ESAIM: M2AN 55 (2021) 3017–3042
https://doi.org/10.1051/m2an/2021078

ESAIM: Mathematical Modelling and Numerical Analysis
www.esaim-m2an.org

CONSISTENCY AND CONVERGENCE FOR A FAMILY OF FINITE VOLUME DISCRETIZATIONS OF THE FOKKER–PLANCK OPERATOR

Martin Heida*, Markus Kantner and Artur Stephan

Abstract. We introduce a family of various finite volume discretization schemes for the Fokker–Planck operator, which are characterized by different Stolarsky weight functions on the edges. This family particularly includes the well-established Scharfetter–Gummel discretization as well as the recently developed square-root approximation (SQRA) scheme. We motivate this family of discretizations both from the numerical and the modeling point of view and provide a uniform consistency and error analysis. Our main results state that the convergence order primarily depends on the quality of the mesh and in second place on the choice of the Stolarsky weights. We show that the Scharfetter–Gummel scheme has the analytically best convergence properties but also that there exists a whole branch of Stolarsky means with the same convergence quality. We show by numerical experiments that for small convection the choice of the optimal representative of the discretization family is highly non-trivial, while for large gradients the Scharfetter–Gummel scheme stands out compared to the others.

Mathematics Subject Classification. 35Q84, 49M25, 65N08.

Received September 28, 2020. Accepted November 17, 2021.

1. Introduction

The Fokker–Planck equation (FPE), also known as *Smoluchowski equation* or *Kolmogorov forward equation*, is one of the most important equations in theoretical physics and applied mathematics with application in physical chemistry, protein synthesis, plasma physics, semiconductor device simulation and others. Originally, it was introduced to describe the time evolution of the probability density function of a particle exposed to fluctuating forces (*e.g.*, Brownian motion). There is a huge interest in the development of efficient and robust numerical methods for the FPE. In the context of finite volume (FV) methods, the central objective is a robust and accurate discretization of the (particle or probability) flux implied by the FPE.

A particularly important discretization scheme for the flux was derived by Scharfetter and Gummel [47] (in the context of the drift-diffusion model for electronic charge carrier transport in bipolar semiconductor devices [49]) and independently by Allan and Southwell [1] and Il'in [29]. The typically exponentially varying carrier densities at p-n junctions lead to unphysical results (spurious oscillations), if the flux is discretized in a naive way using standard finite difference schemes [45]. The problem was overcome by considering the flux expression as a one-dimensional boundary value problem along each edge between adjacent mesh nodes. The resulting Scharfetter–Gummel (SG) scheme provides a robust discretization of the flux as it asymptotically approaches the

Keywords and phrases. Finite volume, Fokker–Planck, Stolarsky mean.

WIAS Berlin, Mohrenstraße 39, 10117 Berlin, Germany.
*Corresponding author: martin.heida@wias-berlin.de

© The authors. Published by EDP Sciences, SMAI 2021

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

numerically stable discretizations in the drift- (upwind scheme) and diffusion-dominated (central finite difference scheme) limits. The SG-scheme and its several generalizations are nowadays widely used in semiconductor device simulation [21,38] and have been extensively studied in the literature [4,18,20,31].

Recently, an alternative flux discretization method, called *square-root approximation* (SQRA) scheme, has been derived explicitly for high dimensional problems in molecular dynamics [33]. It was independently obtained from a maximum entropy path principle [13] and from discretizing the Jordan–Kinderlehrer–Otto variational formulation of the FPE [40]. In Section 3.2, we derive the SQRA from the theory of gradient flows. In contrast to the SG-scheme, the SQRA is very recent and only sparsely investigated.

We point out that the SG and the SQRA schemes as well as others (*e.g.*, the Chang–Cooper scheme [9]) are special cases of a family of discretization schemes based on weighted Stolarsky means [48], see Section 3.1, allowing for a unified analysis. We provide further mathematical context in Section 1.2 below.

1.1. The FPE and the SG and SQRA discretization schemes

In this work, we consider the stationary Fokker–Planck equation in the formulations

$$-\nabla\cdot(\kappa\nabla u)-\nabla\cdot(\kappa u\nabla V)=f, \tag{1.1}$$

$$\text{or}\qquad \operatorname{div}\mathbf{J}(u,V)=f \tag{1.2}$$

using the flux $\mathbf{J}(u,V)=-\kappa(\nabla u+u\nabla V)$, where $\kappa>0$ is a (possibly space-dependent) diffusion coefficient and $V:\Omega\to\mathbb{R}$ is a given potential. For simplicity of presentation, we study the case $\kappa\equiv 1$ but emphasize that the results also hold for $\kappa\in C^2(\overline{\Omega})$ with constants $0<\underline{\kappa}<\overline{\kappa}<\infty$ and $\underline{\kappa}\le\kappa\le\overline{\kappa}$. The flux $\mathbf{J}$ consists of a diffusive part $\kappa\nabla u$ and a drift part $\kappa u\nabla V$, which compensates for the stationary density $\pi=\mathrm{e}^{-V}$ (Boltzmann distribution) as $\mathbf{J}(\mathrm{e}^{-V},V)=0$. The right-hand side f describes possible sink or source terms.

In what follows, we assume for simplicity of presentation $\kappa\equiv 1$ but we emphasize that the calculations hold true also for the general case. However, super convergence of order 2 on cubic meshes only holds for $\kappa\equiv$ const.

Assumption 1.1. *Unless stated otherwise, we assume $\Omega\subset\mathbb{R}^d$ to be a polygonal convex bounded domain, $V\in C^2(\overline{\Omega})$, $f\in L^2(\Omega)$ real-valued functions. The standard boundary conditions in (1.1) are the homogeneous Dirichlet boundary conditions.*

The assumption $V\in C^2(\overline{\Omega})$ implies strict positivity $\pi>0$. Using a transformation $U=u/\pi$ we find that (1.1) is equivalent with

$$-\nabla\cdot(\pi\nabla U)=f. \tag{1.3}$$

Discretizing (1.3) on an admissible mesh in the sense of Definition 10.1 in Chapter 3 of [17] or in [24] we write $\mathcal{T}=(\mathcal{V},\mathcal{E},\mathcal{P})$ for the mesh consisting of convex polytope control volumes $\mathcal{V}:=\{\Omega_i,i=1,\dots,N\}$ with mass m_i,$(d-1)$-dimensional flat interfaces $\mathcal{E}_\Omega=\{\sigma_{i,j}\}$ with measure $m_{i,j}$ and points $\mathcal{P}_\Omega=\{x_i,i=1,\dots,N\}$ which we sometimes call the cell centers. Two cells Ω_i, Ω_j are neighbors if $\sigma_{i,j}:=\partial\Omega_i\cap\partial\Omega_j$ has positive measure and we write $i\sim j$. If $i\sim j$, the distance of the cell centers is $h_{i,j}:=|x_i-x_j|$.

In order to formulate discrete Dirichlet conditions, we follow [24] and enrich the mesh with finitely many points $\mathcal{P}_{\partial\Omega}=(y_k)_k\subset\partial\Omega$ and virtual interfaces $\mathcal{E}_{\partial\Omega}=\{\sigma_{i,k}\text{ flat}:\exists i\text{ with }\sigma_{i,k}\subset\partial\Omega\cap\partial\Omega_i\}$ *i.e.*, for every flat segment $\sigma_{i,k}\subset\partial\Omega\cap\partial\Omega_i$ we chose $y_k\in\sigma_{i,k}$ such that $(y_k-x_i)\perp\sigma_{i,k}$ and denote $m_{i,k}:=|\sigma_{i,k}|$ with $h_{i,k}:=|y_k-x_i|$. We further generalize the notation $i\sim j$ if $\sigma_{i,j}\subset\partial\Omega_i$ or $\sigma_{i,j}\subset\partial\Omega_j$. Then, when summing up over the interfaces in the calculations below, we do not have to distinguish between inner interface of type $\partial\Omega_i\cap\partial\Omega_j$ and outer interfaces of type $\partial\Omega\cap\partial\Omega_i$.

We finally denote $\mathcal{P}=\mathcal{P}_\Omega\cup\mathcal{P}_{\partial\Omega}$ and $\mathcal{E}=\mathcal{E}_\Omega\cup\mathcal{E}_{\partial\Omega}$ and write $\sum_{j:j\sim i}$ for the sum over all interfaces belonging to Ω_i and $\sum_{j\sim i}$ for the sum over all interfaces $\mathcal{E}$.

Remark 1.2. Since Ω_i are convex polytopes, the cones defined by x_i and $\sigma_{i,j}$ are mutually disjoint. Hence, writing $\mathcal{E}(x_i)=\{\sigma_{i,j}\in\mathcal{E}:\sigma_{i,j}\subset\Omega_i\}$ for $d\ge 2$ there exists a constant C_d depending only on the dimension

TABLE 1. Several mean values expressed as Stolarsky means $S_{\alpha,\beta}$ with corresponding weight functions $B_{\alpha,\beta}$, see equation (1.7). The geometric mean corresponds to the SQRA scheme, the $S_{0,-1}$-mean to the Scharfetter–Gummel discretization.

Mean	α	β	$\alpha+\beta$	$S_{\alpha,\beta}(x,y)$	$B_{\alpha,\beta}(x)$
Max	$+\infty$	1	$+\infty$	$\max(x,y)$	$\begin{cases}\mathrm{e}^{-x}, & x\le 0\\ 1, & x>0\end{cases}$
Quadratic mean	4	2	6	$\sqrt{\frac{1}{2}(x^2+y^2)}$	$\sqrt{\frac{1}{2}(1+\mathrm{e}^{-2x})}$
Arithmetic mean	2	1	3	$\frac{1}{2}(x+y)$	$\frac{1}{2}(1+\mathrm{e}^{-x})$
Logarithmic mean	1	0	1	$(x-y)/\log(x/y)$	$\frac{1}{x}(1-\mathrm{e}^{-x})$
Geometric mean (SQRA)	1	-1	0	$\sqrt{xy}$	$\mathrm{e}^{-x/2}$
Scharfetter–Gummel mean	0	-1	-1	$xy\log(x/y)/(x-y)$	$x/(\mathrm{e}^x-1)$
Harmonic Mean	-2	-1	-3	$2xy/(x+y)$	$2/(\mathrm{e}^x+1)$
Min	$-\infty$	1	$-\infty$	$\min(x,y)$	$\begin{cases}\mathrm{e}^{x}, & x\le 0\\ 1, & x>0\end{cases}$

such that

$$\forall i: \qquad C_d^{-1} m_i \le \sum_{\sigma_{i,j}\in\mathcal{E}(x_i)} m_{i,j} h_{i,j} \le C_d m_i.$$

Definition 1.3. Given a family of admissible meshes $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ we denote for $\Omega_i \in \mathcal{V}_h$ the diameter $h_i = \mathrm{diam}\Omega_i$. The family of meshes is called *quasi uniform* if for every $x_i, x_j \in \mathcal{P}_h$, $i \sim j$, it holds $h_{i,j} < h$ and if there exists $R, r > 0$ independent from $\mathcal{T}_h$ such that the following holds: For every $\Omega_i \in \mathcal{V}_h$ there exists $x \in \Omega_i$ such that $\mathbb{B}_{rh_i}(x) \subset \Omega_i \subset \mathbb{B}_{Rh_i}(x)$.

We make the following proposal for a discretization of (1.3)

$$\forall x_i \in \mathcal{P}_\Omega \qquad -\sum_{j:j\sim i} \frac{m_{i,j}}{h_{i,j}} S_{i,j}(U_{\mathcal{T},j} - U_{\mathcal{T},i}) = m_i f_{\mathcal{T},i}, \tag{1.4}$$

where $\pi_i = \mathrm{e}^{-V_i}$, $V_i = V(x_i)$ resp. $V_i = V(y_i)$, $f_{\mathcal{T},i} = \fint_{\Omega_i} f$ is the average of f over Ω_i, $\sum_{j:j\sim i}$ denotes the sum over all neighbors of cell i and $S_{i,j} = S_{\alpha,\beta}(\pi_i, \pi_j)$ is a Stolarsky mean of π_i and π_j [48]

$$S_{\alpha,\beta}(x,y) = \left(\frac{\beta(x^\alpha - y^\alpha)}{\alpha(x^\beta - y^\beta)}\right)^{\frac{1}{\alpha-\beta}}, \qquad \alpha\neq 0,\ \beta\neq 0,\ \alpha\neq\beta,\ x\neq y. \tag{1.5}$$

Stolarsky means can be extended to the critical points $\alpha = 0$, $\beta = 0$, $\alpha = \beta$, $x = y$ in a continuous way and generalize the logarithmic mean and other means, see Table 1.

It is well known that (1.3) has a unique solution $U \in H^2(\Omega) \cap H_0^1(\Omega)$ satisfying homogeneous Dirichlet boundary conditions and we demand that $U_{\mathcal{T}} : \mathcal{P} \to \mathbb{R}$ as a solution of (1.4) satisfies discrete Dirichlet boundary conditions. Then, the discrete linear operator in the above schemes is an M-matrix and (1.4) with discrete homogeneous Dirichlet conditions has a unique solution $U_{\mathcal{T}} : \mathcal{P} \to \mathbb{R}$.

Finally, we reverse the above discretization $U = u/\pi$ and obtain that $u_{\mathcal{T},i} := U_{\mathcal{T},i}\pi_i$ solves the *discrete FPE*

$$\forall x_i \in \mathcal{P}_\Omega \qquad -\sum_{j:j\sim i} \frac{m_{i,j}}{h_{i,j}} S_{i,j}\left(\frac{u_{\mathcal{T},j}}{\pi_j} - \frac{u_{\mathcal{T},i}}{\pi_i}\right) = m_i f_{\mathcal{T},i}, \tag{1.6}$$

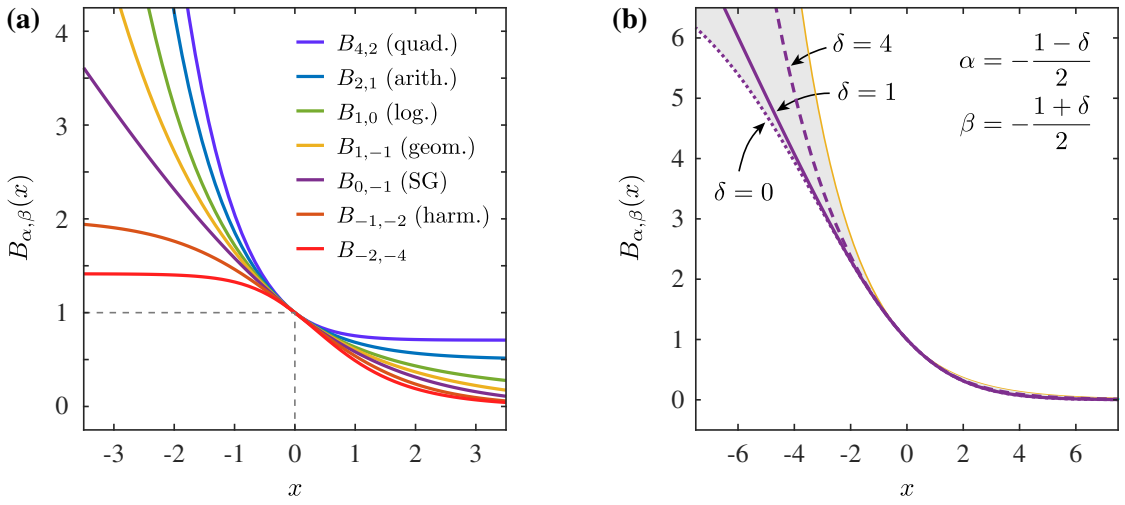

FIGURE 1. (a) Weight functions $B_{\alpha,\beta}$ of the discrete flux scheme for different Stolarsky means $S_{\alpha,\beta}$ according to equation (1.7), *cf.* Table 1. (b) Weight functions for $\alpha+\beta=-1$ using the parametrization $\alpha=-(1-\delta)/2$, $\beta=-(1+\delta)/2$ for $\delta\geq 0$. The SG-mean $(\alpha,\beta)=(0,-1)$ is obtained for $\delta=1$. The grey shaded region indicates the full range $\delta=[0,\infty)$, where the limit $\delta\to\infty$ is given by the weight function $\mathrm{e}^{-x/2}$ of the SQRA scheme.

