ON THE STABILITY OF CHARACTERISTIC SCHEMES FOR THE FERMI EQUATION

M. ASADZADEH

ABSTRACT. We study characteristic schemes for a model problem for the Fermi pencil beam equation. The objective is twofold: (i) To design efficient and accurate numerical schemes based on the principle of solving a particle transport problem, exactly, on each collision free spatial segment combined with a projection on each collision site, from a pre collision angle and energy coordinates (AE) to a post collision AE coordinates. (ii) To prove stability and derive a posteriori error estimates in L_2 and the maximum norms.

INTRODUCTION

The main feature in this paper can be described as follows: Consider a homogeneous infinite slab, $(y, z \in \mathbb{R})$, $\tilde{Q} = (x, y, z)$ of thickness L, $(0 < x < L)$. Let x be the penetration direction of a charged particle beam, $\{x_n\}$ an increasing sequence of discrete points indicating collision sites and $\{\mathcal{V}_n\}$ a corresponding sequence of piecewise polynomial spaces on meshes $\{\mathcal{T}_n\}$ on the transversal variable $x_\perp := (y, z)$. Then given the approximate solution (current) $J^{h,n} \in \mathcal{V}_n$ at the collision site x_n solve the pencil beam equation exactly on the collision free interval (x_n, x_{n+1}) with the data $J^{h,n}$ to give the solution $J_-^{h,n+1}$ at the next collision site x_{n+1}, before the collision. This is an exact transport procedure. Now one may compute $J^{h,n+1} = \mathcal{P}_{n+1} J_-^{h,n+1}$, with $\mathcal{P}_{n+1}$ being a projection into $\{\mathcal{V}_{n+1}\}$, here $J^{h,n+1}$ is the post-collision solution at the (other face of the collision) site x_{n+1}. Thus we have a process of *exact transport + projection*.

The idea of exact transport + projection was first introduced by Johnson, in [11], for the convection problems. Our goal is to extend this process to a simple case of a pencil beam model described by: *a forward-backward, convection dominated convection-diffusion equation of the degenerate type*. One may outline variety of approaches of this type differing in the choice of piecewise polynomial spaces $\{\mathcal{V}_n\}$ (degree of polynomials, orthogonality, continuity or discontinuity) and in the projections $\mathcal{P}_n$, (L_p-projections, $1 \le p \le \infty$, interpolation projections, etc).

Generally the *exact transport* problem, because of the presence of the diffusion term, in the pencil beam equations, if solvable, is highly nontrivial. Besides, simple projections as L_2-projection would create oscillatory behaviour in the presence of discontinuities. This cause a serious drawback in reliability of the beam algorithms in application, (e.g. in the radiative cancer therapy, dealing with discontinuities

1991 *Mathematics Subject Classification.* 65N15, 65N30, 35L80.

Key words and phrases. Fermi equation, pencil beam model, standard Galerkin, streamline diffusion, characteristic Galerkin, characteristic streamline diffusion, numerical diffusion, stability, maximum norm, error analysis.

Supported by TMR, EU contract no. ERB FMRX-CT97-0157.

such as air/tissue and tissue/bone interfaces). To circumvent these difficulties we present an approach leading to an *exact transport solver* for model cases of pencil beam problems by characteristic methods, associated with a modified L_2-projection raising the stability properties. Related studies of this type can be found in [2] and [3] for Vlasov-Poisson, and Fermi and Fokker-Planck equations, respectively.

Our method is obtained through two basic modifications of a standard Galerkin scheme: first, the test functions are modified so that to give a weighted least square control of the residual $\mathcal{R}$, (measuring how well the approximate solution satisfies the considered differential equation locally), of the approximate solution, and secondly artificial viscosity is added to the diffusion coefficient of the form $Ch^2|\mathcal{R}(J^h)|$, where h is the local mesh size. We shall also consider a variant of the streamline diffusion (SD)-method based on using trial functions which are discontinuous in the beams penetration direction x and continuous in the transversal variable $x_\perp = (y, z)$. Orienting the *incident-transversal* mesh approximately along the characteristics we get a particular SD-method suitable for convection dominated convection-diffusion problems referred as characteristic streamline diffusion (CSD).

To construct the numerical schemes, the domain $Q := I_x \times I_y \times I_z$ is subdivided into *slabs* $S_n := I_x^n \times I_y \times I_z$, with $I_x^n := (x_{n-1}, x_n)$, $n = 1, 2, \ldots, N$, corresponding to collision-free paths in the x-direction and I_y and I_z, bounded symmetric intervals representing the transversal domain of $x_\perp$. Each slab S_n has its own *incident-transversal* finite element mesh $\hat{\mathcal{T}}_n$. Consequently, at each collision site x_n we have two transversal meshes $\hat{\mathcal{T}}_n^- = \hat{\mathcal{T}}_n|_{x_n}$ and $\hat{\mathcal{T}}_n^+ = \hat{\mathcal{T}}_{n+1}|_{x_n}$, respectively. In general $\hat{\mathcal{T}}_n^- \neq \hat{\mathcal{T}}_n^+$ and the passage of information from one slab to the next is performed through a modified L_2-projection. The CSD-method performs this modified projection along with the *exact transport* solutions satisfying, in model cases, the convection equations exactly, and separately, on each slab.

