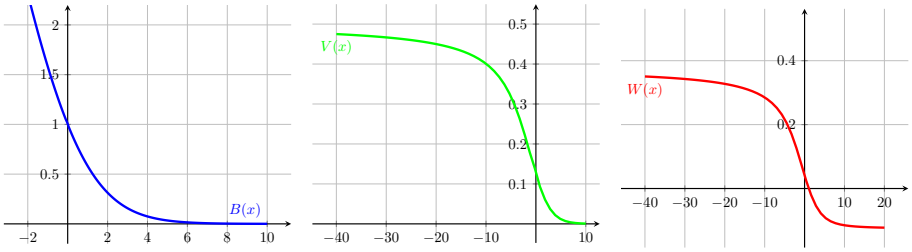

Figure 3: Bernoulli function, $V(x)$ and $W(x)$.

Definition 1. *Let $s \in C(\Omega)$ be the right-hand side of* (1) *and suppose we are given a (possibly unstructured) grid of the form* (4)*. We define two approximation schemes: one for piecewise constant source term (SPC) and one for piecewise linear source term (SPL). In both cases, we start by evaluating s on the grid. If we use* (12) *as right-hand side in the linear system* (11)*, we obtain the SPC scheme. Its discrete solution is given by*

$$\mathbf{u}_{SPC} = A_{SG}^{-1}\mathbf{b}(\bar{s}). \tag{14}$$

Accordingly, the right-hand side (13) *yields the SPL scheme. This time the discrete solution is given by*

$$\mathbf{u}_{SPL} = A_{SG}^{-1}\mathbf{b}(\bar{\bar{s}}). \tag{15}$$

For the SPC scheme it is harder to show the uniform second order convergence on unstructured one-dimensional meshes. However, it is considerably easier to generalize to higher dimensions. On the other hand, for the SPL scheme it is easier to show uniform second order convergence. But it is more difficult to generalize it to higher dimensions.

By construction the following property holds.

Remark 2. *The SPC scheme is nodally exact for piecewise constant source terms and the SPL scheme is nodally exact for piecewise linear source terms.*

This final remark has fascinating implications: The discrete solutions of both schemes (though they only need discrete values of the source term) can be interpreted as the *continuous* solutions of the modified boundary value problems (5) and (6) evaluated on the grid. Hence, instead of studying how well the discrete solution vectors $\mathbf{u}_{SPC}$ and $\mathbf{u}_{SPL}$ approximate the unknown continuous solution u of boundary value problem (1), we will examine how the continuous solutions $\bar{u}$ and $\bar{\bar{u}}$ approximate u. This is a very beneficial feature of our discretization schemes since we can exploit continuous tools (e. g. the Green's function) for the convergence proofs.

3 Error Estimates

We will now discuss error estimates for complete flux schemes based on a piecewise constant approximation as well as a linear approximation of the source term. For this it will be useful to express the solution as a convolution of the source

term and the Green's function. For advection-diffusion problems of the form (1) or (2) the Green's function $G\colon \Omega \times \Omega \to \mathbb{R}$ is given by

$$G(x,y) := \begin{cases} g_1(x,y) := \frac{1}{v}\frac{e^{Px}-1}{e^P-1}(e^{P(1-y)}-1), & x \le y \\ g_2(x,y) := \frac{1}{v}\frac{e^P-e^{Px}}{e^P-1}(1-e^{-Py}), & y \le x. \end{cases} \tag{16}$$

Note, $g_1(x,x) = g_2(x,x)$ for $x \in \Omega$. Also by construction G is not differentiable when $x = y$. Using the Green's function, we can write the solutions u, $\bar{u}$, $\bar{\bar{u}}$ to the respective problems (1), (5) and (6) with homogeneous Dirichlet boundary conditions as convolutions of the form

$$u = G * s = \int_0^1 G(x,y)s(y)dy, \quad \bar{u} = G * \bar{s} \quad \text{and} \quad \bar{\bar{u}} = G * \bar{\bar{s}}. \tag{17}$$

Before we proceed with our error discussion, we state useful properties of the Green's function.

Lemma 3. *Let $x, y \in \Omega$. For any positive Péclet number P, the Green's function* (16) *satisfies*

$$0 \le G(x,y) \le \frac{1}{v}. \tag{18}$$

If $0 < P \le 1$, the bound can be sharpened to

$$0 \le G(x,y) \le \frac{P}{4v}. \tag{19}$$

Proof. Since $g_1, g_2 \ge 0$ and $x, y \in \Omega = (0,1)$, we can immediately deduce that G is nonnegative on the whole domain Ω. For $x \le y$ we have

$$g_1(x,y) = \frac{1}{v}\frac{e^{Px}-1}{e^P-1}(e^{P(1-y)}-1) \le \frac{1}{v}\frac{e^{Py}-1}{e^P-1}(e^{P(1-y)}-1),$$

which is maximised for $y = 1/2$. Hence, we deduce for arbitrary $P > 0$

$$g_1(x,y) \le \frac{1}{v}\frac{(e^{P/2}-1)^2}{e^P-1} = \frac{1}{v}\frac{e^{P/2}-1}{e^{P/2}+1} \le \frac{1}{v}.$$

If $0 < P \le 1$, we find some $\xi \in [0, P/4] \subseteq [0, 1/4]$ such that

$$\frac{e^{P/2}-1}{e^{P/2}+1} = \tanh(P/4) = \tanh'(0)\frac{P}{4} + \frac{\tanh'''(\xi)}{3!}\left(\frac{P}{4}\right)^3 \le \frac{P}{4}.$$

The last inequality follows from the facts that $\tanh'(0) = 1$ and $\tanh'''(y) < 0$ for $y \in [0, 1/4]$. In the same fashion, we see that for $y \le x$

$$g_2(x,y) \le \frac{1}{v}\frac{e^P-e^{Py}}{e^P-1}(1-e^{-Py}) \le \frac{1}{v}\frac{e^{P/2}-1}{e^{P/2}+1},$$

which yields by the same arguments as before the second claim.

□

□

We show now that the solution to problem (6) is an $\mathcal{O}(\Delta x^2)$ approximation to the original solution u. The bound works for nonuniform grids and is independent of the Péclet number.