To highlight the relation to the Scharfetter–Gummel (SG) scheme, we use the relation $S_{\alpha,\beta}(x,y) = x\,S_{\alpha,\beta}(1,y/x)$ and introduce the weight function

$$B_{\alpha,\beta}(x) = S_{\alpha,\beta}\big(1,\mathrm{e}^{-x}\big) \quad \text{with} \quad B_{\alpha,\beta}(-x) = \mathrm{e}^{x} B_{\alpha,\beta}(x), \tag{1.7}$$

such that equation (1.6) can equally be reformulated as

$$-\sum_{j:i\sim j} \frac{m_{i,j}}{h_{i,j}} \big(B_{\alpha,\beta}(V_i-V_j)u_j - B_{\alpha,\beta}(V_j-V_i)u_i\big) = m_i f_{\mathcal{T},i}.$$

Two special cases of particular interest are

$$B_{0,-1}(V_i-V_j) = \frac{V_i-V_j}{\mathrm{e}^{V_i-V_j}-1} = S_{0,-1}(\pi_i,\pi_j)\pi_j^{-1}, \tag{1.8}$$

$$B_{1,-1}(V_i-V_j) = \mathrm{e}^{-\frac{1}{2}(V_i-V_j)} = S_{1,-1}(\pi_i,\pi_j)\pi_j^{-1}. \tag{1.9}$$

With regard to Table 1, these coefficients are known as the Bernoulli function $B_{0,-1}$ (for SG) and the SQRA-coefficient $B_{1,-1}$. FV schemes with general weight functions B have been investigated in [7, 35] (B-schemes). We emphasize here that the case $\alpha=0$ and $\beta=-1$ is indeed very special in the analysis but all schemes with $\alpha+\beta=-1$ behave very similar.

To make this more clear, we write $J_{i,j} := \frac{S_{i,j}}{h_{i,j}}\Big(\frac{u_{\mathcal{T},j}}{\pi_j} - \frac{u_{\mathcal{T},i}}{\pi_i}\Big)$. In the diffusion regime $V\approx$ const we observe for fixed α, β that $B_{\alpha,\beta}(V_i-V_j)\approx \mathrm{e}^0 S_{\alpha,\beta}(1,1)=1$ and $J_{i,j}\approx (u_i-u_j)/h_{i,j}$ is reduced to the discrete diffusive flux (Fig. 1).

In the drift-dominated regime, *i.e.*, for $|V_j-V_i|\gg 1$, the various $B_{\alpha,\beta}$ behave differently. While $B_{1,-1}(V_i-V_j)$ cannot be controlled in a reasonable way, we may introduce

$$J^{\infty}_{i,j} := -\frac{V_j-V_i}{h_{i,j}}\begin{cases} u_j & \text{if } V_j > V_i \\ u_i & \text{if } V_j < V_i \end{cases}$$

which is a robust discretization of the drift part of the flux, with u being evaluated in the donor cell of the flux. For $\alpha = 0$, $\beta = -1$ we recover the upwind scheme, *i.e.*,

$$\left|J^{\infty}_{i,j}\right|^{-1}\left|J_{i,j} - J^{\infty}_{i,j}\right| \to 0 \qquad \text{as} \quad |V_j - V_i| \to \infty. \tag{1.10}$$

Hence, the Bernoulli function $B_{0,-1}$ interpolates between the appropriate discretizations for the drift- and diffusion-dominated limits, which is why the SG scheme is the preferred FV scheme for Fokker–Planck type operators. Mathematically, this is formulated in Section 5.2.

For convenience, we often write $O(h^k)$ which means

$$\text{there exists a constant } C \text{ not depending on } h, \mathcal{T}, \pi \text{ s.t. } \left|O(h^k)\right| \le C|h|^k \text{ for } h \to 0. \tag{1.11}$$

In this introduction, C in (1.11) also depends on $\|f\|_{L^2}$ and $\|\pi\|_{C^2}$. In recent years, convergence order has been derived for many different schemes. In [32], quantitative convergence of order $O(h^2)$ for several upwind schemes on rectangular grids has been shown. In [2] the finite volume Scharfetter–Gummel discretization (of steady convection diffusion equations) is connected to a finite element method and convergence of order $O(h)$ is obtained by using results from [53]. Investigating general B-schemes, Chainais-Hillairet and Droniou [7] proved strong convergence in L^2 for the solutions of the FV scheme to the continuous solution. Recently, convergence of order $O(h)$ for general B-schemes including SG, SQRA as well as Stolarsky means has been proved in 1D [35]. Independently, convergence for the SQRA discretization has been investigated in [40] in 1D, Donati *et al.* [14] (formally, rectangular meshes) and [26] using G-convergence on grids with random weights.

1.2. Major contributions of this work

We derive the order of convergence in the energy norm for general Stolarsky schemes benefiting from analytical properties of Stolarsky means and using consistency theory in the sense of [11]. The error naturally splits into the consistency error for the discretization of the Laplace operator (the consistency of the elliptic operator) plus an error which is due to the convective part. Clearly, we have the possibility to study the error in terms of U and of u. While the error in terms of U can be directly inferred from the diffusive estimate in Lemma 2.6, one can also apply a splitting into diffusion- and convection-part of the error in terms of u. The order of convergence is in general limited by the consistency of the mesh but can be improved up to order $O(h)$ in u (on all Voronoï grids), resp. $O(h^2)$ in U (on cubic grids). It is interesting to observe that the optimal Stolarsky mean can be different in the variables u and U for the same problem on the same mesh. This is indicated by the numerical experiment of Example 7.2.

The choice of the Stolarsky mean does basically not affect the rate of convergence in the variable U. However, in the variable u the scheme $S_{0,-1}$ (SG scheme) is special among all schemes as the additional error term which is solely due to the convection and not the consistency of the grid is of order $O(h^2)$ (Thm. 1.5), compared to $O(h)$ for the other schemes. Due to a perturbation result (Cor. 4.2), the good convergence properties of the SG scheme carry over to every Stolarsky scheme where $\alpha + \beta = -1$. However, in our 1-dimensional sample calculations, this effect cannot be seen due to part 2 of Theorem 5.2 and the relation (1.16) below. On the other hand, extensive 2d or 3d studies are beyond the scope of this work.

In what follows, we denote $L^2(\mathcal{P}) := \{U : \mathcal{P}_\Omega \to \mathbb{R}\}$ and $H_\mathcal{T} := \{U : \mathcal{P} \to \mathbb{R} \mid U|_{\mathcal{P}_{\partial\Omega}} \equiv 0\}$ with the natural imbedding $H_\mathcal{T} \hookrightarrow L^2(\mathcal{P})$ and introduce the norms

$$\forall \tilde{v} \in L^2(\mathcal{P}),\, v \in H_{\cdot\mathcal{T}}: \qquad \|v\|^2_{H_\mathcal{T}} := \sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}}(v_j - v_i)^2, \qquad \|\tilde{v}\|^2_{L^2(\mathcal{P})} := \sum_{\Omega_i} m_i \tilde{v}_i^2. \tag{1.12}$$

We recall the discrete Poincaré inequality for zero boundary values ([17], Sect. 10.2)

$$\forall v \in H_\mathcal{T}: \qquad \|v\|_{L^2(\mathcal{P})} \le C\|v\|_{H_\mathcal{T}}, \tag{1.13}$$

where C depends only on Ω but not on $\mathcal{T}$. Now, the discrete Poincaré inequality applied to (1.4) implies that the solution $U_\mathcal{T}$ satisfies uniformly the discrete *a priori* estimate

$$\|U_\mathcal{T}\|_{H_\mathcal{T}} \leq C\|f_\mathcal{T}\|_{L^2(\mathcal{P}_\Omega)}. \tag{1.14}$$

Of course we cannot expect solutions of (1.4) to approximate solutions of (1.3) better than in the case $\pi = \text{const}$. We will acknowledge this in terms of φ-consistency in Definition 2.8 below in a rigorous way. For the moment, the reader may refer to $\varphi(h)\|U\|_{H^2(\Omega)}$ as the speed of convergence of the scheme (1.4) for constant $\pi = 1$ and continuous solution $U \in H^2(\Omega)$. Then, as we will see, we can express the general order of convergence for $u_\mathcal{T}$ (sol. of (1.6)) towards u (sol of (1.1)) in terms of φ plus some correcting terms.

Finally, given $u \in H^2(\Omega)$ in the following we set $(\mathcal{R}_\mathcal{T} u)(y_j) = u(y_j)$ for $y_j \in \mathcal{P}_{\partial\Omega}$ and

$$(\mathcal{R}_\mathcal{T} u)_i := u(x_i) \quad \text{is the pointwise evaluation of } u \text{ in the centers of the cells } \Omega_i.$$

As a consequence of the Poincaré inequality (1.13) we can transfer order of convergence estimates on $u_\mathcal{T}$ directly to $U_\mathcal{T}$ and *vice versa* through the following relations: for $\tilde{U} \in H_\mathcal{T}$ with $\tilde{u} = \tilde{U}\mathcal{R}_\mathcal{T}\pi$

$$\tilde{U}_j - \tilde{U}_i = \frac{\tilde{u}_j}{\pi_j} - \frac{\tilde{u}_i}{\pi_i} = \frac{1}{2}(\tilde{u}_j - \tilde{u}_i)\left(\frac{1}{\pi_j} + \frac{1}{\pi_i}\right) + \frac{1}{2}(\tilde{u}_j + \tilde{u}_i)\left(\frac{1}{\pi_j} - \frac{1}{\pi_i}\right), \tag{1.15}$$

$$\tilde{u}_j - \tilde{u}_i = \pi_j\tilde{U}_j - \pi_i\tilde{U}_i = \frac{1}{2}\left(\tilde{U}_j - \tilde{U}_i\right)(\pi_j + \pi_i) + \frac{1}{2}\left(\tilde{U}_j + \tilde{U}_i\right)(\pi_j - \pi_i), \tag{1.16}$$

which imply with help of Remark 1.2 that:

$$\left\|\tilde{U}\right\|_{H_\mathcal{T}} \leq \left\|\pi^{-1}\right\|_\infty \|\tilde{u}\|_{H_\mathcal{T}} + 2C_d\left\|\nabla\pi^{-1}\right\|_\infty \|\tilde{u}\|_{L^2(\mathcal{P})}, \tag{1.17}$$

$$\|\tilde{u}\|_{H_\mathcal{T}} \leq \|\pi\|_\infty \left\|\tilde{U}\right\|_{H_\mathcal{T}} + 2C_d\|\nabla\pi\|_\infty \left\|\tilde{U}\right\|_{L^2(\mathcal{P})}. \tag{1.18}$$

Hence it is, in principle, sufficient to derive order of convergence estimates in U for (1.4) and transfering these estimates to the solutions of (1.6).

Theorem 1.4. *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes and let V satisfy Assumption 1.1. Moreover, let $\mathcal{T}_h$ be φ-consistent (Def. 2.8). If $U \in H^2(\Omega) \cap H^1_0(\Omega)$ is the solution of (1.3) and $U_{\mathcal{T}_h}$ the solution of (1.4) with discrete homogeneous Dirichlet boundary conditions then*

$$\|U_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} U\|^2_{H_{\mathcal{T}_h}} \leq C_1\|\pi\|^2_\infty \varphi(h)^2 + C_2 h^k,$$

where $k = 2$ in general and $k = 4$ if the grid is cubic or $d = 1$. Here, C_1 and C_2 depend only on d and Ω, r and R.

We provide a proof of Theorem 1.4 at the end of Section 6.

The convergence rate of U is directly related to the convergence rate of the flux since $\mathbf{J}(u, V) = -\pi\nabla U$. Through (1.18) and the Poincaré inequality, we also get a rate of convergence for the variable u. However, the second term on the right hand side of (1.16) is non-local since the $L^2(\mathcal{P})$-norm is controlled by the $H_\mathcal{T}$-norm. Furthermore, we want to track the influence of the choices of the Stolarsky mean onto the quality of convergence of $u_\mathcal{T}$. Also the proof of convergence of $U_\mathcal{T}$ cannot explain the observed supremacy of the SG coefficient. Therefore, we spend some effort in direct calculations based on u, despite the calculations in U are much easier. The result is the following

Theorem 1.5. *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes and let V satisfy Assumption 1.1. Moreover, let $\mathcal{T}_h$ be φ-consistent (Def. 2.8). If $u \in H^2(\Omega) \cap H^1_0(\Omega)$ is the solution of (1.1) and $u_{\mathcal{T}_h}$ the solution of (1.6) with discrete homogeneous Dirichlet boundary conditions then*

$$\|u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} u\|^2_{H_{\mathcal{T}_h}} \leq C_1\left(\|u\|^2_{H^2} + \|u\|^2_\infty \|V\|^2_{H^2}\right)\varphi(h)^2 + C_2 h^k,$$

where $k = 2$ in general and $k = 4$ if $\alpha + \beta = -1$ and where C_1 depends on Ω,d, r and R and C_2 additionally depends on $\|V\|_{C^2}$ and $\|u\|_{H^2}$.

We provide a proof of Theorem 1.5 at the end of Section 5.2.

This theorem compared to Theorem 1.4 shows that for non-cubic grids, the "higher order" error, which is not due to the grid consistency, can be of one order smaller for the SG scheme than for the other schemes. This is an information which was not possible to obtain from the considerations in U.

Remark 1.6. As a consequence of former works (see Props. 2.9 and 2.10) it holds $\varphi(h) = |h|$ on Voronoï grids and $\varphi(h) = h^2$ on cubic grids. This explains the next result.

Theorem 1.7. *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a sequence of cubic grids $h\mathbb{Z}^d$ and let V satisfy Assumption 1.1. If $u \in H^2(\Omega) \cap H_0^1(\Omega)$ is the solution of (1.1) and $u_{\mathcal{T}_h}$ the solution of (1.6) with discrete homogeneous Dirichlet boundary conditions then*

$$\|u_{\mathcal{T}_h} - \mathcal{R}_{\mathcal{T}_h} u\|^2_{H_{\mathcal{T}_h}} \leq C h^k,$$

where $k = 2$ in general and $k = 4$ if $\alpha + \beta = -1$ and where C depends on on Ω, d, $\|V\|_{C^2}$ and $\|u\|_{H^2}$.

We provide a proof of Theorem 1.7 at the end of Section 6.

We note at this point, that these estimates are only "worst case" estimates, while the true rate of convergence could also be better. In Section 4 we will see that the rate of convergence is close for different Stolarsky schemes that share the same value of $\alpha+\beta$. *i.e.*, the difference in the error due to switching $S_{\alpha,\beta}$ with $S_{\tilde{\alpha},\tilde{\beta}}$ is of order h^3 if $\tilde{\alpha} + \tilde{\beta} = \alpha + \beta$, see Proposition 4.1. This explains the shape of the error graphs in Figures 2a, 2c and 3a, 3c.

Furthermore, we observe that in case $d = 1$ we always get order 2 convergence.

Although we treat the Stolarsky means as an explicit example, note that some of the main results also hold for other smooth means.

1.3. Outline of this work

We recall the consistency theory of [11] in Section 2 and afterwards present two different points of view on the derivation of the above numerical scheme in Section 3.

The mathematical investigation starts in Section 4. In Section 4 we observe that the variational consistency error is close for two different Stolarsky means when they share the same value for $\alpha+\beta$. In Section 5 we prove Theorem 1.5 in a version that uses the language of the variational consistency error and in Section 6 we do the same with Theorem 1.7.