An outline of this paper is as follows: In Section 1 we introduce a continuous model problem. In section 2 we formulate the characteristic schemes for the model problem. In Section 3 we derive L_2 estimates for smooth solutions. Section 4 is devoted to determine the amount of numerical dissipation introduced by the discretizations. Section 5 concerns with the study of stability in the maximum norm. In Section 6 we prove a posteriori error estimates underlying the adaptive algorithm and finally in our concluding Section 7 we introduce some numerical examples to justify smoothing properties of the new approaches. Throughout the paper C will denote an absolute constant not necessarily the same at each accurence.

1. A model Problem

In this section we sketch the derivation of the pencil beam equations and introduce a model problem. Detailed derivation strategy can be found in [10] relaying on Fourier techniques, [12] using spherical harmonics, and [6] based on statistical physics approaches. We start from the steady-state, monoenergetic transport equation in a homogeneous slab $\tilde{Q} :=\in [0, L] \times \mathbb{R} \times \mathbb{R}$, given by

$$\omega \cdot \nabla_{\mathbf{x}} \psi(\mathbf{x}, \omega) + \sigma_t(\mathbf{x})\psi(\mathbf{x}, \omega) = \int_{S^2} \sigma_s(\mathbf{x}, \omega \cdot \omega')\psi(\mathbf{x}, \omega)\, d\omega', \quad \text{in} \quad \tilde{Q} \times S^2, \tag{1.1}$$

and associated with the boundary conditions

$$\begin{cases} \psi(L,y,z,\omega) = 0, & \xi < 0, \\ \psi(0,y,z,\omega) = \frac{1}{2\pi}\delta(1-\xi)\delta(y)\delta(z), & \xi > 0, \end{cases} \tag{1.2}$$

with $\mathbf{x} = (x,y,z) \in \tilde{Q}$, $\omega = (\xi,\eta,\zeta) \in S^2$, describing the spreading of a pencil beam of particles normally incident upon the purely scattering, source-free, slab $\tilde{Q}$ of thickness L. Here ψ is the density of particles at the point $\mathbf{x}$ moving in the direction of ω, σ_t, and σ_s are total and scattering cross-sections, respectively. Assuming *forward peaked* scattering, the transport equation (1.1) may, asymptotically, be approximated by the following Fokker-Planck equation

$$\omega \cdot \nabla_{\mathbf{x}} \psi^{FP} = \sigma \left[\frac{\partial}{\partial \xi}(1-\xi^2)\frac{\partial}{\partial \xi} + \frac{1}{1-\xi^2}\frac{\partial^2}{\partial \vartheta^2} \right] \psi^{FP}, \tag{1.3}$$

where ϑ is the azimuthal angle with respect to the z-axis and

$$\sigma \equiv \frac{1}{2}\sigma_{tr}(\mathbf{x}) = \pi \int_{-1}^{1} (1-\xi)\sigma_s(\mathbf{x},\xi)\, d\xi, \tag{1.4}$$

is the transport cross-section for a purely scattering medium. In the asymptotic expansions leading to the Fokker-Planck equation the absorption term $\sigma_t\psi$ on the left-hand side of (1.1) associated with a Taylor expansion of ψ on the right-hand side would give the right-hand side of (1.3) and a neglected remainder term of order $\mathcal{O}(\sigma^2)$, see [6] for the details. A further approximation, assuming thin slab by letting

$$L \times \sigma \ll 1, \tag{1.5}$$

and performing some algebraic manipulations, see [6] and [10] yields a perturbation of (1.3) to the following Fermi equation;

$$\begin{cases} \omega_0 \cdot \nabla_{\mathbf{x}} \psi^F = \sigma \Delta_{\eta\zeta} \psi^F, \\ \psi^F(0,y,z,\eta,\zeta) = \delta(y)\delta(z)\delta(\eta)\delta(\zeta), & \xi > 0, \\ \psi^F(L,y,z,\eta,\zeta) = 0, & \xi < 0, \end{cases} \tag{1.6}$$

here $\omega_0 = (1,\eta,\zeta)$, where $(\eta,\zeta) \in \mathbb{R}\times\mathbb{R}$ and $\Delta_{\eta\zeta} = \partial^2/\partial\eta^2 + \partial^2/\partial\zeta^2$. Geometrically, the equation (1.6) corresponds to projecting $\omega \in S^2$ in the equation (1.3), along $\omega = (\xi,\eta,\zeta)$, on the tangent plane to S^2 at the point $(1,0,0)$. In this way the Laplacian operator, on the unit sphere, in the right-hand side of the Fokker-Planck equation (1.3) is transfered to the Laplacian operator on this tangent plane, as on the right-hand side of the Fermi equation (1.6).

The equations (1.3)-(1.6) are formulated for the flux ψ, while usually the measured quantity, e.g. in the radiation therapy applications (dose), is related to the current function

$$j = \xi\psi. \tag{1.7}$$

Now we consider a two dimensional version of Eqs. (1.1)-(1.3) leading to the following Fokker-Planck problem, see also [3]: For $0 < x < L$ and $-\infty < y < \infty$, find $\psi^{FP} \equiv \Psi^{FP}(x,y,\theta)$ such that

$$\begin{cases} \omega \cdot \nabla_{\mathbf{x}} \psi^{FP} = \sigma \psi^{FP}_{\theta\theta}, & \theta \in (-\pi/2, \pi/2), \\ \psi^{FP}(0,y,\theta) = \frac{1}{2\pi}\delta(1-\cos\theta)\delta(y), & \theta \in S^1_+, \\ \psi^{FP}(L,y,\theta) = 0, & \theta \in S^1_-, \end{cases} \tag{1.8}$$

where $\omega = (\xi,\eta) \equiv (\cos\theta, \sin\theta)$, $S^1_+ = \{\omega \in S^1 : \xi > 0\}$ and $S^1_- = S^1 \setminus S^1_+$.