Theorem 4 (Second order convergence for nonuniform linear spline source term)**.** *Let $s \in W^2_\infty(\Omega)$. Then, we have for positive Péclet numbers P the bound*

$$\|u - \bar{\bar{u}}\|_{L_\infty(\Omega)} \leq \frac{5}{12v} \min\{1, P\} \Delta x^2 \|s''\|_{L_\infty(\Omega)}.$$

Proof. Let $s \in C^\infty(\Omega)$. Using the expansion

$$s(y) = s(x_i) + s'(x_i)(y - x_i) + \frac{1}{2} s''(\xi_i(y))(y - x_i)^2 \tag{20}$$

for some $\xi_i(y) \in [x_i, x_{i+1}]$, we can rewrite the difference between both functions as follows

$$\begin{aligned}
u(x) - \bar{\bar{u}}(x) &= \int_0^1 G(x,y) \left\{ s(y) - \bar{\bar{s}}(y) \right\} dy \\
&= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} G(x,y) \left\{ s(y) - s_{i+1/2}(y) \right. \quad dy \\
&= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} G(x,y) \Bigg\{ \frac{1}{2} s''(\xi_i(y))(y - x_i)^2 \\
&\qquad + \left(s'(x_i) - \frac{s(x_{i+1}) - s(x_i)}{\Delta x_{i+1/2}} \right) (y - x_i) \Bigg\} dy.
\end{aligned}$$

Expanding $s(x_{i+1})$ on the right-hand side according to (20), we find

$$\begin{aligned}
u(x) - \bar{\bar{u}}(x) = \frac{1}{2} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} G(x,y) \left\{ s''(\xi_i(y))(y - x_i)^2 \right. \\
\left. - s''(\xi_i(x_{i+1})) \Delta x_{i+1/2} (y - x_i) \right. \quad dy.
\end{aligned}$$

We note that by Lemma 3 we can bound the Green's function by

$$G(x,y) \leq \frac{1}{v} \min\{1, P\}$$

for any $P > 0$. Hence, taking the absolute value, we derive the bound

$$\begin{aligned}
|u(x) - \bar{\bar{u}}(x)| &\leq \frac{1}{2} \|s''\|_{L_\infty(\Omega)} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} G(x,y) \left((y - x_i)^2 + \Delta x_{i+1/2} (y - x_i) \right) dy \\
&\leq \frac{1}{2v} \min\{1, P\} \|s''\|_{L_\infty(\Omega)} \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} \left\{ (y - x_i)^2 + \Delta x_{i+1/2} (y - x_i) \right. \quad dy \\
&= \frac{5}{12v} \min\{1, P\} \|s''\|_{L_\infty(\Omega)} \sum_{i=0}^{N} \Delta x_{i+1/2}^3 \\
&\leq \frac{5}{12v} \min\{1, P\} \Delta x^2 \|s''\|_{L_\infty(\Omega)}.
\end{aligned}$$

The claim follows by the usual density argument for weakly differentiable Sobolev functions which shows that the theorem holds for general $W^2_\infty(\Omega)$ source terms.

□

□

Our goal is now to prove a similar result for piecewise constant source terms. On the one hand, we will need to be able to bound a sum by an integral. For this we use a generalised version of an estimate often used in the context of the integral test of convergence. On the other hand, we need to understand the asymptotic bounds of a particular function.

Lemma 5. *Let $f \in L_1(\mathbb{R})$ be monotonically decreasing chosen such that for a given sequence of real points $X = \{x_i\}_{i=K}^{M+1}$ with $x_K < \cdots < x_{M+1}$ the corresponding function values at these points exist. Then*

$$\int_{x_K}^{x_{M+1}} f(x)dx \leq \sum_{j=K}^{M} (x_{j+1} - x_j) f(x_j)$$

and

$$\sum_{j=K+1}^{M} (x_j - x_{j-1}) f(x_j) \leq \int_{x_K}^{x_M} f(x)dx.$$

Proof. Due to the assumed monotone decrease we have for any $x_j \in X$

$$f(x) \leq f(x_j) \quad \text{for} \quad x \in [x_j, \infty) \quad \text{and} \quad f(x_j) \leq f(x) \quad \text{for} \quad x \in [x_K, x_j].$$

Hence, for all $K \leq j \leq M$,

$$\int_{x_j}^{x_{j+1}} f(x)dx \leq \int_{x_j}^{x_{j+1}} f(x_j)dx = (x_{j+1} - x_j) f(x_j) \tag{21}$$

and for all $K + 1 \leq j \leq M + 1$

$$\int_{x_{j-1}}^{x_j} f(x)dx \geq \int_{x_{j-1}}^{x_j} f(x_j)dx = (x_j - x_{j-1}) f(x_j). \tag{22}$$

Using (21), we obtain

$$\int_{x_K}^{x_{M+1}} f(x)dx = \sum_{j=K}^{M} \int_{x_j}^{x_{j+1}} f(x)dx \leq \sum_{j=K}^{M} (x_{j+1} - x_j) f(x_j)$$

and with (22), we find

$$\sum_{j=K+1}^{M} (x_j - x_{j-1}) f(x_j) \leq \sum_{j=K+1}^{M} \int_{x_{j-1}}^{x_j} f(x)dx = \int_{x_K}^{x_M} f(x)dx.$$

□

□

Furthermore, we study the asymptotic behavior of the following function.

Lemma 6. *The function*

$$r(x) := 1 - e^{-x} - xe^{-x/2} \tag{23}$$

is monotonically increasing and for $x \geq 0$, we have the bounds

$$0 \leq r(x) \leq 1 \quad \textit{and} \quad r(x) \leq \frac{1}{24}x^3.$$

Proof. Using the inequality $e^{-x/2} \geq 1 - x/2$, we find that

$$r'(x) = e^{-x} - \left(1 - \frac{x}{2}\right) e^{-x/2} \geq 0,$$

which implies that r is monotonically increasing and thus

$$0 = r(0) \leq r(x) \leq \lim_{x \to \infty} r(x) = 1.$$

Furthermore, due to the upper bound we immediately see that $r(x) \leq \frac{1}{24}x^3$ for $x_0 := 24^{1/3} \leq x$. On the other hand, a Taylor expansion around zero yields

$$r(x) = \sum_{j=0}^{3} \frac{r^{(j)}(0)}{j!}x^j + \frac{r^{(4)}(\xi)}{4!}x^4 = \frac{1}{24}x^3 + \frac{r^{(4)}(\xi)}{4!}x^4$$

for $\xi \in [0, x_0]$. If the last term is not positive the desired inequality follows. Since

$$r^{(4)}(\xi) = -e^{-\xi} + \frac{1}{2}e^{-\xi/2} - \frac{1}{16}\xi e^{-\xi/2} \leq 0$$

for any $\xi \in [0, x_e]$, where $x_e \approx 1.94262$ is the unique root of $r^{(4)}$ in the interval $[0, x_0]$. Hence, it only remains to show the inequality for the regime $x_e \leq x \leq x_0$. In this case it follows from

$$r(x) \leq r(x_0) \approx 0.26 \leq \frac{1}{24}x_e^3 \leq \frac{1}{24}x^3.$$

□

□

The two previous lemmas help us now to prove second order convergence of the original solution of problem (1) to the solution of (5) with piecewise constant source term. The key idea of the proof is to show that the difference between the solutions for discontinuous piecewise constant and continuous piecewise linear source terms is of order $\mathcal{O}(\Delta x^2)$ on the grid points.