Finally, our main results are illustrated in Section 7 by numerical simulations.

2. Consistency and inf-sup stability

We use the framework of [11] and consider on an admissible mesh $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ the space

$$H_{\mathcal{T}} := \{v \in L^2(\mathcal{P}) : v \text{ satisfies hom. Dir. b.c.}\}$$

with the norm $\|\cdot\|_{H_{\mathcal{T}}}$ given in (1.12). We sometimes later also use the following notation for some positive coefficient field Ω on $\mathcal{E}$:

$$\|v\|_{H_{\mathcal{T}},\omega} := \left(\sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} \omega_{i,j} (v_j - v_i)^2\right)^{\frac{1}{2}}.$$

Definition 2.1 (inf-sup stability)**.** Let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes. A family of bilinear forms a_h on $H_{\mathcal{T}_h}$ is called *uniformly inf-sup stable* with respect to two norms $\|\cdot\|_{h,1}$, $\|\cdot\|_{h,2}$ if there exists $\gamma > 0$ (independent from h) such that

$$\forall u \in H_{\mathcal{T}_h} : \quad \gamma \|u\|_{h,1} \leq \sup_{v \in H_{\mathcal{T}_h}} \frac{a_h(u,v)}{\|v\|_{h,2}}.$$

Throughout this paper we write $\mathcal{R}_h := \mathcal{R}_{\mathcal{T}_h}$ for simplicity and sometimes $u_i = (\mathcal{R}_{\mathcal{T}} u)_i$ if the meaning is clear and no confusion with discrete variables occurs. For a continuous and coercive bilinear form $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$, the associated linear operator $A : H^2(\Omega) \to L^2(\Omega)$ is defined by

$$\forall u \in H^2(\Omega) \cap H_0^1(\Omega),\, v \in H_0^1(\Omega) : \quad a(u,v) = \int_\Omega v\, Au. \tag{2.1}$$

Definition 2.2 (Consistency)**.** Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be bilinear and continuous with linear operator A such that (2.1) holds and let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a family of admissible meshes with $a_h : H_{\mathcal{T}_h} \times H_{\mathcal{T}_h} \to \mathbb{R}$ continuous bilinear forms. The *variational consistency error* of a_h in $u \in H^2(\Omega) \cap H_0^1(\Omega)$ is the linear form $\mathfrak{E}_h(u;\cdot) : H_{\mathcal{T}_h} \to \mathbb{R}$ where

$$\forall v \in H_{\mathcal{T}_h} : \quad \mathfrak{E}_h(u;v) := \sum_i v_i \int_{\Omega_i} Au - a_h(\mathcal{R}_h u,\, v). \tag{2.2}$$

We say for the norm $\|\cdot\|_{h,2}$ on $H_{\mathcal{T}_h}$ and $u \in H^2(\Omega) \cap H_0^1(\Omega)$ that *consistency holds* if

$$\|\mathfrak{E}_h(u;\cdot)\|_{h,2,*} := \sup_{v \in H_{\mathcal{T}_h} \setminus \{0\}} \frac{|\mathfrak{E}_h(u;v)|}{\|v\|_{h,2}} \to 0 \quad \text{as} \quad h \to 0.$$

Proposition 2.3 ([11], Thm. 10)**.** *Let $a : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R}$ be bilinear and continuous with A such that (2.1) holds and let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a family of admissible meshes with $a_h : H_{\mathcal{T}_h} \times H_{\mathcal{T}_h} \to \mathbb{R}$ bilinear and uniformly inf-sup stable forms. If $u \in H^2(\Omega) \cap H_0^1(\Omega)$ and $u_h \in H_{\mathcal{T}_h}$ are solutions to*

$$\forall v \in H_0^1(\Omega) : a(u,v) = \int f\, v = \int Au\, v; \qquad \forall v \in H_{\mathcal{T}_h} : a_h(u_h, v) = \sum_i v_i \int_{\Omega_i} Au,$$

then it holds

$$\|u_h - \mathcal{R}_h u\|_{h,1} \le \gamma^{-1} \|\mathfrak{E}_h(u;\cdot)\|_{h,2,*}. \tag{2.3}$$

Using the above general insights, we introduce for $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ the bilinear forms

$$a_{\mathrm{FPE}} : H_0^1(\Omega) \times H_0^1(\Omega) \to \mathbb{R} \qquad (u,v) \mapsto \int_\Omega \nabla u \cdot \nabla v + u \nabla V \cdot \nabla v,$$

$$a_{h,\mathrm{FPE}}(u,v) : H_{\mathcal{T}_h} \times H_{\mathcal{T}_h} \to \mathbb{R} \qquad (u,v) \mapsto \sum_{i \sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \left(\frac{u_j}{\pi_j} - \frac{u_i}{\pi_i} \right) (v_j - v_i),$$

with $A_{\mathrm{FPE}}\, u := -\nabla \cdot (\nabla u + u \nabla V)$ and following (2.2)

$$\forall v \in H_{\mathcal{T}_h} : \quad \mathfrak{E}_{h,\mathrm{FPE}}(u;v) := \sum_i v_i \int_{\Omega_i} A_{\mathrm{FPE}}\, u - a_{h,\mathrm{FPE}}(\mathcal{R}_h u,\, v). \tag{2.4}$$

Lemma 2.4. *Under the Assumption 1.1 let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes. Then $a_{h,\mathrm{FPE}}$ is uniformly inf-sup stable for $\|\cdot\|_{h,1} = \|\cdot\|_{h,2} = \|\cdot\|_{H_{\mathcal{T}_h}}$, where γ depends on Ω, $\inf|\pi|$, $\|\pi\|_\infty$ and $\|\nabla \pi\|_\infty$.*

Remark 2.5. We could also consider inf-sup stability of $a_{h,\mathrm{FPE}}$ for $\|u\|_{h,1} := \left(\sum_{i \sim j} \frac{m_{i,j}}{h_{i,j}} \left(\frac{u_j}{\pi_j} - \frac{u_i}{\pi_i} \right)^2 \right)^{\frac{1}{2}}$ and $\|\cdot\|_{h,2} = \|\cdot\|_{H_{\mathcal{T}_h}}$. This will first lead to an estimate of $\|U_{\mathcal{T}} - \mathcal{R}_{\mathcal{T}} U\|_{H_{\mathcal{T}_h}}$ which then has to be transformed into one on $\|u_{\mathcal{T}} - \mathcal{R}_{\mathcal{T}} u\|_{H_{\mathcal{T}_h}}$ using (1.17) and the Poincaré inequality. However, the speed of convergence for U and

u may be different, which is why we study both separately in Sections 5.1 and 5.2. Hence in what follows, we will always consider for $u_{\mathcal{T}} \in H_{\mathcal{T}_h}$ and $\mathfrak{E} \in H^*_{\mathcal{T}_h}$

$$\|\cdot\|_{h,1} = \|\cdot\|_{h,2} = \|\cdot\|_{H_{\mathcal{T}_h}}, \qquad \|\mathfrak{E}\|_{H^*_{\mathcal{T}_h}} := \sup_{v \in H_{\mathcal{T}_h}\setminus\{0\}} \|v\|^{-1}_{H_{\mathcal{T}_h}} |\mathfrak{E}(v)|.$$

Proof of Lemma 2.4. Introducing $\bar{u}_{i,j} := \frac{1}{2}(u_i + u_j)$ and using (1.15), we obtain with the triangle inequality

$$2\sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j}\Big((U_j - U_i)(U_j - U_i) + \bar{u}^2_{i,j}\big(\pi_j^{-1} - \pi_i^{-1}\big)^2\Big) \geq \sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \frac{1}{4}\left(\frac{(\pi_i + \pi_j)}{\pi_i \pi_j}\right)^2 (u_j - u_i)^2. \tag{2.5}$$

Observing that

$$\sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \bar{u}^2_{i,j}\big(\pi_j^{-1} - \pi_i^{-1}\big)^2 \leq 2\sum_i U_i^2 \sum_{j:j\sim i} \frac{m_{i,j}}{h_{i,j}} S_{i,j} \pi_i^2 \big(\pi_j^{-1} - \pi_i^{-1}\big)^2$$

and exploiting the discrete Poincaré inequality we observe that

$$\sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j}(U_j - U_i)(U_j - U_i) \geq C\|u\|^2_{H_{\mathcal{T}_h}},$$

where C depends on Ω, $0 < r < R$, $\inf|\pi|$, $\|\pi\|_\infty$ and $\|\nabla\pi\|_\infty$. On the other hand

$$\sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j}(U_j - U_i)(U_j - U_i) = a_{h,\mathrm{FPE}}(u,U) \leq \sup_{v\in H_{\mathcal{T}_h}} \frac{a_{h,\mathrm{FPE}}(u,v)}{\|v\|_{H_{\mathcal{T}_h}}} \|U\|_{H_{\mathcal{T}_h}},$$

which together with (1.17) and (2.5) and the Poincaré inequality implies uniform inf-sup stability. □

Next we derive an estimate for $\|\mathfrak{E}_{h,\mathrm{FPE}}(u;\cdot)\|_{H^*_{\mathcal{T}_h}}$. We introduce the diffusive part

$$a_{\mathrm{D}}(u,v) = \int_\Omega \nabla u \cdot \nabla v, \qquad a_{h,\mathrm{D}}(u,v) = \sum_{i\sim j} \frac{m_{i,j}}{h_{i,j}}(u_j - u_i)(v_j - v_i),$$

with $A_{\mathrm{D}}u := -\nabla\cdot(\nabla u)$ and $\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)$ according to (2.2).

Lemma 2.6. *Under the Assumption 1.1 let $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes. It holds*

$$\mathfrak{E}_{h,\mathrm{FPE}}(u;v) = \mathfrak{E}_{h,\mathrm{D}}(u;v) + \mathfrak{E}_{h,\mathrm{conv}}(u;v), \tag{2.6}$$

where the convective part of the consistency error is given by

$$\mathfrak{E}_{h,\mathrm{conv}}(u;v) = \sum_{i\sim j}(v_j - v_i)\left(\int_{\sigma_{i,j}} u\nabla V\cdot\boldsymbol{\nu}_{i,j} - \frac{m_{i,j}}{h_{i,j}}\left(\frac{S_{i,j}-\pi_j}{\pi_j}u_j - \frac{S_{i,j}-\pi_i}{\pi_i}u_i\right)\right). \tag{2.7}$$

In particular, we obtain

$$\|\mathfrak{E}_{h,\mathrm{FPE}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}} \leq 2\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}} + 2\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}},$$

$$\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}} \leq \sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}\nabla u\cdot\boldsymbol{\nu}_{i,j} - \frac{m_{i,j}}{h_{i,j}}\big((\mathcal{R}_h u)_j - (\mathcal{R}_h u)_i\big)\right)^2, \tag{2.8}$$

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}} := \sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}} u\nabla V\cdot\boldsymbol{\nu}_{i,j} - \frac{m_{i,j}}{h_{i,j}}\left(\frac{S_{i,j}-\pi_j}{\pi_j}u_j - \frac{S_{i,j}-\pi_i}{\pi_i}u_i\right)\right)^2. \tag{2.9}$$

We remark that estimate (2.8) was explicitly proved in [11].

Proof. In what follows, we combine ideas of the proofs of Theorem 27 and 33 in [11]. However, since our grid and our coefficients have a simple structure, our calculations are much shorter. Then we obtain

$$S_{i,j}\left(\frac{u_j}{\pi_j}-\frac{u_i}{\pi_i}\right)=(u_j-u_i)+\left(\frac{S_{i,j}-\pi_j}{\pi_j}u_j-\frac{S_{i,j}-\pi_i}{\pi_i}u_i\right)$$

and hence

$$\begin{aligned}\mathfrak{E}_{h,\mathrm{FPE}}(u;v)=&\sum_{i\sim j}(v_j-v_i)\left(\int_{\sigma_{i,j}}\nabla u\cdot\boldsymbol{\nu}_{i,j}-\frac{m_{i,j}}{h_{i,j}}(u_j-u_i)\right)\\&+\sum_{i\sim j}(v_j-v_i)\left(\int_{\sigma_{i,j}}u\nabla V\cdot\boldsymbol{\nu}_{i,j}-\frac{m_{i,j}}{h_{i,j}}\left(\frac{S_{i,j}-\pi_j}{\pi_j}u_j-\frac{S_{i,j}-\pi_i}{\pi_i}u_i\right)\right).\end{aligned}$$

From here we conclude by the definition of $\mathfrak{E}_h(u;\cdot)$ and a direct calculation. □

A particular focus of the calculations below will lie on the following structure. For a functions $g\in C(\overline{\Omega})$ and $g^{\mathcal{T}}:\mathcal{E}_h\to\mathbb{R}$ with $g^{\mathcal{T}}(\sigma_{i,j})=g_{i,j}=g_{j,i}$, we introduce

$$a_g(u,v)=\int_\Omega\nabla u\cdot g\nabla v,\qquad a_{h,g}(u,v)=\sum_{i\sim j}\frac{m_{i,j}}{h_{i,j}}g_{i,j}(u_j-u_i)(v_j-v_i),$$

with $A_gu:=-\nabla\cdot(g\nabla u)$ and $\mathfrak{E}_{h,g}(u;\cdot)$ accordingly.

Lemma 2.7. *Let $\mathcal{T}=(\mathcal{V},\mathcal{E},\mathcal{P})$ be a mesh and $d\le 3$. Let $g\in C(\overline{\Omega})$ and let $g^{\mathcal{T}}\in\mathcal{E}^*$ with $g^{\mathcal{T}}(\sigma_{i,j})=g_{i,j}=g_{j,i}$. Then for every $u\in H^2(\Omega)$ it holds*

$$\|\mathfrak{E}_{h,g}(u;\cdot)\|^2_{H^*_{\mathcal{T}}}\le\left(\sup_{i,j}|g_{i,j}|\right)\|\mathfrak{E}_h(u;\cdot)\|^2_{H^*_{\mathcal{T}}}+\sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}(g-g_{i,j})\nabla u\cdot\boldsymbol{\nu}_{i,j}\right)^2.$$

Proof. This follows from decomposing $g\nabla u\cdot\boldsymbol{\nu}_{i,j}=g_{i,j}\nabla u\cdot\boldsymbol{\nu}_{i,j}+(g-g_{i,j})\nabla u\cdot\boldsymbol{\nu}_{i,j}$ on $\sigma_{i,j}$. □

With regard to (2.3), the above considerations motivate the following definition.

Definition 2.8 (φ-consistency)**.** Let $\mathcal{T}_h=(\mathcal{V}_h,\mathcal{E}_h,\mathcal{P}_h)$ be a quasi uniform family of admissible meshes. We say that $\mathcal{T}_h$ is *φ-consistent* if for every $u\in H^2(\Omega)\cap H^1_0(\Omega)$ there exists $C\ge 0$ such that for every $h>0$

$$\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|_{H^*_{\mathcal{T}_h}}\le C\|u\|_{H^2}\varphi(h).$$

Hence, we immediately obtain the following.

Proposition 2.9 (A consistency result [24])**.** *Let $d\le 3$ and $\mathcal{T}_h=(\mathcal{V}_h,\mathcal{E}_h,\mathcal{P}_h)$ be a family of admissible meshes. Then $(\mathcal{T}_h)_{h>0}$ is φ-consistent with $\varphi(h)=h$, i.e.,*

$$\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|_{H^*_{\mathcal{T}_h}}\le C\|u\|_{H^2}h,$$

where C depends only on Ω and d. We say that the mesh is h-consistent.