We use the scaling substitution

$$z = \tan\theta, \quad \theta \in (-\pi/2, \pi/2), \tag{1.9}$$

and introduce the scaled current function J as:

$$J(x, y, z) \equiv j(x, y, \tan^{-1} z)/(1 + z^2). \tag{1.10}$$

Note that, now z corresponds to the angular variable θ. Below we shall keep θ away from the poles $\pm\pi/2$, and correspondingly formulate a problem for the current function J, in the bounded domain $Q \equiv I_x \times I_y \times I_z = [0, L] \times [-y_0, y_0] \times [-z_0, z_0]$:

$$\begin{cases} J_x + zJ_y = \sigma AJ, & (x, x_\perp) \in Q, \\ J_z(x, y, \pm z_0) = 0, & \text{for} \quad (x, y) \in I_x \times I_y, \\ J(x, \pm y_0, z) = 0, & \text{on} \quad \Gamma^-_{\tilde\beta} \setminus \{\text{supp} f\}, \\ J(0, x_\perp) = f(x_\perp), \end{cases} \tag{1.11}$$

where $\Gamma^-_{\tilde\beta} := \{(x, x_\perp) \in \partial Q : \tilde\beta \cdot \mathbf{n} < 0\}$, $\tilde\beta = (1, z, 0)$, $x_\perp \equiv (y, z)$ is the transversal variable and $\mathbf{n} := \mathbf{n}(x, x_\perp)$ is the outward unit normal to Γ at $(x, x_\perp) \in \Gamma$. Further we have replaced the product of δ-functions (the source term) at the boundary by a smoother L_2-function f. The diffusion operator in (1.11) is:

$$A = \partial^2/\partial z^2, \qquad \textit{(Fermi)}, \tag{1.12}$$

$$A\cdot = \partial/\partial z[a(z)\partial/\partial z\,(b(z)\cdot)], \qquad \textit{(Fokker-Planck)} \tag{1.13}$$

where $a(z) = 1 + z^2$ and $b(z) = (1 + z^2)^{3/2}$. We shall study the Fermi equation. The Fokker-Planck case follows , basically, the same idea except somewhat involved algebraic labour and therefor is omitted. Detailed Fokker-Planck studies can be found in [3]. We note that the transport cross section depends on energy and therefor on the spatial variables: $\sigma \equiv \sigma(x, y) = 1/2\sigma_{tr}(E(x, y))$.

The equation (1.11) is a forward-backward (z changes the sign), convection dominating (σ is small), convection diffusion equation of degenerate type (convection in (x, y) and diffusion in z). A corresponding non-degenerate equation can be formulated as follows:

$$\mathcal{L}(J) := J_x + \beta \cdot \nabla_\perp J - \varepsilon\Delta_\perp J = 0, \tag{1.14}$$

where $\varepsilon \approx C\sigma/2 = C\sigma_{tr}/4$, $C \approx (C_1 + C_2)/2$, $\Delta_\perp := \partial^2/\partial y^2 + \partial^2/\partial z^2$, is the transversal Laplacian operator, and from now on $\beta \equiv (z, 0)$. In our studies below A is given by (1.12) corresponding to the Fermi equation.

2. Characteristic Schemes

We focus on the equation (1.14) and introduce the change of coordinates $(x, \bar x_\perp) = (x, x_\perp - x\beta)$. Then writing $\bar J(x, \bar x_\perp) = J(x, x_\perp)$, we can reformulate the equation (1.14) as:

$$\bar J_x - \varepsilon\Delta_\perp \bar J = 0, \quad \textit{in} \quad [0, L] \times I_y \times I_z, \qquad \bar J(0, \bar x_\perp) = f(x_\perp), \tag{2.1}$$

since $\frac{\partial \bar J}{\partial x} = \frac{\partial}{\partial x} J(x, \bar x_\perp + x\beta) = \frac{\partial J}{\partial x} + \beta \cdot \nabla_\perp J$. If $\varepsilon = 0$, then, recalling the boundary data in (1.11), the solution of (2.1) is given by $\bar J(x, \bar x_\perp) = f(\bar x_\perp)$ and that of (1.14) by

$$J(x, x_\perp) = f(x_\perp - x\beta). \tag{2.2}$$

Clearly the characteristics of equation (1.14) with $\varepsilon = 0$ are given by $x_\perp = \bar{x}_\perp + x\beta$, $x > 0$, and in this case ($\varepsilon = 0$) the solution $J(x, x_\perp)$ is constant along the characteristics. Let now $\{x_n\}$, $n = 0, 1, \ldots, N$, be an increasing sequence of x-values with $x_0 = 0$, and let for each $0 < n \leq N$, $\{\mathcal{T}_n\}$ be a corresponding sequence of triangulation $\mathcal{T}_n$ of $\{x_n\} \times I_y \times I_z$ into triangles K and let $\mathcal{V}_n$ be the space of continuous piecewise linear functions on $\mathcal{T}_n$, i.e. $\mathcal{V}_n = \{v \in \mathcal{C}(I_y \times I_z) : v \text{ is linear on } K, K \in \mathcal{T}_n\}$. Here and below $\mathcal{C}(\Omega)$ denotes the set of continuous functions on Ω.