Theorem 7 (Second order convergence for piecewise constant source term)**.** *Let $s \in W_\infty^2(\Omega)$ and $P > 0$. Then, we have for $0 \leq j \leq N+1$ the bound*

$$|u(x_j) - \bar{u}(x_j)| \leq \frac{1}{v}\left\{\left(C_1 + C_2\right)\|s'\|_{L_\infty(\Omega)} + \frac{5}{12}\min\{1, P\}\|s''\|_{L_\infty(\Omega)}\right\}\Delta x^2,$$

where $C_1 = C_1(x_j, P)$ *and* $C_2 = C_2(x_j, P)$ *are bounded by*

$$\begin{aligned} C_1(x_j, P) &\leq \frac{e}{24} \min\{1, P\} \left(1 + \tanh(P/4)\right) \leq \frac{e}{12} \min\{1, P\} \quad \textit{and} \\ C_2(x_j, P) &\leq 2 \left(\frac{e}{e-1}\right)^2 (1 - e^{-P/2}) \leq 2 \left(\frac{e}{e-1}\right)^2 \min\{1, P/2\}. \end{aligned} \tag{24}$$

Proof. First we note that since we assume homogeneous boundary conditions for u and $\bar{u}$, we only have to show the stated estimates for $j = 1, \ldots, N$. Let $s \in C^\infty(\Omega)$. We start by splitting the error into two parts

$$u(x_j) - \bar{u}(x_j) = G * (s - \bar{s})(x_j) = G * (s - \bar{\bar{s}})(x_j) + G * (\bar{\bar{s}} - \bar{s})(x_j). \tag{25}$$

We will bound the last two contributions to the error separately. We start with the second one. We note that the difference $\bar{\bar{s}} - \bar{s}$ is a superposition of pulses of the form

$$z_{i+1/2}(x) := \begin{cases} \dfrac{s_{i+1} - s_i}{x_{i+1} - x_i}(x - x_i), & x \in [x_i, x_{i+1/2}), \\ \dfrac{s_{i+1} - s_i}{x_{i+1} - x_i}(x - x_{i+1}), & x \in [x_{i+1/2}, x_{i+1}) \\ 0 & \text{otherwise.} \end{cases}$$

With this definition we can write

$$\begin{aligned} G * (\bar{\bar{s}} - \bar{s})(x_j) &= \sum_{i=0}^{N} \int_{x_i}^{x_{i+1}} G(x_j, y) z_{i+1/2}(y) dy \\ &= \sum_{i=0}^{j-1} \int_{x_i}^{x_{i+1}} g_2(x_j, y) z_{i+1/2}(y) dy + \sum_{i=j}^{N} \int_{x_i}^{x_{i+1}} g_1(x_j, y) z_{i+1/2}(y) dy. \end{aligned}$$

When evaluating the integrals in this expression, we obtain

$$\begin{aligned} G * (\bar{\bar{s}} - \bar{s})(x_j) = \frac{1}{vP^2} \Bigg\{ &- \left(\frac{e^P - e^{Px_j}}{e^P - 1}\right) \sum_{i=0}^{j-1} \left(\frac{s_{i+1} - s_i}{x_{i+1} - x_i}\right) e^{-Px_i} r(P\Delta x_{i+1/2}) \\ &+ e^P \left(\frac{e^{Px_j} - 1}{e^P - 1}\right) \sum_{i=j}^{N} \left(\frac{s_{i+1} - s_i}{x_{i+1} - x_i}\right) e^{-Px_i} r(P\Delta x_{i+1/2}) \Bigg\}, \end{aligned}$$

where r denotes the function in (23). The factor in front of the first sum approaches one for x_j close to the left boundary. Similarly, the factor in front of the second sum approaches one for x_j close to the right boundary. Furthermore, we note that the differential quotient appears in both sums. Hence, when applying the absolute value we can bound the error to derive

$$\begin{aligned} |G * (\bar{\bar{s}} - \bar{s})(x_j)| \leq \frac{\|s'\|_{L_\infty(\Omega)}}{vP^2} \Bigg\{ &\left(\frac{e^P - e^{Px_j}}{e^P - 1}\right) \sum_{i=0}^{j-1} e^{-Px_i} r(P\Delta x_{i+1/2}) \\ &+ e^P \left(\frac{e^{Px_j} - 1}{e^P - 1}\right) \sum_{i=j}^{N} e^{-Px_i} r(P\Delta x_{i+1/2}) \Bigg\}. \end{aligned} \tag{26}$$

Before proving the general case, we will investigate two extreme cases, namely dominating diffusion $P\Delta x < 1$ and dominating advection $P\Delta x_{i+1/2} \geq 1$ for $0 \leq i \leq N$. That is, we study the case where the global numerical Péclet number (i. e. also all local ones) is smaller than one separately from the case where the local numerical Péclet number is bigger than one.