Proposition 2.10 ([50])**.** *Let $d\le 3$ and let the mesh $\mathcal{T}_h$ be cubic with all cubes of equal size h. Then the family of meshes is h^2-consistent with $\varphi(h)=h^2$, i.e.,*

$$\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|_{H^*_{\mathcal{T}_h,1}}\le C\|u\|_{H^2}h^2.$$

The following proposition explains why we can expect order 2 convergence in the 1d case when the grid is Voronoi.

Proposition 2.11. *Let $d = 1$ such that $\sigma_{i,j} = x_{i,j}$ is one single point. Let then the mesh $\mathcal{T}_h$ be Voronoi, i.e., $|x_i - x_{i,j}| = |x_j - x_{i,j}|$ for every neighboring pair i, j. Then*

$$\|\mathfrak{E}_{h,\mathrm{D}}(u;\cdot)\|^2_{H^*_{\mathcal{T}_h}} \le C \begin{cases} \|u\|^2_{H^2} \sum_{i\sim j} h^4_{i,j} & \text{if } u \in H^2(\Omega) \\ \|u\|^2_{C^2} \sum_{i\sim j} h^5_{i,j} & \text{if } u \in C^2(\overline{\Omega}) \end{cases}, \tag{2.10}$$

where $C > 0$ depends only on Ω, d. Furthermore,

$$\sum_{i\sim j} \frac{h_{i,j}}{m_{i,j}} \left(\int_{\sigma_{i,j}} (g - g_{i,j}) \nabla u \cdot \boldsymbol{\nu}_{i,j} \right)^2 \le C \|u\|^2_{H^2} \|g\|^2_{H^2} \sum_{i\sim j} h^5_{i,j} \tag{2.11}$$

provided $g \in H^2(\Omega)$ and $g_{i,j} = S_{\alpha,\beta}(g_i, g_j)$ for some α, β and for $g_i = (\mathcal{R}_{\mathcal{T}} g)_i$, $g_j = (\mathcal{R}_{\mathcal{T}} g)_j$ and where $S_{\alpha,\beta}$ is twice boundedly differentiable on $R_g := \{(g(x), g(y)) \in \mathbb{R}^2 : x, y \in \overline{\Omega}\}$. The constant C then depends on $\|S_{\alpha,\beta}\|_{C^2(\overline{R_g})}$.

Remark. We observe that $S_{\alpha,\beta}$ is not twice differentiable in $(0,0)$ for most choices of α and β. However, for $\alpha = 2$, $\beta = 1$ this is the case.

Proof. Throughout this proof, $C(\alpha, \beta, R_g)$ is a constant changing from line to line depending only on (α, β, R_g). Note that $m_{i,j} = 1$ and $\sigma_{i,j} = x_{i,j}$ consists of one single point. Furthermore, since $S_{\alpha,\beta}$ is twice boundedly differentiable on R_g we find $\sup_{i,j} \left| D^2 S_{\alpha,\beta}(g_i, g_j) \right| \le C(\alpha, \beta, R_g)$.

Step 1. We first prove the second statement. Since

$$g_i - g(x_{i,j}) = g'(x_{i,j}) \cdot (x_i - x_{i,j}) + \int_0^{x_i - x_{i,j}} s^2 g''(s + x_{i,j}) \, \mathrm{d}s \tag{2.12}$$

and $x_i - x_{i,j} = x_{i,j} - x_j$ it follows

$$\left| \frac{1}{2}(g_i - g(x_{i,j})) + \frac{1}{2}(g_j - g(x_{i,j})) \right| \le \left| \int_{-h_{i,j}/2}^{h_{i,j}/2} s^2 g''(s + x_{i,j}) \, \mathrm{d}s \right|. \tag{2.13}$$

Furthermore, since $g_{i,j} = S_{\alpha,\beta}(g_i, g_j)$ and $g(x_{i,j}) = S_{\alpha,\beta}(g(x_{i,j}), g(x_{i,j}))$ and for every $x > 0$ it holds $\partial_x S_{\alpha,\beta}(x,x) = \partial_y S_{\alpha,\beta}(x,x) = \frac{1}{2}$ and $S_{\alpha,\beta}$ is twice boundedly differentiable on R_g we find from Taylors formula for some constant $C(\alpha, \beta, R_g)$

$$\left| g_{i,j} - g(x_{i,j}) - \frac{1}{2}(g_i - g(x_{i,j})) - \frac{1}{2}(g_j - g(x_{i,j})) \right| \le C(\alpha, \beta, R_g) \left| \begin{pmatrix} g_i - g(x_{i,j}) \\ g_j - g(x_{i,j}) \end{pmatrix} \right|^2, \tag{2.14}$$

From (2.12) to (2.14) we conclude (2.11) using $\sigma_{i,j} = x_{i,j}$ and

$$|g(x_{i,j}) - g_{i,j}| \le \left| g_{i,j} - g(x_{i,j}) - \frac{1}{2}(g_i - g(x_{i,j})) - \frac{1}{2}(g_j - g(x_{i,j})) \right| + \left| \frac{1}{2}(g_i - g(x_{i,j})) + \frac{1}{2}(g_j - g(x_{i,j})) \right|.$$

Step 2. In view of (2.8) the first statement follows from

$$\left| \nabla u \cdot \boldsymbol{\nu}_{i,j} - \frac{1}{h_{i,j}} \Big((\mathcal{R}_h u)_j - (\mathcal{R}_h u)_i \Big) \right| \le h^{-1}_{i,j} \int_{-h_{i,j}/2}^{h_{i,j}/2} s^2 u''(s + x_{i,j}) \, \mathrm{d}s$$

and a similar argument as in Step 1.

□

3. Derivation of the methods and heuristic comparison

In this section, we repeat the original derivation of the Scharfetter–Gummel scheme in a more general way and show that both the SG and the SQRA scheme are members of a huge family of discretization schemes. Then we provide a physically motivated derivation of the SQRA scheme which assigns the SQRA a special place in the family of Stolarsky discretizations.

3.1. A family of discretization schemes

In one dimension, the Scharfetter–Gummel scheme for the discrete flux on the interval $[0,h]$ is derived in [47] under the assumption of constant flux J and constant diffusion coefficient κ on $[0,h]$. In particular, we consider the two-point boundary value problem

$$J = -\kappa(u'(x) + u(x)V'(x)) \quad \text{on } [0,h], \qquad u(0) = u_0, \quad u(h) = u_h, \tag{3.1}$$

for a general potential $V : [0,h] \to \mathbb{R}$. The solution reads

$$u(x) = -\left(\frac{1}{\kappa} J \int_0^x \mathrm{e}^V + u_0 \mathrm{e}^{V(0)}\right)\mathrm{e}^{-V(x)}.$$

The flux can be computed explicitly under the assumption $J = \text{const}$ and setting $x = h$ in the above formula. Writing $V_0 = V(0)$ and $V_h = V(h)$ this yields

$$J = -\kappa \frac{u_h \mathrm{e}^{V_h} - u_0 \mathrm{e}^{V_0}}{\int_0^h \mathrm{e}^V} = -\kappa \frac{1}{h}\left(\frac{1}{h}\int_0^h \pi^{-1}\right)^{-1}\left(\frac{u_h}{\pi_h} - \frac{u_0}{\pi_0}\right) = -\kappa \pi_{\mathrm{mean}} \frac{1}{h}\left(\frac{u_h}{\pi_h} - \frac{u_0}{\pi_0}\right)$$

for the averaged $\pi_{\mathrm{mean}} = \left(\frac{1}{h}\int_0^h \pi^{-1}\right)^{-1}$. In particular, for affine $V(x) = \frac{V_h - V_0}{x_h - x_0}(x - x_0) + V_0$, one easily calculates $\pi_{\mathrm{mean}} = (V_h - V_0)/\left(\mathrm{e}^{V_h} - \mathrm{e}^{V_0}\right)$, which yields the Scharfetter–Gummel discretization. However, a potential can also be approximated not by piecewise affine interpolation but in other ways, resulting in different means π_{mean}. We provide an example of such an approximation for the SQRA in the Appendix A.2.

Generalizing the later considerations to higher dimensions, we find for the flux on the edge between two neighboring points in the discretization the expression

$$J^S_{i,j} u^{\mathcal{T}} := -\frac{}{h_{i,j}} S_{i,j}\left(\frac{u_j^{\mathcal{T}}}{\pi_j} - \frac{u_i^{\mathcal{T}}}{\pi_i}\right), \tag{3.2}$$

where relates to κ and $S_{i,j}$ relates to π_{mean}.

We aim to express π_{mean} by means of the values π_0 and π_h at the boundaries. The choice of this average is non-trivial and determines the quality of the discretization scheme, as we will see below. In the present work, we focus on the (weighted) Stolarsky mean, putting $\pi_{\mathrm{mean}} = S(\pi_i, \pi_j)$ although there are also other means like general f-means ($M_f(x,y) = f\left(\left[f^{-1}(x) + f^{-1}(y)\right]/2\right)$ for a strictly increasing function f). The Stolarsky mean has the advantage that it is a closed formula for a broad family of popular means and that its derivatives can be computed explicitly. Moreover, we can – at least in theory – choose different $S_{\alpha,\beta}$ on each interface.

Interestingly, the derivation of the SQRA in Section 2.2 of [33] relies on the assumption that the flux through a FV-interface has to be proportional to $\left(u_j^{\mathcal{T}}/\pi_j - u_i^{\mathcal{T}}/\pi_i\right)$ with the proportionality factor given by a suitable mean of π_i and π_j. The choice of $S_{-1,1}$ in [33] seems arbitrary, yet it yields very good results [14, 19, 51].

3.2. The Wasserstein gradient structure of the Fokker–Planck operator and the SQRA method

The choice of $S_{\alpha,\beta}$ is crucial for the convergence properties but also from a physical point of view. A physically reasonable discretization is not necessarily the best from the rate of convergence point of view and *vice versa*,

compare with numerical simulations in Section 7. However, the physical consideration is helpful to understand the family of discretizations from a different point of view.

In [30] it was proved that the Fokker–Planck equation

$$\dot{u} = \nabla \cdot (\kappa \nabla u + \kappa u \nabla V) \tag{3.3}$$

has the gradient flow formulation $\dot{u} = \partial_\xi \Psi^*(u, -\mathrm{D}E(u))$ where

$$E(u) = \int_\Omega u \log u + Vu - u + 1 = \int_\Omega u \log\left(\frac{u}{\pi}\right) - u + 1, \qquad \Psi^*(u,\xi) = \frac{1}{2}\int_\Omega \kappa u |\nabla \xi|^2, \tag{3.4}$$

and $\pi = \mathrm{e}^{-V}$ is the stationary solution of (3.3). Indeed, one easily checks that $\mathrm{D}E(u) = \log u + V = \log(u/\pi)$ and $\partial_\xi \Psi^*(u,\xi) = -\nabla \cdot (\kappa u \nabla \xi)$ such that it formally holds

$$\partial_\xi \Psi^*(u,\xi)|_{\xi = -\mathrm{D}E(u)} = -\nabla \cdot (\kappa u \nabla \xi)|_{\xi = -\mathrm{D}E(u)} = \nabla \cdot \left(\kappa u \left(\frac{\nabla u}{u} + \nabla V\right)\right) = \nabla \cdot (\kappa \nabla u + \kappa u \nabla V) = \dot{u}.$$

However, the simple parabolic equation $\partial_t u = \Delta u$ can be described either by (3.4) with $V = 0$ or by the choice $E(u) = \int u^2$ with $\Psi^*(\xi) = \int |\nabla \xi|^2$, which plays a role in phase field modeling (see [28] and references therein) or $E(u) = -\int \log u$ with $\Psi^*(\xi) = \int u^2 |\nabla \xi|^2$.

Due to this non uniqueness, one might pose the question about "natural" gradient structures of the discretization schemes that incorporate the underlying physical principles in a discretized way. The discrete energy functional is clearly prescribed by (3.4) with the natural discrete analogue

$$E_{\mathcal{T}}(u) = \sum_i m_i \left(u_i \log\left(\frac{u_i}{\pi_i}\right) - u_i + 1\right). \tag{3.5}$$

Since we identified the continuous flux to be $\mathbf{J} = -\kappa \pi \nabla U$ with $U = u/\pi$, we expect the form

$$\dot{u}_i m_i = \partial_\xi \Psi^*_{\mathcal{T}}(u, -\mathrm{D}E_{\mathcal{T}}(u)) = \sum_{j:i\sim j} \frac{m_{i,j}}{h_{i,j}} \pi_{i,j} \left(\frac{u_j}{\pi_j} - \frac{u_i}{\pi_i}\right) \tag{3.6}$$

for some suitably averaged $\pi_{i,j}$. Equation (3.6) can be understood as a time-reversible (or detailed balanced) Markov process on the finite state space $\mathcal{P}$. Recently, various different gradient structures have been suggested for (3.6): [10, 16, 36, 39, 40] for a quadratic dissipation as a generalization of the Jordan–Kinderlehrer–Otto approach; and [43, 44], where a dissipation of cosh-type was appeared in the large deviation rate functional for a hydrodynamic limit of an interacting particle system. All of them can be written in the abstract form

$$\Psi^*_{\mathcal{T}}(u,\xi) = \frac{1}{2} \sum_i \frac{1}{m_i} \sum_{j:i\sim j} \frac{m_{i,j}}{h_{i,j}} S_{i,j} a_{i,j}(u,\pi) \psi^*(\xi_i - \xi_j), \tag{3.7}$$

$$a_{i,j}(u,\pi) = \left(\frac{u_i}{\pi_i} - \frac{u_j}{\pi_j}\right) \partial_\xi \psi^* \left(\log\left(\frac{u_i}{\pi_i}\right) - \log\left(\frac{u_j}{\pi_j}\right)\right)^{-1}. \tag{3.8}$$

Any positive, normalized and convex function ψ^* defines a mathematically valid dissipation functional Ψ^* by (3.7) and (3.8). A special case is when choosing for ψ^* and exponentially fast growing function $\psi^*(r) := \mathsf{C}^*(r) := 2(\cos h(r/2) - 1)$. Then $a_{i,j}$ simplifies to

$$a_{i,j}(u,\pi) = \sqrt{\frac{u_i u_j}{\pi_i \pi_j}},$$

and hence, providing the square roots. Choosing $S_{i,j} = \sqrt{\pi_i \pi_j}$, we obtain the form

$$\Psi^*_{\mathcal{T}}(u,\xi) = \sum_i \sum_{j:i\sim j} m_{i,j} h_{i,j} \sqrt{u_i u_j}\, \frac{1}{h_{i,j}^2} \mathsf{C}^*(\xi_i - \xi_j). \tag{3.9}$$

The gradient structure of discrete equations has recently attracted a lot of interest. An intense research has developed after [5,6], first highlighting the benefit of gradient structures for the convergence analysis. Besides the classical convergence analysis, further benefits are, *e.g.*, the preservation of the large time behavior [8]. A further recent study on the discrete gradient flows from the more analytical point of view is [46]. There are (at least) three further good reasons why choosing this gradient structure and modeling fluxes in exponential terms: a historical, a mathematical and a physical:

(1) Already in Marcelin's Ph.D. thesis from 1915 [37] exponential reaction kinetics have been derived, which are still common in chemistry literature.
(2) Recently, convergence for families of gradient systems has been derived based on the energy-dissipation principle (the so-called EDP-convergence [15,34,41]). *Vice versa*, the above cosh-gradient structure appears as an effective gradient structure applying EDP-convergence to Wasserstein gradient flow problems [23,34].
(3) The dissipation mechanism Ψ^* of (3.4) is totally independent of the particular form of the energy $\mathcal{E}$, which is determined by the potential V. This is physically understandable, since a change of the potential energy should not influence the dissipation structure. The same holds for the discretized version (3.9). In fact it was shown in [42] (with a similar proof in an earlier version of our paper), that the only discrete gradient structure, where the dissipation does not depend on V, is the cosh-gradient structure with $S_{i,j} = S_{-1,1}(\pi_i, \pi_j)$.