The *Characteristic Galerkin method* for (1.14), in the case of $\varepsilon = 0$, is formulated as follows: For $n = 1, 2, \ldots, N$, find $J^{h,n} \in \mathcal{V}_n$ such that

(2.3)
$$\int_{I_y \times I_z} J^{h,n}(x_\perp) v(x_\perp)\, dx_\perp = \int_{I_y \times I_z} J^{h,n-1}(x_\perp - \hbar_n \beta) v(x_\perp)\, dx_\perp, \quad \forall v \in \mathcal{V}_n,$$

where $\hbar = x_n - x_{n-1}$ and $J^{h,0} = f$. In other words

$$J^{h,n} = \mathcal{P}_n T_n J^{h,n-1}, \tag{2.4}$$

where $\mathcal{P}_n : L_2(I_y \times I_z) \to \mathcal{V}_n$ is the L_2-projection defined by $(\mathcal{P}_n w, v) = (w, v), \quad \forall v \in \mathcal{V}_n$, where $(\cdot, \cdot)$ denotes the inner product in $L_2(I_y \times I_z)$, and $T_n v(x_\perp) = v(x_\perp - \hbar_n \beta)$. Thus (2.4) may be expressed as *exact transport* T_n+ projection $\mathcal{P}_n$.

Next we formulate the *streamline diffusion* (SD)-method, and the *characteristic streamline diffusion* (CSD) method (as a special case with oriented phase-space mesh elements) for the equation (1.14) as follows: For $n = 1, 2, \ldots, N$, let $\hat{\mathcal{T}}_n = \{\hat{K}\}$ be a finite element subdivision of the slab $S_n = I_x^n \times I_\perp$, $I_x^n = (x_{n-1}, x_n)$, $I_\perp = I_y \times I_z$, into elements $\hat{K}$ and let $\hat{\mathcal{V}}_n$ be a space of continuous piecewise polynomials on $\hat{\mathcal{T}}_n$ of degree at most k. For $k = 1$ and small ε the SD-method may be formulated as follows: For $n = 1, 2, \ldots, N$, find $\hat{J}^h \equiv \hat{J}^h|_{S_n} \in \hat{\mathcal{V}}_n$ such that

$$\begin{aligned} &\int_{S_n} (\hat{J}_x^h + \beta \cdot \nabla_\perp \hat{J}^h)\,(v + \delta(v_x + \beta \cdot \nabla_\perp v))\; dx dx_\perp \\ &\quad + \int_{S_n} \hat{\varepsilon} \nabla_\perp \hat{J}^h \cdot \nabla_\perp v\, dx dx_\perp + \int_{I_\perp} \hat{J}_+^{h,n} v_+^n\, dx_\perp \\ &\quad = \int_{I_\perp} \hat{J}_-^{h,n} v_+^n\, dx_\perp, \quad \forall v \in \hat{\mathcal{V}}_n, \end{aligned} \tag{2.5}$$

where $v_\pm^n(x_\perp) = \lim_{\Delta x \to 0^+} v(x \pm \Delta x, x_\perp)$, $\hat{\varepsilon} = \max(\varepsilon, \mathcal{F}(Ch^\alpha \mathcal{R}(\hat{J}^h))/M_n)$, with

$$\mathcal{R}(\hat{J}^h) = |\hat{J}_x^h + \beta \cdot \nabla_\perp \hat{J}^h| + |[\hat{J}^h]|/\hbar_n, \qquad on \quad S_n, \tag{2.6}$$

where $[v^n] = v_+^n - v_-^n$, $\mathcal{F}(v)$ is the element-wise average of v and δ is a small parameter in general of order $\mathcal{O}(h)$ locally and $\alpha = 2 - \kappa$, κ small and positive. Here $h(x, x_\perp)$ is a continuous function measuring the local size of elements $\hat{K} \in \hat{\mathcal{T}}_n$. Further $M_n = \max_{x_\perp} |J_+^{h,n}(x_\perp)|$, is a normalization factor. Note that equation (2.5) is nonlinear in $\hat{J}^h|_{S_n}$ since $\hat{\varepsilon}$ depends on $\hat{J}^h$. By a fixed point argument using monotonicity, it is possible to show the existence of a solution to the equation (2.5). The streamline diffusion modification is given by $\delta(v_x + \beta \cdot \nabla_\perp v)$ and the degenerate-shock-capturing modification by $\hat{\varepsilon}$. Approximating β by piecewise constants on each slab, the streamline diffusion modification will disappear in the CSD-method.

We now make a special choice of the finite element subdivision $\hat{\mathcal{T}}_n = \{\hat{K}\}$ of S_n and the corresponding finite element space $\hat{\mathcal{V}}_n$ to obtain the CSD-method. Let $\hat{\mathcal{T}}_n = \{\hat{K}\}$ be a subdivision of S_n given by the prismatic elements oriented along the characteristics

$$\hat{K}_n = \{(x, \bar{x}_\perp + (x - x_n)\beta) : \bar{x}_\perp \in K \in \mathcal{T}_n,\ x \in I_x^n\}, \tag{2.7}$$

where $\mathcal{T}_n = \{K\}$ is a triangulation of $I_\perp$ given above. Further, let $\hat{\mathcal{V}}_n$ be defined by

$$\hat{\mathcal{V}}_n = \{\hat{v} \in \mathcal{C}(S_n) : \hat{v}(x, x_\perp) = v(x_\perp - (x - x_n)\beta),\ v \in \mathcal{V}_n\}, \tag{2.8}$$