(a) The case: $P\Delta x < 1$

Using Lemma 6, Lemma 5 and the assumption $P\Delta x < 1$, we can bound the first sum in (26) by

$$\begin{aligned}\sum_{i=0}^{j-1} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq e^{P\Delta x} \sum_{i=0}^{j-1} e^{-Px_{i+1}} r(P\Delta x_{i+1/2}) \\ &\leq \frac{1}{24} P^3 \Delta x^2 e^{P\Delta x} \sum_{i=1}^{j} e^{-Px_i} \Delta x_{i-1/2} \\ &\leq \frac{1}{24} P^3 \Delta x^2 e^{P\Delta x} \int_{x_0}^{x_j} e^{-Px} dx \\ &\leq \frac{e}{24} P^2 (1 - e^{-Px_j}) \Delta x^2\end{aligned} \tag{27}$$

and similarly the second sum in (26) by

$$\begin{aligned}\sum_{i=j}^{N} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \frac{1}{24} P^2 e^{P\Delta x} (e^{-Px_j} - e^{-Px_{N+1}}) \Delta x^2 \\ &\leq \frac{e}{24} P^2 e^{-Px_j} (1 - e^{-P(1-x_j)}) \Delta x^2.\end{aligned} \tag{28}$$

Therefore, we obtain for the error in (26) the estimate

$$|G * (\bar{\bar{s}} - \bar{s})(x_j)| \leq \frac{1}{v} C_1(x_j, P) \|s'\|_{L_\infty(\Omega)} \Delta x^2,$$

where

$$C_1(x_j, P) := \frac{1}{12} \frac{e}{e^P - 1} (e^P - e^{Px_j})(1 - e^{-Px_j}).$$

We note that C_1 is nonnegative, vanishes if x_j approaches the boundary and is uniformly bounded by

$$\begin{aligned}C_1(x_j, P) &= \frac{e}{24} \left\{ \frac{e^P - e^{Px_j}}{e^P - 1} (1 - e^{-Px_j}) + \frac{e^P (1 - e^{-Px_j})}{e^P - 1} \left(1 - e^{-P(1-x_j)}\right) \right\} \\ &\leq \frac{e}{24} \max\left\{1 - e^{-Px_j}, 1 - e^{-P(1-x_j)}\right\} \frac{2e^P - e^{Px_j} - e^{P(1-x_j)}}{e^P - 1} \\ &\leq \frac{e}{24} \min\{1, P\} 2 \frac{e^P - e^{P/2}}{e^P - 1} \\ &= \frac{e}{24} \min\{1, P\} \left(1 + \tanh(P/4)\right)\end{aligned}$$

since $1 - e^x \leq -x$ for all $x \in \mathbb{R}$.

(b) The case: $P\Delta x_{i+1/2} \geq 1$

If on the other hand, we know $P\Delta x_{i+1/2} \geq 1$ for $1 \leq i \leq N$, then we bound the function r in (26) simply by one. The error is then bounded by

$$|G * (\bar{\bar{s}} - \bar{s})(x_j)| \leq \frac{\|s'\|_{L_\infty(\Omega)}}{vP^2} \frac{e^P}{e^P - 1} \left\{ \left(1 - e^{-P(1-x_j)}\right) \sum_{i=0}^{j-1} e^{-Px_i} + \left(1 - e^{-Px_j}\right) \sum_{i=j}^{N} e^{-P(x_i - x_j)} \right\}. \tag{29}$$

Define the shortest length between any two neighboring nodes

$$\Delta x_{\min} := \min_{0 \leq j \leq N} \Delta x_{j+1/2},$$

then

$$\max \left\{ \sum_{i=0}^{j-1} e^{-Px_i}, \sum_{i=0}^{N-j} e^{-P(x_{i+j} - x_j)} \right\} \leq \sum_{i=0}^{\infty} e^{-Pi\Delta x_{\min}} = \frac{e^{P\Delta x_{\min}}}{e^{P\Delta x_{\min}} - 1} \leq \frac{e}{e-1}. \tag{30}$$

The first inequality follows from

$$x_i \geq i\Delta x_{\min}$$

for $0 \leq i \leq j-1$ as well as

$$x_{i+j} = x_j + (x_{i+j} - x_j) \geq x_j + i\Delta x_{\min}$$

for $0 \leq i \leq N - j$ and the last inequality follows from the fact that even for the shortest length between two neighboring nodes we have $P\Delta x_{\min} \geq 1$. Combining inequality (30) with $1/P \leq \Delta x$, we obtain from (29) for the error this time

$$\begin{aligned} |G * (\bar{\bar{s}} - \bar{s})(x_j)| &\leq \frac{1}{vP^2} \frac{e^P}{e^P - 1} \|s'\|_{L_\infty(\Omega)} \left\{ 2 - e^{-P(1-x_j)} - e^{-Px_j} \right\} \frac{e}{e-1} \\ &\leq \frac{1}{v} C_2(x_j, P) \|s'\|_{L_\infty(\Omega)} \Delta x^2 \end{aligned} \tag{31}$$

where

$$C_2(x_j, P) := \left(\frac{e}{e-1}\right)^2 \left\{ 2 - e^{-P(1-x_j)} - e^{-Px_j} \right\} \leq 2 \left(\frac{e}{e-1}\right)^2 (1 - e^{-P/2}).$$

In the final inequality in (31) we have used that our assumption implies that $P \geq 1$ which allows us to bound $e^P/(e^P - 1)$ by $e/(e-1)$.

(c) The general case:

Finally, let us consider the case between these two extreme cases. We assume that the nodes are distributed in such a way that neither of the previous cases holds for the entire grid. In particular, this implies $1/P \leq \Delta x$.

For $0 \leq m \leq N$, we define the point sets

$$\begin{aligned} X_m &:= \{x_k \in X \mid m \leq k\}, \\ \bar{Y}_m &:= \{x_k \in X \mid P\Delta x_{k+1/2} < 1, m \leq k \leq N\}, \\ Z_m &:= \{x_k \in X \mid P\Delta x_{k+1/2} \geq 1, m \leq k \leq N\} \end{aligned}$$

and corresponding index sets

$$\begin{aligned} I_{\bar{Y}_m} &:= \{k \in \mathbb{N} \mid P\Delta x_{k+1/2} < 1, m \leq k \leq N\}, \\ I_{Z_m} &:= \{k \in \mathbb{N} \mid P\Delta x_{k+1/2} \geq 1, m \leq k \leq N\}. \end{aligned}$$

The sets $\bar{Y}_m$ and Z_m are disjoint and we have

$$X \setminus \{x_0, \ldots, x_{m-1}, x_{N+1}\} = X_m \setminus \{x_{N+1}\} = \bar{Y}_m \cup Z_m.$$

We want to bound the two sums

$$\sum_{i=0}^{j-1} e^{-Px_i} r(P\Delta x_{i+1/2}) \qquad \text{and} \qquad \sum_{i=j}^{N} e^{-Px_i} r(P\Delta x_{i+1/2}) \tag{32}$$

in (26) for $1 \leq j \leq N$. Using the new notation, the second sum can be expressed as a sum of two sums:

$$\sum_{i=j}^{N} e^{-Px_i} r(P\Delta x_{i+1/2}) = \sum_{i \in I_{\bar{Y}_j}} e^{-Px_i} r(P\Delta x_{i+1/2}) + \sum_{i \in I_{Z_j}} e^{-Px_i} r(P\Delta x_{i+1/2}).$$