Remark 3.1 (Convergence of energy and dissipation functional)**.** Γ-convergence of $E_{\mathcal{T}} \xrightarrow{\Gamma} E$ can be shown if the fineness of $\mathcal{T}$ tends to 0 since $u \mapsto u\log(u/\pi) - u$ is convex. For the dissipation potentials $\Psi^*_{\mathcal{T}}(u,\xi)$ we observe for smooth functions u and ξ that $\frac{1}{h_{i,j}^2}\mathsf{C}^*(\xi_i - \xi_j) = \frac{1}{2}\Big(\frac{\mathbf{x}_i - \mathbf{x}_j}{|x_i - x_j|}\cdot\nabla\xi\Big)^2 + O(h_{i,j}^2)$ and $\sqrt{u_i u_j} \approx u\big(\frac{1}{2}(x_i + x_j)\big)$. For small mesh size, we get approximately $\Psi^*_{\mathcal{T}}(u,\xi) \approx \frac{1}{2}\int_\Omega \kappa u|\nabla\xi|^2$.

For quadratic dissipation, qualitative convergence results using the underlying gradient structure and the energy-dissipation principle are obtained in [12] in 1D, and in [22] for multiple dimensions. In [25] convergence of the associated metric is proved.

4. Comparison of discretization schemes

We consider two different smooth mean coefficients $S_{i,j} = S(\pi_i, \pi_j)$ and $\tilde{S}_{i,j} = \tilde{S}(\pi_i, \pi_j)$ for two different Stolarsky means S and $\tilde{S}$. In view of (2.4) they both come up with their own consistency $\mathfrak{E}^S_{h,\mathrm{FPE}}(u;\cdot)$ resp. $\mathfrak{E}^{\tilde{S}}_{h,\mathrm{FPE}}(u;\cdot)$ from Lemma 2.6 and a short calculation reveals that

$$\mathfrak{E}^S_{h,\mathrm{FPE}}(u;v) = \mathfrak{E}^{\tilde{S}}_{h,\mathrm{FPE}}(u;v) + \sum_{i\sim j}\frac{m_{i,j}}{h_{i,j}}\Big(S_{i,j}\mathcal{R}_h\frac{u}{\pi} - \tilde{S}_{i,j}\mathcal{R}_h\frac{u}{\pi}\Big)(v_j - v_i). \tag{4.1}$$

Relating to (1.11) we introduce $O_\pi(x^k)$ through

$$\text{there exists } C \text{ depending only on } d,\Omega,\alpha,\beta \text{ and } \|\pi\|_\infty \text{ s.t. } \big|O_\pi(x^k)\big| \le C|x|^k.$$

The derivatives of a general Stolarsky mean $S_{\alpha,\beta}$ satisfy in $x = y \neq 0$ (see Appendix A.1)

$$\begin{aligned}
&\partial_x S_{\alpha,\beta}(x,x) = \partial_y S_{\alpha,\beta}(x,x) = \frac{1}{2},\\
&\partial_x^2 S_{\alpha,\beta}(x,x) = \partial_y^2 S_{\alpha,\beta}(x,x) = -\partial^2_{xy} S_{\alpha,\beta}(x,x) = -\partial^2_{yx} S_{\alpha,\beta}(x,x) = \frac{1}{12x}(\alpha+\beta-3),
\end{aligned} \tag{4.2}$$

so we have the following expansion of $S_{i,j}$: writing $\pi_{i,j} = \frac{1}{2}(\pi_i + \pi_j)$, $\pi_+ = \pi_- = \frac{1}{2}(\pi_i - \pi_j)$ and $\pi_i = \pi_0 + \pi_+$ and $\pi_j = \pi_0 - \pi_-$ we obtain from Taylor's formula

$$S_{i,j} = S_{\alpha,\beta}(\pi_{i,j},\pi_{i,j}) + \frac{1}{2}(\pi_+ - \pi_-) + \frac{1}{2}\partial_x^2 S_{\alpha,\beta}(\pi_{i,j},\pi_{i,j})(\pi_+ + \pi_-)^2 + O_\pi\big(\pi_\pm^3\big)$$

$$= \pi_{i,j} + \frac{\frac{1}{3}(\alpha+\beta)-1}{8\,\pi_{i,j}}(\pi_i-\pi_j)^2 + O_\pi(\pi_i-\pi_j)^3, \tag{4.3}$$

and hence

$$S_{i,j} - \tilde{S}_{i,j} = \frac{(\alpha+\beta)-\left(\tilde{\alpha}+\tilde{\beta}\right)}{24\,\pi_{i,j}}(\pi_i-\pi_j)^2 + O_\pi(\pi_i-\pi_j)^3. \tag{4.4}$$

Proposition 4.1. *Let $\mathcal{T}$ be an admissible mesh. Then*

$$\left|\mathfrak{E}^S_{h,\mathrm{FPE}}(u;v)\right| \le \left|\mathfrak{E}^{\tilde{S}}_{h,\mathrm{FPE}}(u;v)\right| + 2\left(\frac{\left|(\alpha+\beta)-\left(\tilde{\alpha}+\tilde{\beta}\right)\right|}{24\,\pi_{i,j}}(\pi_i-\pi_j)^2 + O_\pi(\pi_i-\pi_j)^3\right)\|u\|_{H^2(\Omega)}\|v\|_{H_\mathcal{T}}.$$

In particular, the last result shows that convergence rates are similar up to order three for different α, β which satisfy $\alpha+\beta = \text{const}$.

Corollary 4.2. *Let $\mathcal{T}_h$ be a quasi uniform family of admissible meshes and let $S_{i,j} = S_{\alpha,\beta}(\pi_i,\pi_j)$ and $\tilde{S}_{i,j} = S_{\tilde{\alpha},\tilde{\beta}}(\pi_i,\pi_j)$ with $\alpha+\beta = \tilde{\alpha}+\tilde{\beta}$ be two different Stolarsky mean coefficients. Then there is a constant C depending only on d, Ω and $\|\pi\|_{C^2(\Omega)}$ such that*

$$\left\|\mathfrak{E}^S_{\mathcal{T},\mathrm{FPE}}(u;v)\right\|^2_{H^*_\mathcal{T}} \le 2\left\|\mathfrak{E}^{\tilde{S}}_{\mathcal{T},\mathrm{FPE}}(u;v)\right\|^2_{H^*_\mathcal{T}} + Ch^6.$$

Proof. This follows from (4.1) and (4.4). □

5. Convergence of the discrete FPE

Throughout this section, we assume that the mesh satisfies the consistency property of Definition 2.8 with a suitable consistency function $\varphi : \mathbb{R}_{\ge 0} \to \mathbb{R}_{\ge 0}$ and discretization operator $\mathcal{R}_h : H^2(\Omega) \to L^2(\mathcal{P}_h)$ as introduced in Section 2. The parameters π_i and u_i below are then given in terms of

$$\pi_i = (\mathcal{R}_h\pi)_i, \qquad u_i = (\mathcal{R}_h u)_i, \qquad U_i = (\mathcal{R}_h U)_i. \tag{5.1}$$

We derive consistency errors for U in Section 5.1 and consistency errors for u in Section 5.2. For both calculations we will need the following result.

Lemma 5.1. *Let $d \le 3$, $\mathcal{T}_h$ be a quasi uniform family of meshes on a polygonal domain $\Omega \subset \mathbb{R}^d$ and let $S_{\alpha,\beta}$ be a Stolarsky mean, $\varpi \in H^2(\Omega)$ and let $R_\varpi := \{(\varpi(x),\varpi(y)) : x,y \in \overline{\Omega}\}$ such that $S_{\alpha,\beta} \in C^2(R_\varpi)$. Then there exists $C > 0$ depending on Ω, d, α, β and ϖsuch that for every h it holds: for every functions $U \in H^2(\Omega)$ with $\varpi_i := \varpi(x_i)$ and $S_{i,j} := S_{\alpha,\beta}(\varpi_i,\varpi_j)$*

$$\left|\int_{\sigma_{i,j}} (\varpi - S_{i,j})\nabla U\cdot\boldsymbol{\nu}_{i,j}\right| \le C\begin{cases}\sum_{k=i,j}\|\nabla\varpi\|_{H^1(\Omega_k)}\|\nabla U\|_{H^1(\Omega_k)} \\ \sum_{k=i,j} h_k^{\frac{1}{2}}\|\nabla\varpi\|_{H^1(\Omega_k)}\left(\int_{\sigma_{i,j}}|\nabla U|^2\right)^{\frac{1}{2}}\end{cases}. \tag{5.2}$$

In particular, we find

$$\sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}(\varpi - S_{i,j})\nabla U\cdot\boldsymbol{\nu}_{i,j}\right)^2 \le Ch^2\|\nabla\varpi\|^2_{H^1}\|\nabla U\|^2_{H^1(\Omega)}. \tag{5.3}$$

Proof. Due to Definition 1.3, *i.e.*, the condition $\mathbb{B}_{rh_i}(x) \subset \Omega_i \subset \mathbb{B}_{Rh_i}(x)$ uniformly over all cells, and the convex polytope shape of the $(\Omega_i)_i$ we find $C > 0$ such that for every cell Ω_i

$$\forall f \in H^1(\Omega_i): \qquad \|f\|^2_{L^2(\sigma_{i,j})} \leq \frac{1}{h_i} C^2 \|f\|^2_{H^1(\Omega_i)}, \tag{5.4}$$

$$\forall f \in H^2(\Omega_i): \qquad \|f - f_i\|^2_{L^2(\sigma_{i,j})} \leq h_i C^2 \|\nabla f\|^2_{H^1(\Omega_i)}. \tag{5.5}$$

Here, equation (5.4) follows *e.g.*, from Section 2.1 of [27] and (5.5) follows from $\|f - f_i\|^2_{C(\Omega_i)} \leq h_i C^2 \|\nabla f\|^2_{H^1(\Omega_i)}$. Observe that

$$\int_{\sigma_{i,j}} |\varpi - S_{i,j}||\nabla U \cdot \boldsymbol{\nu}_{i,j}| \leq \left(\int_{\sigma_{i,j}} |\varpi - S_{i,j}|^2 \right)^{\frac{1}{2}} \left(\int_{\sigma_{i,j}} |\nabla U \cdot \boldsymbol{\nu}_{i,j}|^2 \right)^{\frac{1}{2}}. \tag{5.6}$$

Using (5.4), (5.5) and the C^2-regularity of $S_{\alpha,\beta}$ and R_ϖ we obtain (5.2). Equation (5.3) follows from summing up. □

5.1. Error analysis in U

In view of (2.6) and (2.8) we observe that the natural variational consistency error for a given Stolarsky mean S equivalently takes the form

$$\mathfrak{E}_{h,\mathrm{FPE}}(u;v) = \widetilde{\mathfrak{E}}_{h,\mathrm{FPE}}(U;v) := \sum_{i\sim j} (v_j - v_i) \left(\int_{\sigma_{i,j}} \pi \nabla U \cdot \boldsymbol{\nu}_{i,j} - S_{i,j} \frac{m_{i,j}}{h_{i,j}} \Big((\mathcal{R}_h U)_j - (\mathcal{R}_h U)_i \Big) \right).$$

And as a consequence of Lemma 2.7 we find

$$\left\| \widetilde{\mathfrak{E}}_{h,\mathrm{FPE}}(U;\cdot) \right\|^2_{H^*_{\mathcal{T}_h,S}} \leq \|\pi\|^2_\infty \|\mathfrak{E}_{h,D}(U;\cdot)\|^2_{H^*_{\mathcal{T}_h,S}} + \sum_{i\sim j} \frac{h_{i,j}}{m_{i,j}}^{-1} S_{i,j}^{-1} \left(\int_{\sigma_{i,j}} (\pi - S_{i,j}) \nabla U \cdot \boldsymbol{\nu}_{i,j} \right)^2. \tag{5.7}$$

Using the strict positivity of π on $\overline{\Omega}$ we may apply Lemma 5.1 for $\varpi = \pi$ for every $d \leq 3$ or Proposition 2.11 for strictly positive $g = \pi$ and $d = 1$. Then we immediately infer from (5.7) the main result of the section.

Proposition 5.2 (Localized order of convergence)**.** *Let $d \leq 3$ and the mesh $\mathcal{T}_h$ be a quasi uniform family of admissible meshes and φ-consistent in sense of Definition 2.8. Then for every $U \in H^2(\Omega) \cap H^1_0(\Omega)$ it holds*

$$\|\mathfrak{E}_{h,\mathrm{FPE}}(U;\cdot)\|^2_{H^*_{\mathcal{T},S}} \leq \|\pi\|^2_\infty \|\mathfrak{E}_{h,D}(U;\cdot)\|^2_{H^*_{\mathcal{T}_h,S}} + C(\pi, d, \|\nabla U\|_{H^1}) \times h^2.$$

If $d = 1$ and the mesh $\mathcal{T}_h$ is Voronoi and $U \in C^2(\overline{\Omega})$, then

$$\|\mathfrak{E}_{h,\mathrm{FPE}}(U;\cdot)\|^2_{H^*_{\mathcal{T},S}} \leq C(\pi, d, \|U\|_{C^2})\, h^4.$$

5.2. Error analysis in u

We will now derive an alternative estimate for the consistency error which accounts more for the convective aspect of the FPE and which directly aims at u instead of U. In Lemma 2.6 we have split the consistency error $\mathfrak{E}_{h,\mathrm{FPE}}(u;\cdot)$ into the two parts $\mathfrak{E}_h(u;\cdot)$ and $\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)$.