with $\mathcal{V}_n$ the space of continuous piecewise linear functions on $\mathcal{T}_n$ as above. In other words $\hat{\mathcal{V}}_n$ consists of the continuous functions $\hat{v}(x, x_\perp)$ on S_n such that $\hat{v}$ is constant along characteristics $x_\perp = \bar{x}_\perp + x\beta$ parallel to the sides of the prismatic elements $\hat{K}_n$ and v_+^n is piecewise linear on $\mathcal{T}_n$ for $x = x_n$. With this choice the SD-method (2.5) reduces to the following method since $\frac{\partial \hat{v}}{\partial x} + \beta \cdot \nabla_\perp \hat{v} = 0$ if $\hat{v} \in \hat{\mathcal{V}}_n$: For $n = 1, 2, \ldots, N$, find $\hat{J}^h \equiv \hat{J}^h|_{S_n} \in \hat{\mathcal{V}}_n$ such that

$$\int_{S_n} \hat{\varepsilon} \nabla_\perp \hat{J}^h \cdot \nabla_\perp v \, dx dx_\perp + \int_{I_\perp} \hat{J}_+^{h,n} v_+^n \, dx_\perp = \int_{I_\perp} \hat{J}_-^{h,n} v_+^n \, dx_\perp, \quad \forall \hat{v} \in \hat{\mathcal{V}}_n, \tag{2.9}$$

where

$$\hat{\varepsilon} = \max\left(\varepsilon, \mathcal{F}(Ch^\alpha \frac{|[\hat{J}^{h,n}]|}{\hbar_n})/M_n\right), \quad on \quad S_n,$$

and $h(x, x_\perp) = h_n(x_\perp - (x - x_n)\beta)$, where $h_n(x_\perp)$ gives the local element size of $\mathcal{T}_n$. If now ε is small, then (2.9) may be written as

$$\int_{I_\perp} \tilde{\varepsilon} \nabla_\perp \hat{J}_+^{h,n} \cdot \nabla_\perp v \, dx_\perp + \int_{I_\perp} \hat{J}_+^{h,n} v \, dx_\perp = \int_{I_\perp} \hat{J}_-^{h,n} v \, dx_\perp, \quad \forall v \in \mathcal{V}_n, \tag{2.10}$$

where $\tilde{\varepsilon} = \mathcal{F}(Ch^\alpha |[\hat{J}^{h,n}]|)/M_n$. Writing $\hat{J}_+^{h,n} = J^{h,n}$, we can thus restate (2.9) as follows (since $\hat{J}_-^{h,n} = T_n J^{h,n-1}$): For $n = 1, 2, \ldots, N$, find $J^{h,n} \in \mathcal{V}_n$ such that

$$\int_{I_\perp} \tilde{\varepsilon} \nabla_\perp J^{h,n} \cdot \nabla_\perp v \, dx_\perp + \int_{I_\perp} J^{h,n} v \, dx_\perp = \int_{I_\perp} T_n J^{h,n-1} v \, dx_\perp, \quad \forall v \in \mathcal{V}_n, \tag{2.11}$$

where $J^{h,0} = f$ and $\tilde{\varepsilon} = \mathcal{F}(Ch_n^\alpha |J^{h,n} - T_n J^{h,n-1}|)/M_n$. Introducing the operator $\tilde{\mathcal{P}}_n : L_2(I_\perp) \cap L_\infty(I_\perp) \to \mathcal{V}_n$ defined by

$$(\tilde{\mathcal{P}}_n w, v) + (\tilde{\varepsilon} \nabla_\perp \tilde{\mathcal{P}}_n w, \nabla_\perp v) = (w, v), \quad \forall v \in \mathcal{V}_n, \tag{2.12}$$

where $\tilde{\varepsilon} = \mathcal{F}(Ch_n^\alpha |\tilde{\mathcal{P}}_n w - w|)/\max|\tilde{\mathcal{P}}_n w|$, and $(\cdot, \cdot)$ denotes the $L_2(I_\perp)^m$ inner product with $m = 1, 2$, we can write (2.11) as

$$J^{h,n} = \tilde{\mathcal{P}}_n T_n J^{h,n-1}. \tag{2.13}$$

Obviously, $\tilde{\mathcal{P}}_n$ may be viewed as a modification of the usual L_2-projection $\mathcal{P}_n : L_2(I_\perp) \to \mathcal{V}_n$ defined above by $(\mathcal{P}_n w, v) = (w, v)$, $\forall v \in \mathcal{V}_n$, obtained by adding an artificial viscosity term with coefficient $\tilde{\varepsilon} = \mathcal{F}(Ch_n^\alpha |\tilde{\mathcal{P}}_n w - w|)/\max|\tilde{\mathcal{P}}_n w|$.

Note that the mesh size h_n of the triangulation $\mathcal{T}_n$ may vary with $x_\perp$ (and, evidently also, with n); it is reasonable to require that $|\nabla_\perp h_n(x_\perp)| \le c$, $x_\perp \in I_\perp$, where c is a sufficiently small constant and assume that $|K| \sim h_n(x_\perp)$, if $x_\perp \in K \in \mathcal{T}_n$. For simplicity we assume in this note that $\mathcal{T}_n$ is quasiuniform so that

we may take h_n constant. The extensions to the general non-uniform mesh is straightforward.