Thus, we have split the sum into an diffusion-dominated part and a advection-dominated. Again, we will bound the final two sums separately. We start with the first sum. We modify the set $\bar{Y}_j$ in such a way that we can use our previous estimate. The problem is that currently two neighboring points, say $x_k < x_\ell$ in $\bar{Y}_j$, will violate the smallness condition $P(x_\ell - x_k) < 1$ if at least one point of the set Z_j lies between them. In this case, $\ell > k + 1$. Hence, we construct a new set $Y_j \supseteq \bar{Y}_j$. Apart from $\bar{Y}_j$, the new set shall include the point $y_j := x_j$ if it is not already included and as many points (all of them bigger than y_j) as necessary until

$$P\Delta y_j < 1$$

is satisfied where

$$\Delta y_j := \max_{j \leq k \leq N_{Y_j}} \left\{ y_{k+1} - y_k \right\} \leq \Delta \bar{y}_j := \max \left\{ \Delta x_{i+1/2} \mid i \in I_{\bar{Y}_j} \right\},$$

for $Y_j := \{y_j, \ldots, y_{N_{Y_j}} \mid y_j < \ldots < y_{N_{Y_j}} < y_{N_{Y_j}+1} = 1\}$ and some natural number $N_{Y_j} \geq j$. These additional points do not belong to the original mesh X. However, every *diffusion-dominated* pair of original mesh points in the sense of $\bar{Y}_j$ is included. Note that $\Delta \bar{y}_j$ is the biggest difference between any two points whose $I_{\bar{Y}_j}$ indices are neighbors. Hence, the previously mentioned difference between x_k and x_ℓ would not contribute to the set Y_j over which the maximum is taken since new points are filled between them. As before we set $\Delta y_{i+1/2} = y_{i+1} - y_i$.

With the help of this new set we derive the bound

$$\begin{aligned}\sum_{i\in I_{\bar{Y}_j}} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \sum_{i=j}^{N_{Y_j}} e^{-Py_i} r(P\Delta y_{i+1/2}) \\ &\leq \frac{e}{24} P^2 e^{-Px_j}(1-e^{-P(1-x_j)})\Delta y_j^2 \\ &\leq \frac{e}{24} P^2 e^{-Px_j}(1-e^{-P(1-x_j)})\Delta x^2,\end{aligned}$$

where we have used that $x_j = y_j$ and (28). On the other hand, we see similarly as before that

$$\begin{aligned}\sum_{i\in I_{Z_j}} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \sum_{i\in I_{Z_j}} e^{-Px_i} = e^{-Px_j} \sum_{i\in I_{Z_j}} e^{-P(x_i-x_j)} \\ &\leq e^{-Px_j}\left(1+\sum_{i=1}^{\infty} e^{-Pi\Delta z_{\min}}\right) \\ &= e^{-Px_j}\sum_{i=0}^{\infty} e^{-Pi\Delta z_{\min}} \leq e^{-Px_j}\frac{e}{e-1}\end{aligned}$$

where $\Delta z_{\min} = \min_{i\in I_{Z_j}}\{\Delta x_{i+1/2}\}$ which satisfies by construction $P\Delta z_{\min} \geq 1$. Since x_j is not necessarily in Z_j, we cannot deduce in the lines above that $\min\{Z_j\} - x_j \geq \Delta z_{\min}$. Instead we bound the corresponding exponential $e^{-P(\min\{Z_j\}-x_j)}$ by one. With regard to the first sum in (32), we split it again in two parts:

$$\sum_{i=0}^{j-1} e^{-Px_i} r(P\Delta x_{i+1/2}) = \sum_{\substack{i\in I_{\bar{Y}_0}\\ i\leq j-1}} e^{-Px_i} r(P\Delta x_{i+1/2}) + \sum_{\substack{i\in I_{Z_0}\\ i\leq j-1}} e^{-Px_i} r(P\Delta x_{i+1/2}).$$

We construct a finer mesh Y_0 with similar properties as before to bound the first sum via (27). The second sum can be estimated directly as in the previous discussion. We obtain for both sums on the right-hand side

$$\begin{aligned}\sum_{\substack{i\in I_{\bar{Y}_0}\\ i\leq j-1}} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \frac{e}{24} P^2 (1-e^{-Px_j})\Delta x^2, \\ \sum_{\substack{i\in I_{Z_0}\\ i\leq j-1}} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \frac{e}{e-1}.\end{aligned}$$

So combining the above results we find

$$\begin{aligned}\sum_{i=0}^{j-1} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \frac{e}{24} P^2 \left(1-e^{-Px_j}\right)\Delta x^2 + \frac{e}{e-1}, \\ \sum_{i=j}^{N} e^{-Px_i} r(P\Delta x_{i+1/2}) &\leq \frac{e}{24} P^2 e^{-Px_j}\left(1-e^{-P(1-x_j)}\right)\Delta x^2 + e^{-Px_j}\frac{e}{e-1}.\end{aligned}$$

Thus by using these bounds in (26), $1/P \leq \Delta x$ as well as $P \geq 1$, we have for $1 \leq j \leq N$ the bound

$$|G * (\bar{\bar{s}} - \bar{s})(x_j)| \leq \frac{1}{v} \left\{ \Big(C_1 + C_2 \Big) \|s'\|_{L_\infty(\Omega)} \right\} \Delta x^2, \tag{33}$$

where the constants $C_1 = C_1(x_j, P)$ and $C_2 = C_2(x_j, P)$ are defined as before. We recall from cases (a) and (b) the bounds

$$C_1(x_j, P) \leq \frac{e}{24} \min\{1, P\} \left(1 + \tanh(P/4)\right) \leq \frac{e}{12} \min\{1, P\} \quad \text{and}$$
$$C_2(x_j, P) \leq 2 \left(\frac{e}{e-1} \right)^2 (1 - e^{-P/2}) \leq 2 \left(\frac{e}{e-1} \right)^2 \min\{1, P/2\}.$$

Finally, we bound the first part in (25) using Theorem 4 by

$$|G * (s - \bar{\bar{s}})(x_j)| \leq \frac{5}{12v} \min\{1, P\} \Delta x^2 \|s''\|_{L_\infty(\Omega)}. \tag{34}$$

The claim follows by the usual density argument for Sobolev spaces after taking the absolute value and inserting the bounds (33) and (34) in (25).