Proposition 5.3. *Let $d \leq 3$ and $\mathcal{T}_h = (\mathcal{V}_h, \mathcal{E}_h, \mathcal{P}_h)$ be a quasi uniform family of admissible meshes and let the assumptions of Lemma 5.1 hold. Using the notation of Lemma 2.6 let $u_{i,j} := \frac{1}{2}(u_i + u_j)$. Then for some constant C depending on $S_{\alpha,\beta}$ and $\|V\|_{C^2}$*

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}}} \leq 2\|u\|_\infty \|\mathfrak{E}_{h,D}(V;\cdot)\|^2_{H^*_{\mathcal{T}_h}} + 2 \sum_{i\sim j} \frac{h_{i,j}}{m_{i,j}} \left(\int_{\sigma_{i,j}} (u - u_{i,j}) \nabla V \cdot \boldsymbol{\nu}_{i,j} \right)^2 + Ch^2.$$

In case $\alpha+\beta=-1$ the above can be improved to

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}}} \le 2\|u\|_\infty \|\mathfrak{E}_{h,D}(V;\cdot)\|^2_{H^*_{\mathcal{T}_h}} + 2\sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}(u-u_{i,j})\nabla V\cdot\boldsymbol{\nu}_{i,j}\right)^2 + Ch^4.$$

Proof. Following (4.2) we find

$$\frac{S_{i,j}-\pi_j}{\pi_j}u_j - \frac{S_{i,j}-\pi_i}{\pi_i}u_i = \frac12\frac{S_{i,j}}{\pi_i\pi_j}(\pi_i-\pi_j)(u_i+u_j) + \frac12\frac{1}{\pi_i\pi_j}(S_{i,j}\pi_i + S_{i,j}\pi_j - 2\pi_i\pi_j)(u_i-u_j)$$

$$(S_{i,j}\pi_i + S_{i,j}\pi_j - 2\pi_i\pi_j) = \left(\frac12 + C_{\alpha,\beta}\left(\frac{\pi_j}{\pi_i}+\frac{\pi_i}{\pi_j}\right)\right)(\pi_i-\pi_j)^2 + O_\pi(\pi_i-\pi_j)^3$$

for $C_{\alpha,\beta}=\frac{1}{12}(\alpha+\beta-3)$. Hence, we conclude from

$$\begin{aligned}\mathfrak{E}_{h,\mathrm{conv}}(u;v) = &\sum_{i\sim j}\left(\frac{m_{i,j}}{h_{i,j}}\frac12\frac{S_{i,j}}{\pi_i\pi_j}(\pi_i-\pi_j)(u_i+u_j) - \int_{\sigma_{i,j}}u\nabla V\cdot\boldsymbol{\nu}_{i,j}\right)(v_j-v_i)\\ &+\sum_{i\sim j}\left(\frac{m_{i,j}}{h_{i,j}}\frac12\frac{1}{\pi_i\pi_j}\left(\left(\frac12 + C_{\alpha,\beta}\left(\frac{\pi_j}{\pi_i}+\frac{\pi_i}{\pi_j}\right)\right)(\pi_i-\pi_j)^2(u_i-u_j) + O_\pi(\pi_i-\pi_j)^3\right)\right)(v_j-v_i)\end{aligned}$$

that there is a constant C depending on $\|\pi\|_\infty$, $\|\nabla\pi\|_\infty$, $\|u\|_\infty$, $\|\nabla u\|_\infty$ such that we have

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}}} \le \sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}u\nabla V\cdot\boldsymbol{\nu}_{i,j} - \frac{m_{i,j}}{h_{i,j}}\frac12\frac{S_{i,j}}{\pi_i\pi_j}(\pi_i-\pi_j)(u_i+u_j)\right)^2 + C\,h^4.$$

To estimate the right-hand side, we use that for a general Stolarsky mean we have

$$\frac{S_{i,j}}{\pi_i\pi_j}(\pi_i-\pi_j) = \frac12 S_{i,j}\left(\frac{1}{\pi_i}+\frac{1}{\pi_j}\right)(V_j-V_i) + O(h).$$

Defining $g:=u$ and $g_{i,j}:=\frac14 S_{i,j}\left(\frac{1}{\pi_i}+\frac{1}{\pi_j}\right)(u_i+u_j)$ and applying Lemma 2.7 we now obtain

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T}}} \le 2\left(\sup_{i,j}|g_{i,j}|\right)\|\mathfrak{E}_h(V;\cdot)\|^2_{H^*_{\mathcal{T}_h}} + 2\sum_{i\sim j}\frac{h_{i,j}}{m_{i,j}}\left(\int_{\sigma_{i,j}}(u-g_{i,j})\nabla V\cdot\boldsymbol{\nu}_{i,j}\right)^2 + Ch^2.$$

We observe that $\frac12 S_{i,j}\left(\frac{1}{\pi_i}+\frac{1}{\pi_j}\right) = 1 + O(h|\nabla\pi|)$, which implies that

$$\left|\int_{\sigma_{i,j}}(u-g_{i,j})\nabla V\cdot\boldsymbol{\nu}_{i,j}\right| \le \left|\int_{\sigma_{i,j}}(u-u_{i,j})\nabla V\cdot\boldsymbol{\nu}_{i,j}\right| + C|h|.$$

So, the first claim now follows for general $S_{\alpha,\beta}$.

For the case $S_{\alpha,\beta}=S_{0,-1}$, we have $S_{0,-1}(x,y)=\frac{xy}{x-y}\log(x/y)$. Hence, we observe that

$$\frac{S_{i,j}}{\pi_i\pi_j}(\pi_i-\pi_j) = (V_j-V_i).$$

For general $S_{\alpha,\beta}$ with $\alpha+\beta=-1$, we apply Corollary 4.2, which proves the second estimate. □

We conclude the section by the following proof.

Proof of Theorem 1.5. Using the Definition 2.8, the result is an immediate consequence of Lemma 2.6, Proposition 5.3 and Lemma 5.1. □

6. Cubic Meshes

Throughout this section we consider $d \leq 3$ and a polygonal domain $\Omega \subset \mathbb{R}^d$ with a cubic mesh where $\Omega_i = x_i + [-h/2, h/2]^d$, $x_i \in h\mathbb{Z} \cap \Omega$.

Lemma 6.1. *Let $\Omega \subset \mathbb{R}^d$ be a polygonal domain with $d \leq 3$ and a cubic mesh where $\Omega_i = x_i + [-h/2, h/2]^d$, $x_i \in h\mathbb{Z} \cap \Omega$. Then for every functions $V \in C^1(\overline{\Omega})$ and $\varpi \in C^2$ with $\varpi_i := \varpi(x_i)$ and $S_{i,j} := S_{\alpha,\beta}(\varpi_i, \varpi_j)$ there exist $C_\varpi, C_V > 0$ depending only on $\|\varpi\|_{C^2}$ and $\|\nabla V\|_{C^1}$ respectively, such that for every $U, u \in H^2(\Omega) \cap H_0^1(\Omega)$ it holds*

$$\left| \int_{\sigma_{i,j}} (\varpi - S_{i,j}) \nabla U \cdot \boldsymbol{\nu}_{i,j} \right|^2 \leq C_\varpi h^{d+2} \|\nabla U\|^2_{H^1(\Omega_i)}, \tag{6.1}$$

$$\left| \int_{\sigma_{i,j}} (u - u_{i,j}) \nabla V \cdot \boldsymbol{\nu}_{i,j} \right|^2 \leq C_V h^{d+2} \|\nabla u\|^2_{H^1(\Omega_i)}, \tag{6.2}$$

where $u_{i,j} = \frac{1}{2}(u_i + u_j)$.

Proof. Let $Q = [0,1]^d$ with midpoint $\overline{x}$. There exists $C > 0$ such that for every $f \in H^1(Q)$ with $\int_Q f = 0$ and every $g \in H^2(Q)$ it holds

$$\int_{\partial Q} f^2 \leq C \int_Q |\nabla f|^2, \qquad \int_{\partial Q} (g - g(\overline{x}))^2 \leq C \int_Q \left(|\nabla g|^2 + \left|\nabla^2 g\right|^2 \right).$$

Hence for each $\Omega_i \in \mathcal{V}_h$ and $G_{U,i} := \fint_{\Omega_i} \nabla U$ we find by a scaling argument

$$\int_{\partial \Omega_i} |\nabla U - G_{U,i}|^2 \leq hC \int_{\Omega_i} |\nabla(\nabla U)|^2, \tag{6.3}$$

$$\int_{\partial \Omega_i} |u - u_i|^2 \leq hC \int_{\Omega_i} \left(|\nabla u|^2 + \left|\nabla^2 u\right|^2 \right). \tag{6.4}$$

Proof of (6.1)*:* We first observe

$$\left| \int_{\sigma_{i,j}} (\varpi - S_{i,j}) \nabla U \cdot \boldsymbol{\nu}_{i,j} \right| \leq \left| \int_{\sigma_{i,j}} (\varpi - S_{i,j}) G_{U,i} \right| + \left| \int_{\sigma_{i,j}} |\varpi - S_{i,j}| |\nabla U - G_{U,i}| \right|.$$

We then find for some C depending on d that $|\varpi - S_{i,j}| \leq C \|\nabla \varpi\|_\infty h$ and hence using also (6.3)

$$\begin{aligned} \left| \int_{\sigma_{i,j}} |\varpi - S_{i,j}| |\nabla U - G_{U,i}| \right|^2 &\leq \left(\int_{\sigma_{i,j}} |\varpi - S_{i,j}|^2 \right) \left(\int_{\sigma_{i,j}} |\nabla U - G_{U,i}|^2 \right) \\ &\leq C h^{d+2} \int_{\Omega_i} |\nabla(\nabla U)|^2. \end{aligned}$$

We have for $x \in \sigma_{i,j}$ and $\xi_x \in \left\{ \nabla^2 \varpi(y) : y \in \overline{\Omega} \right\}$, $\xi_{x,i,j}, \xi_x \in \left\{ \nabla^2 S(\varpi(y), \varpi(y))) : y \in \overline{\Omega} \right\}$ with

$$\begin{aligned} |\xi_{x,i,j}|_\infty \leq C_{\alpha,\beta,\varpi} &:= \sup_{y \in \Omega} \left\| \nabla^2 S(\varpi(y), \varpi(y))) \right\|_\infty, \\ |\xi_x| &\leq \|\varpi\|_{C^2} \end{aligned}$$

that for $G_{\varpi,i} := \fint_{\Omega_i} \nabla \varpi$ it holds

$$S_{i,j} - \varpi(x) = \frac{1}{2}(\varpi_i - \varpi(x)) + \frac{1}{2}(\varpi_j - \varpi(x)) + \begin{pmatrix} \varpi_i - \varpi(x) \\ \varpi_j - \varpi(x) \end{pmatrix} \cdot \xi_{x,i,j} \begin{pmatrix} \varpi_i - \varpi(x) \\ \varpi_j - \varpi(x) \end{pmatrix},$$

$$\varpi_i - \varpi(x) = G_{\varpi,i}\cdot(x_i - x) + (\nabla\varpi(x) - G_{\varpi,i})(x_i - x) + (x_i - x)\xi_x(x_i - x).$$

Thus we find

$$\left|\int_{\sigma_{i,j}} (\varpi - S_{i,j})G_{U,i}\right| \le \sum_{k=i,j}\left|\int_{\sigma_{i,j}} (\varpi - \varpi_k)G_{U,i}\right| + C\|\nabla\varpi\|_\infty^2 h^{d+1}|G_{U,i}|,$$

where $C = C_{\alpha,\beta,\varpi}$ in general and $C_{\alpha,\beta,\varpi} = 0$ if $S_{\alpha,\beta}(a,b) = \frac{1}{2}(a+b)$. Due to the anti symmetry of $G_{\varpi,i}\cdot(x_i - x)$ on $\sigma_{i,j}$, $|\nabla\varpi(x) - G_{\varpi,i}| \le \|\varpi\|_{C^2} h$ and $|\xi_x| \le \|\varpi\|_{C^2}$ we obtain

$$\left|\int_{\sigma_{i,j}} (\varpi - S_{i,j})G_{U,i}\right| \le h^{d+1}\|\varpi\|_{C^2}|G_{U,i}|$$

Now with $|G_{U,i}| \le h^{-\frac{d}{2}}\|\nabla U\|_{L^2(\Omega_i)}$ it follows in total

$$\left|\int_{\sigma_{i,j}} (\varpi - S_{i,j})\nabla U\cdot\boldsymbol{\nu}_{i,j}\right|^2 \le Ch^{d+2}\|\nabla U\|^2_{H^1(\Omega_i)}.$$

Proof of (6.2)*:* We start from

$$\left|\int_{\sigma_{i,j}} (u - u_{i,j})\nabla V\cdot\boldsymbol{\nu}_{i,j}\right| \le \left|\int_{\sigma_{i,j}} (u - u_{i,j})G_{V,i}\right| + \left|\int_{\sigma_{i,j}} |u - u_{i,j}||\nabla V - G_{V,i}|\right|.$$

We find for some C depending on d that $|\nabla V - G_{V,i}| \le C\|\nabla(\nabla V)\|_\infty h$ and using (6.4)

$$\left|\int_{\sigma_{i,j}} |u - u_{i,j}||\nabla V - G_{V,i}|\right|^2 \le Ch^{d+2}\int_{\Omega_i}|\nabla u|^2.$$

For the second term, we make use of $G_{u,i} := \fint_{\Omega_i}\nabla u$ and

$$u_i - u(x) = G_{u,i}\cdot(x_i - x) + (\nabla u(x) - G_{u,i})(x_i - x) + \int_0^1 (x_i - x)\nabla^2 u(tx + (1-t)x_i)(x_i - x)\,\mathrm{d}t.$$

By anti-symmetry of $G_{u,i}\cdot(x_i - x)$ on $\sigma_{i,j}$ we obtain using (6.3)

$$\begin{aligned}\left|\int_{\sigma_{i,j}} (u - u_i)G_{V,i}\right| &\le h^{1+\frac{d-1}{2}}|G_{V,i}|\left(\int_{\sigma_{i,j}}|\nabla u(x) - G_{u,i}|^2\right)^{\frac{1}{2}} + h|G_{V,i}|\int_{\Omega_i}\left|\nabla^2 u\right| \\ &\le h^{1+\frac{d}{2}}|G_{V,i}|\left(\int_{\Omega_i}\left|\nabla^2 u\right|^2\right)^{\frac{1}{2}}.\end{aligned}$$

Together, this implies (6.2). □

Proposition 6.2 (Consistency on cubic meshes)**.** *Let $\Omega\subset\mathbb{R}^d$ with $d\le 3$ be a polygonal domain with a family of cubic meshes where for each h we set $\Omega_i = x_i + [-h/2, h/2]^d$, $x_i\in h\mathbb{Z}\subset\Omega$. Then for some constant C depending only on d, Ω and $\|\pi\|_{C^2(\Omega)}$*

$$\|\mathfrak{E}_{h,S}(u;\cdot)\|^2_{H^*_{\mathcal{T},S}} \le Ch^4\|u\|^2_{H^2(\Omega)}.$$

Proof. The statement follows from the application of Lemma 2.7 twice for $g = S$ together with Lemma 6.1 and Proposition 2.10. □

In view of Theorem 5.3 combined with Theorem 6.2 and Lemma 6.1 with $\frac{1}{2}(u_i + u_j) = S_{2,1}(u_i, u_j)$ we also obtain the following.

Proposition 6.3. *Let $d \leq 3$. On a polygonal domain $\Omega \subset \mathbb{R}^d$ with a cubic mesh where $\Omega_i = x_i + [-h/2, h/2]^d$, $x_i \in h\mathbb{Z} \subset \Omega$, it holds: Using the notation of Lemma 2.6 it holds for some constant C depending only on d, Ω and $\|\pi\|_{C^2(\Omega)}$*

$$\|\mathfrak{E}_{h,\mathrm{conv}}(u;\cdot)\|^2_{H^*_{\mathcal{T},S}} \leq Ch^k \|u\|^2_{H^2(\Omega)},$$

where $k = 2$ in general and $k = 4$ in case $\alpha + \beta = -1$.

Proof of Theorem 1.4. The claim follows from Proposition 5.2 (general case) or Proposition 2.10 together with twice application of Lemma 6.1 in the cubic case. □

Proof of of Theorem 1.7. This is a consequence of Lemma 2.6, Propositions 2.10 (resp. Prop. 6.2) and 6.3. □

7. Numerical tests and convergence analysis

In this section, we provide a numerical convergence analysis of the discretization schemes based on Stolarsky means described above. As the central problem of flux discretization is in the context of the finite volume method essentially one-dimensional (*cf.* Sect. 3), we restrict ourselves to the analysis of one-dimensional test problems on iteratively refined grids, for which already non-trivial results can be observed. We consider non-equidistant grids in order to rule out possible cancellation effects and spurious convergence properties which might occur on uniform grids.

In the examples below, the non-equidistant grids are generated with the help of a *mesh density function* $\rho : [0,1] \to [0,1]$. We choose

$$\rho(x) = \frac{1}{1 + \left(\frac{1-x}{x}\right)^a}, \tag{7.1}$$

where $a > 0$ is a shape parameter. The mesh density function equation (7.1) transforms an equidistant mesh $\{x_i\}_{i=1\ldots N}$ with $x_i = (i-1)h$ and $h = 1/(N-1)$ into a non-equidistant one $\{\rho(x_i)\}_{i=1\ldots N}$, where N is the number of nodes. For $a > 1$, the mesh density function is S-shaped, which implies small grid spacings close to the boundaries and larger grid spacings in the center of the computational domain. Note that the grids satisfy the quasi uniformity condition given in Definition 1.3. The example calculations described below are carried out for $a = 1$ (equidistant grid, $\rho(x)|_{a=1} = x$) and $a = 4$ (non-equidistant grid).

Example 7.1. We consider the potential $V(x) = 30x(1+x)$ and the right hand side $f(x) = x(1-x)$ on the domain $(0,1)$ with diffusion coefficient $\kappa = 1$ and homogeneous Dirichlet boundary conditions $u(0) = u(1) = 0$. The numerical solutions obtained using the Stolarsky mean discretizations are compared point-wise with the exact solution u_{ref} (involving the imaginary error function) that has been obtained analytically with the help of Mathematica [52].