3. Error estimates for smooth solutions

In this section we give the standard error estimates for the characteristic Galerkin method (CG) (2.4) and the CSD-method (2.13), in the case of a smooth exact solution. For CSD, in this case, we may choose $\tilde{\varepsilon} = 0$ in (2.12) so that (2.4) and (2.13) indeed coincide. Our point is that using the CSD-approach we obtain sharper results than through the standard CG-approach, as we shall now see.

Starting with error estimates for the CG-method we have for $J^n = J(x_n, \cdot)$ that

$$\begin{aligned} \|J^n - J^{h,n}\| &\leq \|T_n J^{n-1} - \mathcal{P}_n T_n J^{h,n-1}\| \\ &\leq \|T_n J^{n-1} - \mathcal{P}_n T_n J^{n-1}\| + \|\mathcal{P}_n T_n J^{n-1} - \mathcal{P}_n T_n J^{h,n-1}\| \\ &\leq Ch_n^2 \|J^{n-1}\|_{H^2(I_\perp)} + \|J^{n-1} - J^{h,n-1}\|, \end{aligned}$$

using a standard error estimate for $\mathcal{P}_n$ of the form $\|w - \mathcal{P}_n w\| \leq Ch_n^2 \|w\|_{H^2(I_\perp)}$, the boundedness of $\mathcal{P}_n : L_2 \to L_2$ in the form $\|\mathcal{P}_n w\| \leq \|w\|$ and the fact that $\|T_n w\| = \|w\|$. By iteration we get

$$\|J^N - J^{h,N}\| \leq \sum_{n=1}^{N} Ch_n^2 \|J^{n-1}\|_{H^2(I_\perp)} = \mathcal{O}(Nh^2), \tag{3.1}$$

if $h_n \sim h$ for all n and J is smooth.

The standard error estimate for the SD-method (2.5) with $\hat{\mathcal{V}}_n$ given by (2.8) and with $\hat{\varepsilon} = 0$, (see [3] and [11]), states that

$$\begin{aligned} &\|J^N - J^{h,N}\| + \left(\sum_{n=1}^{N} \|J^{h,n} - T_n J^{h,n-1}\|^2 \right)^{1/2} \\ &\qquad \leq \left(\sum_{n=1}^{N} Ch_{n-1}^4 \|J^{n-1}\|_{H^2(I_\perp)}^2 \right)^{1/2} \leq C\sqrt{N}h^2, \end{aligned} \tag{3.2}$$

if J is smooth, which is clearly sharper than (3.1). To prove the estimate (3.2) for (2.4) we note that with $e^{h,n} = J^{h,n} - J^n$, we have by (2.3) for $n = 1, 2, \ldots, N$,

$$(e^{h,n} - T_n e^{h,n-1}, v) = 0, \qquad \forall v \in \mathcal{V}_n. \tag{3.3}$$

Now since $\|T_n e^{h,n-1}\| = \|e^{h,n-1}\|$, we have

$$\begin{aligned} \frac{1}{2}\|e^{h,N}\|^2 + \frac{1}{2}\sum_{n=1}^{N} \|e^{h,n} - T_n e^{h,n-1}\|^2 &= \sum_{n=1}^{N} (e^{h,n} - T_n e^{h,n-1}, e^{h,n}) + \frac{1}{2}\|e^{h,0}\|^2 \\ &= \sum_{n=1}^{N} (e^{h,n} - T_n e^{h,n-1}, J^n - \mathcal{P}_n J^n) \\ &\leq \frac{1}{4}\sum_{n=1}^{N} \|e^{h,n} - T_n e^{h,n-1}\|^2 + \sum_{n=1}^{N} \|J^n - \mathcal{P}_n J^n\|^2, \end{aligned} \tag{3.4}$$

where we used Eq. (3.3) with $v = \mathcal{P}_n J^n - J^{h,n}$, the fact that $e^{h,0} = 0$, and Cauchy's inequality. Recalling now the above standard estimate for $\|J^{h,n} - \mathcal{P}_n J^n\|$, we obtain (3.2).

Note that the stability estimate for (2.4) underlying (3.1) and (3.2), respectively, are as follows

$$\|J^{h,n}\| \le \|f\|, \quad n = 1, 2, \ldots, N, \tag{3.5}$$

$$\|J^{h,N}\|^2 + \sum_{n=1}^{N} \|J^{h,n} - T_n J^{h,n-1}\|^2 = \|f\|^2, \tag{3.6}$$

where (3.5) reflects that $\|\mathcal{P}_n w\| \le \|w\|$ and $\|T_n w\| = \|w\|$ for $w \in L_2(I_\perp)$, and Eq. (3.6) follows by choosing $v = J^{h,n}$ in (2.3) and noting as in (3.4) that

$$\begin{aligned} \frac{1}{2}\|J^{h,N}\|^2 \quad &+ \quad \frac{1}{2}\sum_{n=1}^{N} \|J^{h,n} - T_n J^{h,n-1}\|^2 \\ &= \quad \sum_{n=1}^{N} (J^{h,n} - T_n J^{h,n-1}, J^{h,n}) + \frac{1}{2}\|f\|^2 = \frac{1}{2}\|f\|^2. \end{aligned}$$

The improvement using Eq. (3.6) indicates that the classical stability concept based on (3.5) is not fully adequate. To obtain sharp results it seems to be necessary, and also natural, to include dissipation terms in the stability estimates.