□

□

We point out that the bound on C_1 for small P approaches a similar bound one would obtain for the corresponding diffusion problem i. e. when $v = 0$. By Remark 2, we automatically obtain now the following bounds for our numerical schemes.

Theorem 8. *Let $s \in W^2_\infty(\Omega)$. For $0 \leq j \leq N+1$, we have*

$$|u(x_j) - (\mathbf{u}_{SPL})_j| \leq \frac{5}{12v} \min\{1, P\} \Delta x^2 \|s''\|_{L_\infty(\Omega)}.$$

and

$$|u(x_j) - (\mathbf{u}_{SPC})_j| \leq \frac{1}{v} \left\{ \Big(C_1 + C_2 \Big) \|s'\|_{L_\infty(\Omega)} + \frac{5}{12} \min\{1, P\} \|s''\|_{L_\infty(\Omega)} \right\} \Delta x^2.$$

with C_1 and C_2 satisfying (24).

□

4 Numerical Examples

4.1 1D Test Problem

We verify our theoretically obtained bound for the SPC scheme on unstructured grids now numerically. Consider the source term

$$s(x) = D\pi^2 \sin(\pi x) + \pi v \cos(\pi x) \tag{35}$$

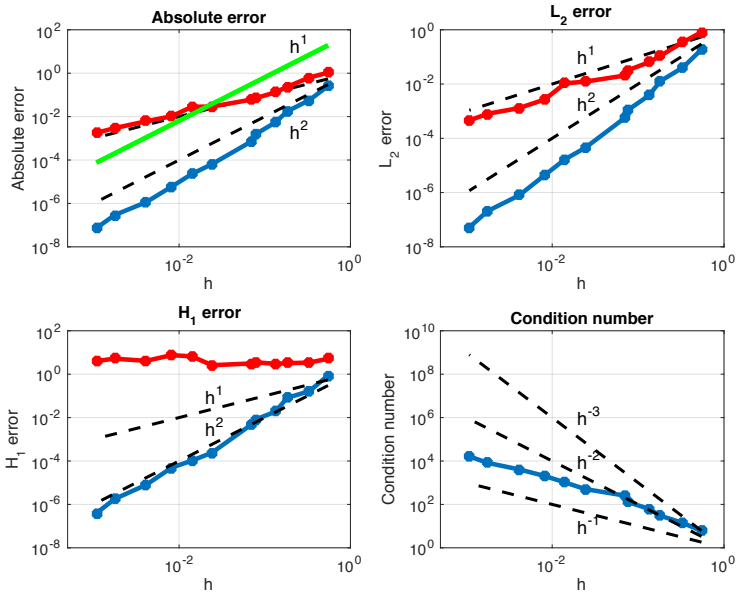

Figure 4: 1D convergence plots for the Scharfetter–Gummel scheme (red) and the complete flux scheme with piecewise constant source term (blue) piecewise constant source term for $D = 10^{-12}$ and $v = 1$. The successively finer nonuniform grids are constructed from uniformly distributed pseudorandom numbers. The green line shows the (weakest) bound derived in Theorem 7.

D	central difference	simple upwind	Scharfetter–Gummel	SPC
10^{-2}	5.94×10^{-1}	$1.18 \times 10^{+0}$	$1.01 \times 10^{+0}$	2.65×10^{-1}
	1.60×10^{-2}	8.31×10^{-2}	4.90×10^{-2}	1.51×10^{-3}
	2.82×10^{-7}	8.95×10^{-4}	7.25×10^{-6}	5.06×10^{-8}
10^{-3}	$1.73 \times 10^{+0}$	$1.21 \times 10^{+0}$	$1.05 \times 10^{+0}$	2.57×10^{-1}
	1.88×10^{-2}	7.27×10^{-2}	7.10×10^{-2}	1.52×10^{-3}
	1.29×10^{-6}	1.26×10^{-3}	1.87×10^{-4}	1.51×10^{-7}
10^{-4}	$1.31 \times 10^{+1}$	$1.21 \times 10^{+0}$	$1.06 \times 10^{+0}$	2.57×10^{-1}
	2.59×10^{-2}	7.46×10^{-2}	6.99×10^{-2}	1.52×10^{-3}
	3.28×10^{-6}	1.61×10^{-3}	1.23×10^{-3}	7.12×10^{-8}
10^{-5}	$1.27 \times 10^{+2}$	$1.21 \times 10^{+0}$	$1.06 \times 10^{+0}$	2.57×10^{-1}
	1.15×10^{-1}	7.46×10^{-2}	6.98×10^{-2}	1.52×10^{-3}
	4.29×10^{-6}	1.76×10^{-3}	1.72×10^{-3}	7.33×10^{-8}
10^{-6}	$1.27 \times 10^{+3}$	$1.21 \times 10^{+0}$	$1.06 \times 10^{+0}$	2.57×10^{-1}
	7.68×10^{-1}	7.46×10^{-2}	6.98×10^{-2}	1.52×10^{-3}
	4.58×10^{-6}	1.78×10^{-3}	1.78×10^{-3}	7.35×10^{-8}

Table 1: Maximum errors for different schemes, decreasing diffusion constants D as well as different grid sizes. The first, second and third values in each box correspond to the mesh sizes $\Delta x = 0.5472, \Delta x = 0.0775$ and $\Delta x = 0.0011$, respectively. The nonuniform grids are constructed from uniformly distributed pseudorandom numbers.

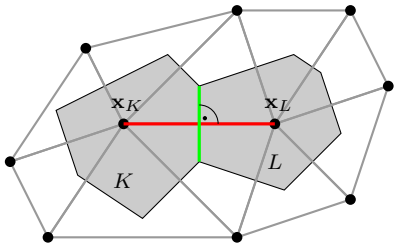

Figure 5: Two Voronoï cells belonging to collocation node $\mathbf{x}_K$ and $\mathbf{x}_L$.

for problem (1) with $D = 10^{-12}$ and $v = 1$. This implies that the solution is given by $u(x) = \sin(\pi x)$. We then solve (5) with the complete flux scheme. Figure 4 shows convergence plots for different errors, comparing the Scharfetter–Gummel scheme with the complete flux scheme. The successively finer grids are constructed from uniformly distributed pseudorandom numbers. Also depicted is the (weakest) bound derived in Theorem 7. It is worth pointing out that not only the maximum error but also the H^1 error converges quadratically. Initially, the condition numbers grow quadratically. On finer grids, however, their growth becomes linear.