The numerical results for Example 7.1 are summarized in Figure 2. In Figure 2a, the logarithmic error $\log_{10}(\|u - u_{\mathrm{ref}}\|_{H_\mathcal{T}})$ is shown in the (α,β)-plane of the Stolarsky-mean parameters for an equidistant grid with $2^{10} + 1 = 1025$ nodes. First, we note that the accuracy for a mean $S_{\alpha,\beta}$ is indeed practically invariant along $\alpha + \beta = \text{const.}$, which supports Corollary 4.2 in Section 4 and our main theorems (see Sect. 1.2), respectively. In the present example, we observe optimal accuracy around $\alpha + \beta = -1$, which includes the SG-scheme (Stolarsky mean $S_{0,-1}$) as a special case. Figure 2b shows the convergence behavior under iterative mesh refinement, where the fastest convergence in the $H_\mathcal{T}$-norm is indeed observed for the SG-scheme. Note that, however, also the other schemes considered in the comparison show a quadratic convergence behavior (as predicted in for one-dimensional problems by Thm. 5.2), but with a larger constant. The results for the non-equidistant grid (shape parameter $a = 4$) shown in Figures 2c and 2d are qualitatively the same as in the equidistant case. We observe that the optimum around $\alpha + \beta = -1$ becomes sharper in the case of non-equidistant grids (compare Figs. 2a and 2c), which we interpret as a result of the improved grid resolution at the domain boundaries.

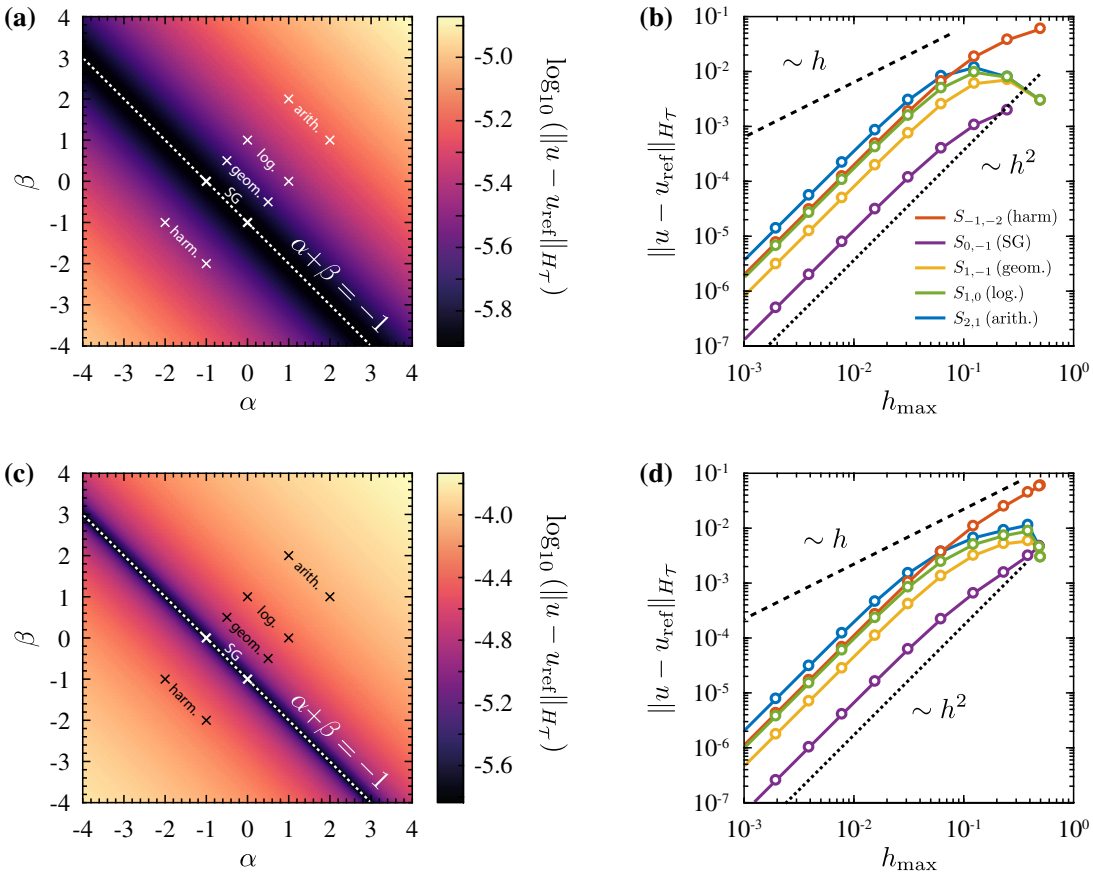

FIGURE 2. Numerical results for Example 7.1. (a) Discretization error $\log_{10}(\|u-u_{\mathrm{ref}}\|_{H_{\mathcal{T}}})$ in the (α,β)-plane on an equidistant grid ($a=1$) with $2^{10}+1$ nodes. The error is color-coded and is minimal around $\alpha+\beta=-1$. Several special Stolarsky means (*cf.* Tab. 1) are highlighted by crosses. Note the symmetry $S_{\alpha,\beta}(x,y)=S_{\beta,\alpha}(x,y)$. (b) Quadratic convergence of the discrete solution to the exact reference solution u_{ref} under mesh refinement in the $H_{\mathcal{T}}$-norm. See the inset for a legend and color-coding of the considered means $S_{\alpha,\beta}$. In the present example, the SG scheme $S_{0,-1}$ provides the fastest convergence under mesh refinement. (c), (d) Same as in panels (a), (b) but on a non-equidistant mesh with shape parameter $a=4$.

Example 7.2. We consider the potential $V(x)=2\exp(2x)$ and keep the right hand side, diffusion constant and boundary conditions as in Example 7.1. The reference solution was computed numerically to a high precision using a shooting method (involving a fourth order Runge–Kutta method together with Brent's root finding algorithm [3]) on a fine grid with 7937 nodes.

The results of the numerical convergence analysis for Example 7.2 are presented in Figure 3. The plot of the discretization errors $\|u-u_{\mathrm{ref}}\|_{H_{\mathcal{T}}}$ in the (α,β)-plane of the Stolarsky-mean parameters shows a minimum around $\alpha+\beta=0$, see Figures 3a and 3c. This optimum includes the SQRA scheme with geometric mean $S_{1,-1}$ and is qualitatively the same in the case of equidistant and non-equidistant grids. Just as in the previous example, we observe quadratic convergence for all considered Stolarsky-mean schemes under iterative mesh refinement, see Figures 3b and 3d. Note that the potential gradient $V'(x)$, which acts as a driving force for the drift-like flux component, in Example 7.2 ($4\le V'(x)<30$) is smaller than in Example 7.1 ($30\le V'(x)\le 90$). Hence, our

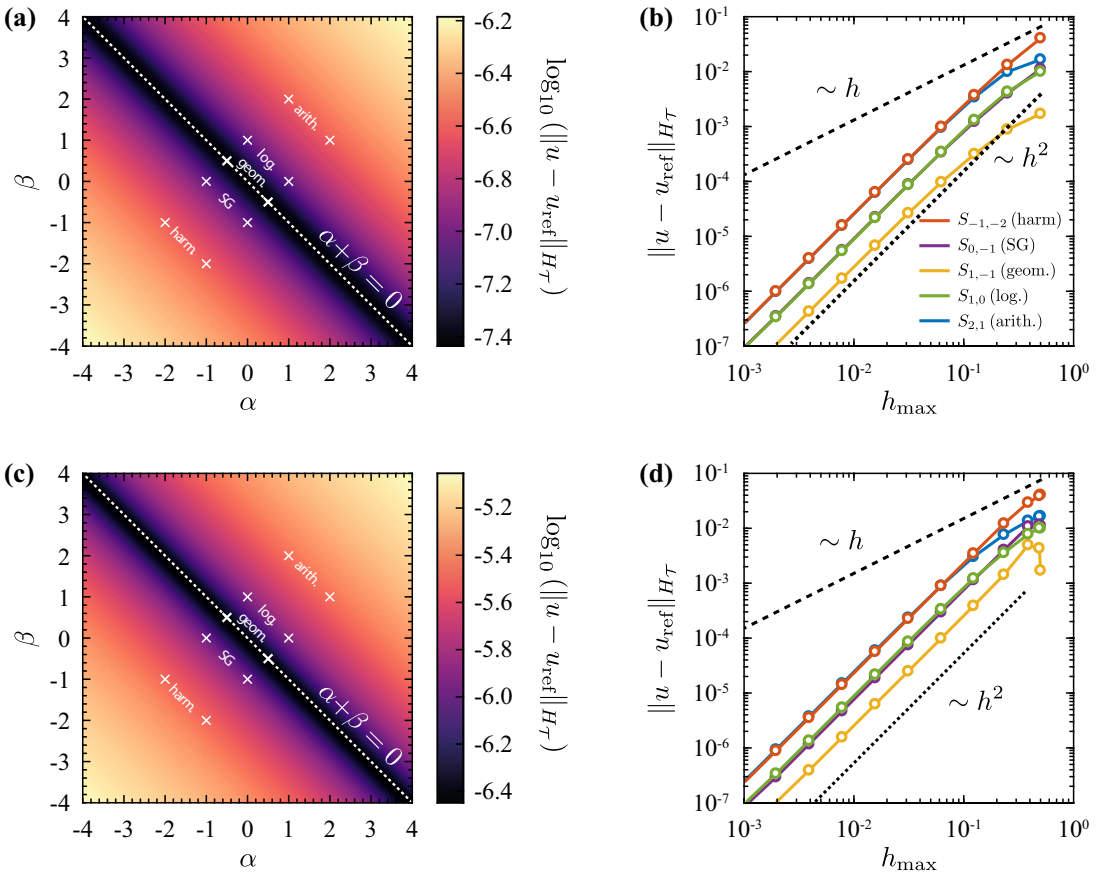

FIGURE 3. Discretization errors and convergence behavior of the numerically computed solution u in the $H_\mathcal{T}$-norm for Example 7.2 using the Stolarsky-mean schemes. The errors in (a) and (c) are color-coded (as before, on grids with $2^{10}+1=1025$ nodes). The coloring of the means in (b) and (d) is the same as in Figure 2b. For the considered example, the results indicate a superior performance of the SQRA scheme (geometric mean $S_{1,-1}$) on the equidistant as well as on the non-equidistant grid (shape parameter $a=4$).

results obtained for Example 7.2 indicate that away from the drift-dominated regime, the SG-scheme might be outperformed by other Stolarsky-mean schemes (*e.g.*, the SQRA scheme). This legitimizes the use of alternative flux discretizations for problems with moderate potential gradients, as carried out in reference [14].

Finally, Figure 4 shows the discretization error $\log_{10}(\|U-U_{\mathrm{ref}}\|_{H_\mathcal{T}})$ obtained using the Stolarsky-mean schemes for Example 7.2. We observe that the optimal parameters are $\alpha+\beta=0.6$ in the equidistant case, see Figure 4a, and $\alpha+\beta=-0.2$ in the non-equidistant case, see Figure 4b, which is clearly different from the optimal parameter set required to obtain maximum accuracy of u, *cf.* Figures 3a and 3c.

8. OUTLOOK

The results of this work suggest to search for "optimal" parameters α and β in the choice of the Stolarsky mean in order to reduce the error of the approximation as much as possible. However, from an analytical point of view, the quest for such optimal α and β is quite challenging. Moreover, since the optimal choice might vary

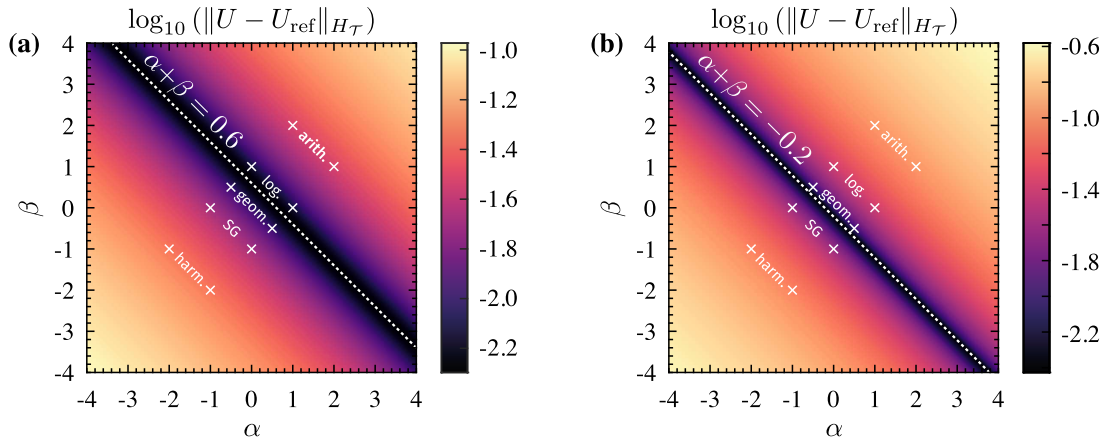

FIGURE 4. Comparison of the discretization errors $\log_{10}(\|U - U_{\text{ref}}\|_{H_\mathcal{T}})$ for Example 7.2 on a grid with 1025 nodes with (a) equidistant and (b) non-equidistant spacing (shape parameter $a = 4$). The optimum Stolarsky-mean parameters (α, β) for minimum error in U are different from those required for minimum error in u, *cf.* Figure 3.

locally, depending on the local properties of the potential V, we suggest to implement a learning algorithm that provides suitable parameters α and β depending on the local structure of V and the mesh.

APPENDIX A.