The estimate (3.2) is sharp as an estimate for $\|J^N - J^{h,N}\|$; for the discontinuous Galerkin method with piecewise linears, which corresponds to (2.4) with $\mathcal{P}_n$ being the L_2-projection onto the piecewise linears, in [3], we have shown that in general the error $\|J^N - J^{h,N}\|$ with $N = \mathcal{O}(h^{-1})$, $\hbar = \mathcal{O}(h)$, is not better than $\mathcal{O}(h^{3/2})$ which corresponds to (3.2) with $N = \mathcal{O}(h^{-1})$.

To sum up, we get for Eq. (2.4) with the standard CG-approach , $\|J^N - J^{h,N}\| = \mathcal{O}(Nh^2)$, while the more careful analysis in the SD-approach gives $\|J^N - J^{h,N}\| = \mathcal{O}(\sqrt{N}h^2)$. With $N = \mathcal{O}(h^{-1})$, we thus have $\|J^N - J^{h,N}\| = \mathcal{O}(h)$ with the CG-approach and $\|J^N - J^{h,N}\| = \mathcal{O}(h^{3/2})$ with the SD-approach if the exact solution J is sufficiently smooth, i.e. in the Sobolev space H^2.

4. Numerical Diffusion

We shall now seek quantitative estimates for the dissipation in (2.4), i.e., the CG-method or equivalently the CSD-method without the shock-capturing perturbation, and in the CSD-method (2.13) with shock-capturing, i.e. $\tilde{\varepsilon} \neq 0$.

For Eq. (2.3) using (3.6) we have

$$\|J^{h,N}\|^2 + D_N = \|f\|^2, \tag{4.1}$$

where

$$D_N = \sum_{n=1}^{N} \|J^{h,n} - T_n J^{h,n-1}\|^2, \tag{4.2}$$

may be taken as a quantitative measure for the dissipation. Introducing $T_n \mathcal{V}_{n-1} = \{T_n v : v \in \mathcal{V}_{n-1}\}$, we have $T_n J^{h,n-1} \in T_n \mathcal{V}_{n-1}$ and to estimate D_N we are led to estimate $\|J^{h,n} - T_n J^{h,n-1}\| = \|(\mathcal{P}_n - I)w\|$ with $w = T_n J^{h,n-1} \in T_n \mathcal{V}_{n-1}$, i.e., the L_2-error in the L_2-projection of a piecewise linear function $T_n J^{h,n-1}$ on one

mesh $T_n\mathcal{V}_{n-1}$ onto a set of piecewise linears $\mathcal{V}_n$ on a different mesh. Obviously, by standard estimates we have for $w \in T_n J^{h,n-1}$ the following first order estimate:

$$\|(\mathcal{P}_n - I)w\| \leq Ch_n \|w\|_{H^1(I_\perp)}, \tag{4.3}$$

with no standard second order counterpart since $w \notin H^2(I_\perp)$ if $w \in T_n\mathcal{V}_{n-1}$. However, there is in fact a second order analogue of (4.3) available which takes the form:

$$\|(\mathcal{P}_n - I)w\| \leq C(h_n^2 + h_{n-1}^2)^2 \|\Delta_{\perp,n-1} w\|, \tag{4.4}$$

where $\Delta_{\perp,n-1} : H^1(I_\perp) \to T_n\mathcal{V}_{n-1}$ is a discrete Laplacian operator defined by $-(\Delta_{\perp,n-1}\varphi, v) = (\nabla_\perp \varphi, \nabla_\perp v),\ \forall v \in T_n\mathcal{V}_{n-1}$, see [4].

Inserting the estimate (4.4) into (4.2) we obtain assuming $h_n \leq h$,

$$D_N \leq C \sum_{n=1}^{N} \frac{h^4}{\hbar_n} \|\Delta_{\perp,n-1} T_n J^{h,n-1}\|^2 \hbar_n. \tag{4.5}$$

With $\hbar_n = h$ the inequality (4.5) suggests that the dissipation in (2.4) corresponds to adding a diffusion term of the form $ch^3\Delta_\perp^2 J$ to the continuous equation. In particular for smooth solution it appears that (2.4) adds little diffusion as compared to a first order upwind scheme with a corresponding continuous diffusion term of the form $Ch\Delta_\perp J$ with much larger diffusion coefficient. Thus, (2.4) does not appear to add excessive numerical diffusion unless of course we take $\hbar_n$ small compared to h_n, so that very many L_2-projections of different meshes will be performed. On the other hand in some sense (2.4) contains too little numerical diffusion since oscillations may occur at discontinuities of the exact solution.

We now turn to the CSD-method (2.13) which obviously adds more numerical diffusion than the CG-method due to modification on ε-term. The stability estimate corresponding to (4.1) in this case takes the form

$$\|J^{h,N}\|^2 + \tilde{D}_N = \|f\|^2, \tag{4.6}$$

where

$$\tilde{D}_N = D_N + 2\sum_{n=1}^{N} \int \tilde{\varepsilon} |\nabla_\perp J^{h,n}|^2 \, dx_\perp, \tag{4.7}$$

where $\tilde{\varepsilon} = \mathcal{F}(Ch_n^\alpha |J^{h,n} - T_n J^{h,n-1}|)/M_n$. It follows that the shock-capturing term in the CSD-method corresponds to adding a viscous term of the form $-div(\hat{\varepsilon}\nabla_\perp J)$ to the continuous equation with $\hat{\varepsilon} = \tilde{\varepsilon}/\hbar_n$ in S_n. If the exact solution is smooth, we expect by (4.5) to have $\hat{\varepsilon} = \mathcal{O}(h^3)$ if $h_n \leq h$ and $\hbar_n = h$, $(\alpha = 2)$, i.e. the same amount of viscosity without the perturbation. However, close to discontinuity of J (assuming f is discontinuous) we may have $|J^{h,n} - T_n J^{h,n-1}| = \mathcal{O}(1)$ at least for n small, and then $\hat{\varepsilon} = \mathcal{O}(1)$, i.e., the shock-capturing term may add significant additional numerical diffusion in regions of non-smoothness of the exact solution.