In Table 1, we compare the maximum errors of the SPC scheme to three other schemes: a central difference, a simple upwind and a Scharfetter–Gummel discretization of the flux [14]. It can be observed that on unstructured grids only the central difference and the SPC scheme converge quadratically. However, the former is not robust with respect to the diffusion constant D. Even though the simple upwind and the Scharfetter–Gummel scheme converge only linearly, their maximum error does not explode when the diffusion constant decreases. The SPC scheme yields for all grids and diffusion constants the smallest error.

4.2 2D Test Problem

In order to study how the presented one-dimensional scheme can be used in higher space dimensions, we study the problem

$$-D\,\mathrm{div}(\mathrm{grad}(u)) + \mathrm{div}(u\mathbf{v}) = s \tag{36}$$

on the two-dimensional domain $\Omega = [0,1]^2$ with $\mathbf{v} = (v, 0)^T$ and

$$s(x,y) = 2D\pi^2 \sin(\pi x)\sin(\pi y) - \pi v \cos(\pi x)\sin(\pi y) \tag{37}$$

as well as homogeneous Dirichlet boundary conditions. The analytical solution is then given by

$$u(x,y) = \sin(\pi x)\sin(\pi y).$$

In order to discretize the domain Ω, we employ a Voronoï box based finite volume method introduced in [10], also known as *box method* due to [1]. It uses a simplical boundary conforming Delaunay grid [18] which allows to obtain control volumes surrounding each given collocation point by joining the circumcenters of the simplices containing it. Figure 5 shows the construction of two Voronoï boxes belonging to collocation nodes $\mathbf{x}_K$ and $\mathbf{x}_L$. The main advantage of constructing grids in this fashion is that one can project the higher dimensional flux on a

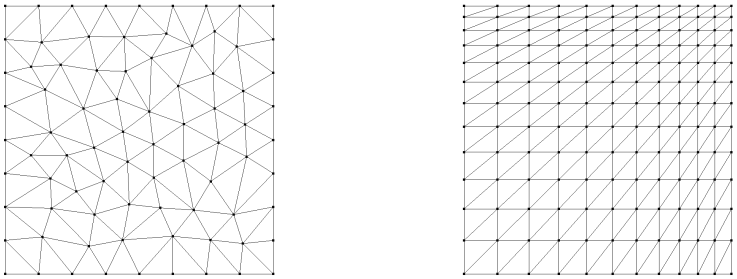

Figure 6: Examples of an unstructured grid (left) and advection-aligned grid (right) used for the numerical computations. Note that for the latter grid the grading along the advection and the grading along the orthogonal direction are different. The two-dimensional finite volume method uses the dual grids of these grids to define the control volumes. The dual grid of the triangulation on the right is locally orthogonal and aligned with the advection vector since the diagonal edges in the primary triangulation have no effect on the dual grid (the circumcenters of the those triangles adjacent to diagonal edges coincide).

one-dimensional edge. Thus, effectively reducing the complexity of the problem. This kind of finite volume method has been extensively studied in [4, 5]. In Figure 6 two different kind of meshes are shown which will be used to study the convergence behavior: an unstructured and a graded advection-aligned grid. Each vertex in both grids corresponds to a node in the mesh. The dual mesh (the so-called Voronoï mesh) is then used to setup the control volumes.

Figure 7 shows the absolute, the L_1, the L_2 and L_∞ errors for $D = 10^{-5}$ and $v = 1$ when using successively finer unstructured triangle grids to define the control volumes. It can be observed that in this case the complete flux scheme converges only linearly. However, if one uses graded advection-aligned grids as in Figure 8 the uniform second order convergence can be recovered. The grading in the direction of the advection vector is different from the grading in the orthogonal direction as can be seen in Figure 6. For this type of mesh, the H^1 error converges again quadratically just as for the one-dimensional example.

This numerical example suggests that in order to keep the second order uniform convergence in 2D, one has to use locally orthogonal grids where one direction is aligned along the advection vector. This observation seems to be reasonable since for very small perturbation parameters, information in the advection-diffusion equations is exchanged only along given characteristics. Aligning the grid along these characteristics leads also to a discretization, where the discrete information cannot spoil the discrete solution at neighboring grid points. We note that the locally orthogonal grids need not be completely structured as the right image in Figure 6 shows. Therefore, there is good hope that automatic grid generation for such grids is feasible. We also conjecture that similar grid restrictions hold for finite element discretizations of advection-diffusion problems.

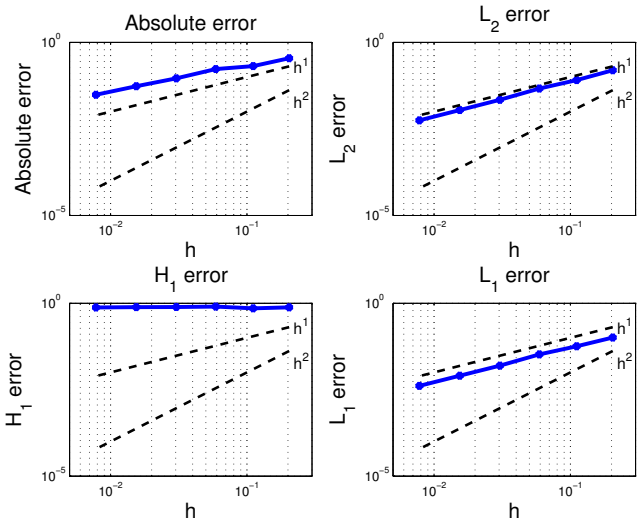

Figure 7: 2D convergence plot for (36) using the complete flux scheme with piecewise constant source term generated from (37) with $D = 10^{-5}$ and $v = 1$ where the control volumes are generated from successively finer unstructured triangle grids.

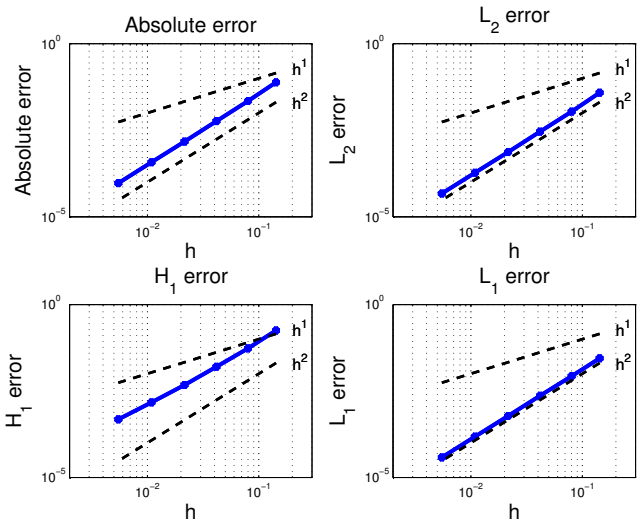

Figure 8: 2D convergence plot for (36) using the complete flux scheme with piecewise constant source term generated from (37) with $D = 10^{-10}$ and $v = 1$ where the control volumes are generated from successively finer, graded advection-aligned grids.