A.1. Properties of the Stolarsky mean

Lemma A.1. *For every of the above Stolarsky means $S_*(x, y)$ it holds*

$$\partial_x S_*(x,x) = \partial_y S_*(x,x) = \frac{1}{2} \quad \text{and} \quad \partial_x^2 S_*(x,x) = \partial_y^2 S_*(x,x) = -\partial_{xy}^2 S_*(x,x) = -\partial_{yx}^2 S_*(x,x).$$

Proof. Since $S_*(x,x) = x$ and S_* is symmetric in x and y, we find from differentiating $\partial_x S_* = \partial_y S_* = \frac{1}{2}$. From the last equality, we find $\partial_x S_*(x,x) - \partial_y S_*(x,x) = 0$ as well as $\partial_x S_*(x,x) + \partial_y S_*(x,x) = 1$ and differentiation yields

$$\partial_x^2 S_*(x,x) - \partial_y^2 S_*(x,x) - \partial_{xy}^2 S_*(x,x) + \partial_{yx}^2 S_*(x,x) = 0, \tag{A.1}$$

$$\partial_x^2 S_*(x,x) + \partial_y^2 S_*(x,x) + \partial_{xy}^2 S_*(x,x) + \partial_{yx}^2 S_*(x,x) = 0. \tag{A.2}$$

Since $-\partial_{xy}^2 S_*(x,x) + \partial_{yx}^2 S_*(x,x) = 0$, equation (A.1) yields $\partial_x^2 S_*(x,x) = \partial_y^2 S_*(x,x)$. Inserting the last two relations into (A.2) yields $\partial_{xy}^2 S_*(x,x) = \partial_{yx}^2 S_*(x,x) = -\partial_x^2 S_*(x,x)$. □

Lemma A.2. *It holds $\partial_x^2 S_{\alpha,\beta}(\pi,\pi) = \frac{1}{12\pi}(\alpha + \beta - 3)$.*

Proof. We know from Lemma A.1 that $\partial_x S_{\alpha,\beta}(x,x) = \frac{1}{2}$ and $\partial_x^2 S_{\alpha,\beta}(x,x) = -\partial_y \partial_x S_{\alpha,\beta}(x,x)$. Given a fixed y, we define $z = x/y$ and find

$$f(z) := yz S_{\alpha,\beta}(1,z) = S_{\alpha,\beta}(yz, y)$$

satisfies

$$\partial_z f(z) = y \partial_x S_{\alpha,\beta}(yz,y), \qquad \partial_{zz} f(z) = y^2 \partial_{xx} S_{\alpha,\beta}(yz,y),$$

and hence $\partial_{zz} f(1) = y^2 \partial_{xx} S_{\alpha,\beta}(y,y)$. In $z \neq 1$ it holds

$$\partial_z f(z) = y\left(\frac{\beta}{\alpha}\right)^{\frac{1}{\alpha-\beta}} \frac{(z^\alpha-1)^{\frac{1}{\alpha-\beta}-1}}{(z^\beta-1)^{\frac{1}{\alpha-\beta}-1}} \frac{\alpha\big(z^\beta-1\big)z^\alpha - \beta(z^\alpha-1)z^\beta}{(\alpha-\beta)\ z\ (z^\beta-1)^2}$$

and in $z=1$ we find $\partial_z f(1) = y\frac{1}{2}$. Using an expansion $z = 1+h$ in

$$\lim_{h\to 0} h^{-1}\left(y^{-1}\partial_z f(1+h) - \frac{1}{2}\right) = y\partial_{xx} S(y,y)$$

we conclude with (4.2). □

A.2. Approximation of potential to get the SQRA mean

The aim of this section is to provide a class of potentials which are easy to handle and which generate the SQRA-mean $S_{-1,1}(\pi_0,\pi_h)$ by $\pi_{\text{mean}} = \left(\frac{1}{h}\int_0^h \pi^{-1}\right)^{-1}$. Clearly, choosing the constant potential $V(x) := V_c := -\log S_{-1,1}(\pi_0,\pi_h)$ we obtain right mean. Although this works for any means, this has two drawbacks

(1) The potential jumps and hence the gradient is somewhere infinite, which means that at these points the force on the particles is infinitely high which is not physical.
(2) Approximating a general function by piecewise constants, on each interval the accuracy is only of order h. However, approximating a function by affine interpolation the accuracy is of order h^2 on each interval (see below for the calculation).

So we want to get a potential which may be used as a good approximation (*i.e.*, approximating of order h^2), is physical (*i.e.*, continuous) and generates the SQRA-mean. Note, that most considerations below also work for other Stolarsky means. For simplicity we focus on the SQRA mean $S_{-1,1}$.

We consider a piecewise affine potential of the form

$$\hat{V}(x) = \begin{cases} \frac{V_c-V_0}{x_1}x + V_0, & x\in[0,x_1] \\ V_c, & x\in[x_1,x_2] \\ \frac{V_h-V_c}{h-x_2}(x-x_2)+V_c, & x\in[x_2,h] \end{cases}$$

where $x_1, x_2 \in [0,h]$ are firstly arbitrary and $V_c = -\log S_{-1,1}(\pi_0,\pi_h) = \frac{1}{2}(V_h+V_0)$. The potential is clearly continuous. Then

$$\frac{1}{h}\int_0^h \mathrm{e}^{\hat{V}(x)}\mathrm{d}x = \frac{x_1}{h}\frac{\mathrm{e}^{V_c}-\mathrm{e}^{V_0}}{V_c-V_0} + \frac{x_2-x_1}{h}\mathrm{e}^{V_c} + \frac{h-x_2}{h}\frac{\mathrm{e}^{V_h}-\mathrm{e}^{V_c}}{V_h-V_c}.$$

Introducing the ratios $\alpha = \frac{x_1}{h}$ and $\beta = \frac{h-x_2}{h}$ (which are in $[0,1/2]$), we want to solve $\frac{1}{h}\int_0^h \mathrm{e}^{\hat{V}(x)}\mathrm{d}x = \mathrm{e}^{\frac{1}{2}(V_h+V_0)}$. Indeed, introducing the difference of the difference of the potentials $\bar{V} = V_h - V_0$, we obtain

$$\lambda = \frac{\alpha}{\beta} = \frac{\mathrm{e}^{\bar{V}/2} - \bar{V}/2 - 1}{\mathrm{e}^{-\bar{V}/2} + \bar{V}/2 - 1} \approx 1 + \frac{1}{3}\bar{V} + \frac{1}{18}\bar{V}^2.$$

Hence, any value α,β satisfying this ratio generates a potential with the SQRA-mean.

Acknowledgements. M.H. is financed by Deutsche Forschungsgemeinschaft (DFG) within SPP 2256 Project 441154659, HE 8716/1-1. A. S. is financed by DFG through Grant CRC 1114 "Scaling Cascades in Complex Systems", Project C05 Effective models for materials and interfaces with multiple scales. The work of M. K. received funding from the DFG under Germany's Excellence Strategy EXC2046: MATH+.

References

[1] D.N.D.G. Allen and R.V. Southwell, Relaxation methods applied to determine the motion in two dimensions of a viscous fluid past a fixed cylinder. *Q. J. Mech. Appl. Math.* **8** (1955) 129–145.

[2] R. Bank, W. Coughran and L.C. Cowsar, The finite volume scharfetter-gummel method for steady convection diffusion equations. *Comput. Visulaization Sci.* **1** (1998) 123–136.

[3] R.P. Brent, An algorithm with guaranteed convergence for finding a zero of a function. *Comput. J.* **14** (1971) 422–425.

[4] F. Brezzi, L.D. Marini and P. Pietra, Numerical simulation of semiconductor devices. *Comput. Methods Appl. Mech. Eng.* **75** (1989) 493–514.

[5] C. Cancès and C. Guichard, Convergence of a nonlinear entropy diminishing control volume finite element scheme for solving anisotropic degenerate parabolic equations. *Math. Comput.* **85** (2016) 549–580.

[6] C. Cancès and C. Guichard, Numerical analysis of a robust free energy diminishing finite volume scheme for parabolic equations with gradient structure. *Found. Comput. Math.* **17** (2017) 1525–1584.

[7] C. Chainais-Hillairet and J. Droniou, Finite-volume schemes for noncoercive elliptic problems with neumann boundary conditions. *IMA J. Numer. Anal.* **31** (2011) 61–85.

[8] C. Chainais-Hillairet and M. Herda, Large-time behaviour of a family of finite volume schemes for boundary-driven convection–diffusion equations. *IMA J. Numer. Anal.* **40** (2020) 2473–2504.

[9] J.S. Chang and G. Cooper, A practical difference scheme for fokker-planck equations. *J. Comput. Phys.* **6** (1970) 1–16.

[10] S.-N. Chow, W. Huang, Y. Li and H. Zhou, Fokker–Planck equations for a free energy functional or Markov process on a graph. **203** (2012) 969–1008.

[11] D.A. Di Pietro and J. Droniou, A third strang lemma and an Aubin–Nitsche trick for schemes in fully discrete formulation. *Calcolo* **55** (2018) 40.

[12] K. Disser and M. Liero, On gradient structures for Markov chains and the passage to Wasserstein gradient flows. *Networks Heterg. Media* **10** (2015) 233–253.

[13] P.D. Dixit, A. Jain, G. Stock and K.A. Dill, Inferring transition rates of networks from populations in continuous-time markov processes. *J. Chem. Theory Comput.* **11** (2015) 5464–5472.

[14] L. Donati, M. Weber and B.G. Keller, Markov models from the square root approximation of the Fokker–Planck equation: calculating the grid-dependent flux. *J. Phys. Condensed Matter* **33** (2021) 115902.

[15] P. Dondl, T. Frenzel and A. Mielke, A gradient system with a wiggly energy and relaxed EDP-convergence. *ESAIM Control Optim. Calc. Var.* **25** (2019) 68.

[16] M. Erbar and J. Maas, Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ratio. Mech. Anal.* **206** (2012) 997–1038.

[17] R. Eymard, T. Gallouët and R. Herbin, Finite volume methods. *Handb. Numer. Anal.* **7** (2000) 713–1018.

[18] R. Eymard, J. Fuhrmann and K. Gärtner, A finite volume scheme for nonlinear parabolic equations derived from one-dimensional local dirichlet problems. *Numer. Math.* **102** (2006) 463–495.

[19] K. Fackeldey, P. Koltai, P. Névir, H. Rust, A. Schild and M. Weber, From metastable to coherent setstime-discretization schemes. *Chaos Interdisciplinary J. Nonlinear Sci.* **29** (2019) 012101.

[20] P. Farrell, T. Koprucki and J. Fuhrmann, Computational and analytical comparison of flux discretizations for the semiconductor device equations beyond Boltzmann statistics. *J. Comput. Phys.* **346** (2017) 497–513.

[21] P. Farrell, N. Rotundo, D.H. Doan, M. Kantner, J. Fuhrmann and T. Koprucki, Drift-Diffusion Models. In: Handbook of Optoelectronic Device Modeling and Simulation: Lasers, Modulators, Photodetectors, Solar Cells, and Numerical Methods, edited by J. Piprek. Vol. 2, Chapter 50, CRC Press, Taylor & Francis Group, Boca Raton (2017) 731–771.

[22] D. Forkert, J. Maas and L. Portinale, Evolutionary γ-convergence of entropic gradient flow structures for fokker-planck equations in multiple dimensions. Preprint arXiv:2008.10962 (2020).

[23] T. Frenzel and M. Liero, Effective diffusion in thin structures via generalized gradient systems and EDP-convergence. *Discr. Cont. Dyn. Syst.-S.* **14** (2021) 395.

[24] T. Gallouët, R. Herbin and M.H. Vignal, Error estimates on the approximate finite volume solution of convection diffusion equations with general boundary conditions. *SIAM J. Numer. Anal.* **37** (2000) 1935–1972.

[25] P. Gladbach, E. Kopfer, J. Maas and L. Portinale, Homogenisation of one-dimensional discrete optimal transport. *J. Math. Pures App.* **139** (2020) 204–234.

[26] M. Heida, Convergences of the squareroot approximation scheme to the Fokker–Planck operator. *Math. Models Methods Appl. Sci.* **28** (2018) 2599–2635.

[27] M. Heida, Stochastic homogenization on randomly perforated domains. Preprint arXiv:2001.10373 (2020).

[28] M. Heida, J. Màlek and K.R. Rajagopal, On the development and generalizations of Allen–Cahn and Stefan equations within a thermodynmic framework. *Z. Angew. Math. Phys. (ZAMP)* **63** (2012) 759–776.

[29] A.M. Il'in, Differencing scheme for a differential equation with a small parameter affecting the highest derivative. *Math. Notes Acad. Sci. USSR* **6** (1969) 237–248. Translated from *Mat. Zametki* **6** (1969) 237–248.

[30] R. Jordan, D. Kinderlehrer and F. Otto, The variational formulation of the Fokker–Planck equation. *SIAM J. Math. Anal.* **29** (1998) 1–17.

[31] M. Kantner, Generalized Scharfetter–Gummel schemes for electro-thermal transport in degenerate semiconductors using the Kelvin formula for the Seebeck coefficient. *J. Comput. Phys.* **402** (2020) 109091.

[32] R.D. Lazarov, I.D. Mishev and P.S. Vassilevski, Finite volume methods for convection-diffusion problems. *SIAM J. Numer. Anal.* **33** (1996) 31–55.
[33] H.C. Lie, K. Fackeldey and M. Weber, A square root approximation of transition rates for a markov state model. *SIAM J. Matrix Anal. App.* **34** (2013) 738–756.
[34] M. Liero, A. Mielke, M.A. Peletier and D.R. Michiel Renger, On microscopic origins of generalized gradient structures. *Discr. Cont. Dynam. Syst. Ser. S* **10** (2017) 1–35.
[35] L. Lu and J.-G. Liu, Large time behaviors of upwind schemes and b-schemes for Fokker–Planck equations on $\mathbb{R}$ by jump processes. *Math. Comput.* **89** (2020) 2283–2320.
[36] J. Maas, Gradient flows of the entropy for finite Markov chains. *J. Funct. Anal.* **261** (2011) 2250–2292.
[37] R. Marcelin, Contribution a l'étude de la cinétique physico-chimique. *Ann. Phys.* **III** (1915) 120–231.
[38] P.A. Markowich, The Stationary Semiconductor Device Equations. Springer, Vienna (1986).
[39] A. Mielke, A gradient structure for reaction-diffusion systems and for energy-drift-diffusion systems. *Nonlinearity* **24** (2011) 1329–1346.
[40] A. Mielke, Geodesic convexity of the relative entropy in reversible markov chains. *Calc. Var. Part. Differ. Equ.* **48** (2013) 1–31.
[41] A. Mielke, On evolutionary Γ-convergence for gradient systems (Ch. 3). In: Macroscopic and Large Scale Phenomena: Coarse Graining, Mean Field Limits and Ergodicity, edited by A. Muntean, J. Rademacher and A. Zagaris. Proc. of Summer School in Twente University, June 2012. Vol. 3 of *Lecture Notes in Applied Math. Mechanics* Springer (2016) 187–249.
[42] A. Mielke and A. Stephan, Coarse-graining via EDP-convergence for linear fast-slow reaction systems. *Math. Models Methods Appl. Sci.* **30** (2020) 1765–1807.
[43] A. Mielke, M.A. Peletier and D.R. Michiel Renger, On the relation between gradient flows and the large-deviation principle, with applications to Markov chains and diffusion. *Potential Anal.* **41** (2014) 1293–1327.
[44] A. Mielke, R.I.A. Patterson, M.A. Peletier and D.R. Michiel Renger, Non-equilibrium thermodynamical principles for chemical reactions with mass-action kinetics. *SIAM J. Appl. Math.* **77** (2017) 1562–1585.
[45] J.J.H. Miller and S. Wang, An analysis of the Scharfetter–Gummel box method for the stationary semiconductor device equations. *ESAIM: M2AN* **28** (1994) 123–140.
[46] M.A. Peletier, R. Rossi, G. Savaré and O. Tse, Jump processes as generalized gradient flows. Preprint arXiv:2006.10624 (2020).
[47] D.L. Scharfetter and H.K. Gummel, Large-signal analysis of a silicon read diode oscillator. *IEEE Trans. Electron. Devices* **16** (1969) 64–77.
[48] K.B. Stolarsky, Generalizations of the logarithmic mean. *Math. Mag.* **48** (1975) 87–92.
[49] W.W. van Roosbroeck, Theory of the flow of electrons and holes in germanium and other semiconductors. *Bell Syst. Tech. J.* **29** (1950) 560–607.
[50] P.S. Vassilevski, S.I. Petrova and R.D. Lazarov, Finite difference schemes on triangular cell-centered grids with local refinement. *SIAM J. Sci. Stat. Comput.* **13** (1992) 1287–1313.
[51] M. Weber and N. Ernst, A fuzzy-set theoretical framework for computing exit rates of rare events in potential-driven diffusion processes. Preprint arXiv:1708.00679 (2017).
[52] Wolfram Mathematica, Version 11.1. Wolfram Research. Inc., Champagne, IL, USA (2017).
[53] J. Xu and L. Zikatanov, A monotone finite element scheme for converction-diffusion equations. *Math. Comput.* **68** (1999) 1429–1446.

Subscribe to Open (S2O)
A fair and sustainable open access model

This journal is currently published in open access under a Subscribe-to-Open model (S2O). S2O is a transformative model that aims to move subscription journals to open access. Open access is the free, immediate, online availability of research articles combined with the rights to use these articles fully in the digital environment. We are thankful to our subscribers and sponsors for making it possible to publish this journal in open access, free of charge for authors.

Please help to maintain this journal in open access!

Check that your library subscribes to the journal, or make a personal donation to the S2O programme, by contacting subscribers@edpsciences.org

More information, including a list of sponsors and a financial transparency report, available at: https://www.edpsciences.org/en/maths-s2o-programme