5 Conclusion

We have shown uniform second order convergence of two one-dimensional complete flux schemes. One of these schemes approximates the source term by a piecewise constant source term. Numerical results suggest that locally orthogonal advection-aligned grids lead to uniform second order convergence for the latter scheme in higher dimensions as well. We conjecture that flow-aligned grids are a universal restriction to finite volume and finite element discretizations if one wishes to obtain uniform second order convergence for singularly perturbed advection-diffusion equations. Future research needs to be done to understand how our approach can be extended to variable velocity fields.

References

[1] Bank, R.E., Rose, D.J.: Some error estimates for the box method. SIAM J. Numer. Anal **24**, 777–787 (1987)

[2] Brooks, A.N., Hughes, T.J.R.: Streamline upwind/Petrov-Galerkin formulations for convection dominated flows with particular emphasis on the incompressible Navier-Stokes equations. Comput. Methods Appl. Mech. Engrg. **32**(1-3), 199–259 (1982). FENOMECH '81, Part I (Stuttgart, 1981)

[3] Chainais-Hillairet, C., Jüngel, A., Shpartko, P.: A finite-volume scheme for a spinorial matrix drift-diffusion model for semiconductors. Numerical Methods for Partial Differential Equations (2015)

[4] Eymard, R., Gallouët, T., Herbin, R.: Finite volume methods. In: Solution of Equation in $\mathbb{R}^n$ (Part 3), Techniques of Scientific Computing (Part 3), *Handbook of Numerical Analysis*, vol. 7, pp. 713 – 1018. Elsevier (2000)

[5] Farrell, P., Rotundo, N., Doan, D.H., Kantner, M., Fuhrmann, J., Koprucki, T.: Electronics: Numerical methods for drift-diffusion models. In: Handbook of Optoelectronic Device Modeling and Simulation. Taylor & Francis (to appear in 2017)

[6] Fuhrmann, J., Langmach, H.: Stability and existence of solutions of time-implicit finite volume schemes for viscous nonlinear conservation laws. Appl. Numer. Math. **37**(1-2), 201–230 (2001)

[7] Gärtner, K.: DEPFET sensors, a test case to study 3d effects. J. Comput. Electron **6**, 275–278 (2007)

[8] van't Hof, B., ten Thije Boonkkamp, J.H.M., Mattheij, R.M.M.: Discretization of the stationary convection-diffusion-reaction equation. Numer. Methods Partial Differential Equations **14**(5), 607–625 (1998)

[9] Liu, L., van Dijk, J., ten Thije Boonkkamp, J., Mihailova, D., van der Mullen, J.: The complete flux scheme—error analysis and application to plasma simulation. Journal of Computational and Applied Mathematics **250**(0), 229 – 243 (2013)

[10] Macneal, R.H.: An asymmetrical finite difference network. Quart. Math. Appl. **11**, 295–310 (1953)

[11] Auf der Maur, M., Povolotskyi, M., Sacconi, F., Pecchia, A., Romano, G., Penazzi, G., Di Carlo, A.: TiberCAD: towards multiscale simulation of optoelectronic devices. Optical and quantum electronics **40**(14-15), 1077–1083 (2008)

[12] Morton, K.: Numerical Solution Of Convection-Diffusion Problems. Applied Mathematics. Taylor & Francis (1996)

[13] Roos, H.G., Stynes, M.: Some open questions in the numerical analysis of singularly perturbed differential equations. Comput. Methods Appl. Math. **15**(4), 531–550 (2015)

[14] Roos, H.G., Stynes, M., Tobiska, L.: Robust numerical methods for singularly perturbed differential equations, *Springer Series in Computational Mathematics*, vol. 24, 2nd edn. Springer, Berlin (2008)

[15] Scharfetter, D., Gummel, H.: Large-signal analysis of a silicon read diode oscillator. IEEE Transactions on Electron Devices **16**(1), 64–77 (1969)

[16] Shewchuk, J.: Triangle: A two-dimensional quality mesh generator and Delaunay triangulator. http://www.cs.cmu.edu/∼quake/triangle.html, University of California at Berkeley

[17] Si, H.: Tetgen, a delaunay-based quality tetrahedral mesh generator. ACM Trans. Math. Softw. **41**(2), 11:1–11:36 (2015)

[18] Si, H., Gärtner, K., Fuhrmann, J.: Boundary conforming Delaunay mesh generation. Computational Mathematics and Mathematical Physics **50**(1), 38–53 (2010)

[19] Stynes, M.: Steady-state convection-diffusion problems. Acta Numerica **14**, 445–508 (2005)

[20] Stynes, M., O'Riordan, E.: A uniformly accurate finite element method for a singular perturbation problem in conservative form. SIAM Journal on Numerical Analysis **23**(2), 369–375 (1986)

[21] Surla, K., Uzelac, Z.: An analysis and improvement of the El Mistikawy and Werle scheme. Publ. Inst. Math. (Beograd) (N.S.) **54(68)**, 144–155 (1993)

[22] Thiart, G.D.: Improved finite-difference scheme for the solution of convection-diffusion problems with the simplen algorithm. Numer. Heat Transfer, Part B **18**(1), 81–95 (1990)

[23] ten Thije Boonkkamp, J., Anthonissen, M.: The finite volume-complete flux scheme for advection-diffusion-reaction equations. Journal of Scientific Computing **46**(1), 47–70 (2011)

[24] ten Thije Boonkkamp, J., Schilders, W.H.: An exponential fitting scheme for the electrothermal device equations specifically for the simulation of avalanche generation. COMPEL-The international journal for computation and mathematics in electrical and electronic engineering **12**(2), 95–111 (1993)

[25] ten Thije Boonkkamp, J.H.M.: A complete flux scheme for one-dimensional combustion simulation. In: Finite volumes for complex applications IV, pp. 573–583. ISTE, London (2005)