5. Stability in the Maximum norm

The stability, in the maximum norm, for the CSD-method being a particular SD-method reads as follows: For a given $L > 0$ there is a constant C such that if $J^{h,n},\ n = 1, 2, \ldots, N$ satisfies (2.13), then if $x_n \leq L$ we have

$$\|J^{h,n}\|_\infty \leq C\|f\|_\infty, \tag{5.1}$$

where $\|v\|_\infty = \sup_{x_\perp \in I_\perp} |v(x_\perp)|$. The estimate (5.1) may alternatively be expressed as follows

$$\|J^{h,n}\|_p \le \|J^{h,n-1}\|_p, \quad \textit{if} \quad p \le ch^{-\kappa/4}, \tag{5.2}$$

where $\kappa = 2 - \alpha > 0$ appears in the definition of $\hat{\varepsilon}$ in (2.5), c is a sufficiently small constant, and $\|\cdot\|_p$ denotes the $L_p(I_\perp)$-norm:

$$\|v\|_p = \left(\int_{I_\perp} |v(x)|^p \, dx_\perp \right)^{1/p}, \quad p \ge 1. \tag{5.3}$$

More specifically, (5.1) follows from (5.2) by an inverse estimate letting $p \to \infty$. To prove (5.2) the essential step is to choose in (2.11), $v = \pi_n \left((J^h)^{p-1} \right)$, where p is an even natural number, and $\pi_n : \mathcal{C}(I_\perp) \to \mathcal{V}_n$ is the standard nodal interpolation operator, so that we get

$$\begin{aligned} \int_{I_\perp} (J^{h,n})^p \, dx_\perp + \int_{I_\perp} \tilde{\varepsilon} \nabla_\perp J^{h,n} \cdot \nabla_\perp \left(\pi_n ((J^{h,n})^{p-1}) \right) \, dx_\perp \\ = \int_{I_\perp} T_n J^{h,n-1} (J^{h,n})^{p-1} \, dx_\perp + E_n, \end{aligned} \tag{5.4}$$

where

$$E_n = \int_{I_\perp} (J^{h,n} - T_n J^{h,n-1}) \left((J^{h,n})^{p-1} - \pi_n ((J^{h,n})^{p-1}) \right) \, dx_\perp. \tag{5.5}$$

Now by standard interpolation error estimates

$$|E_n| \le Cp^2 \int_{I_\perp} |J^{h,n} - T_n J^{h,n-1}| h_n^2 |\nabla_\perp J^{h,n}|^2 \|J^{h,n}\|_{\infty,K}^{p-3} \, dx_\perp, \tag{5.6}$$

where $\|v\|_{\infty,K} = \sup_{x_\perp \in K} |v(x_\perp)|$ on K. On the other hand we have for some constant c independent of $p = 2m$, $m = 1, 2, \ldots$, $n = 1, 2, \ldots, N$,

$$\begin{aligned} \int_{I_\perp} \tilde{\varepsilon} \nabla_\perp J^{h,n} \cdot \nabla_\perp \left(\pi_n ((J^{h,n})^{p-1}) \right) \, dx_\perp \\ \ge \frac{c}{p^2} \int_{I_\perp} \tilde{\varepsilon} |\nabla_\perp J^{h,n}|^2 \|J^{h,n}\|_{\infty,K}^{p-2} \, dx_\perp. \end{aligned} \tag{5.7}$$

For simplicity we now assume that $\tilde{\varepsilon}$ is defined slightly differently compared to the above, assuming now that $M_n = 1 + \|J^{h,n}\|_{\infty,K}$ on $K \in \mathcal{T}_n$, in which case $|E_n|$ is dominated by the right hand side of (5.7) so that recalling (5.4):

$$\int_{I_\perp} (J^{h,n})^p \, dx_\perp \le \int_{I_\perp} T_n J^{h,n-1} (J^{h,n})^{p-1} \, dx_\perp, \quad \textit{if} \quad p \le ch^{-\kappa/4}, \tag{5.8}$$

with c sufficiently small. Finally, (5.2) now follows by applying Hölder inequality to (5.8). Note that the proof of the crucial estimate (5.7) is carried out "element-wise" and uses in an essential way that $\mathcal{V}_n$ consists of piecewise linears.

6. A Posteriori Error Estimates

We return to the original problem and derive a posteriori error estimates. We shall use the high accuracy and good stability features of the streamline diffusion Galerkin method, studied in [2], based on

a) A phase-space discretization based on piecewise polynomial approximation with basis functions being continuous in $x_\perp$ and discontinuous in x. (Discontinuity in all variables are considered in [3]).

b) A *streamline diffusion* modification of the test function giving a weighted least square control of the residual $\mathcal{R}(J^h) = \mathcal{L}(J^h)$ of the finite element solution J^h.

c) Modification of the transport cross-section $\sigma_{tr} = 2\sigma$ so that an artificial transport cross-section $\hat{\sigma}_{tr}$ is obtained modifying ε as

$$\hat{\varepsilon}(x, x_\perp) = \max\left(\varepsilon(x,y), c_1 h\mathcal{R}(J^h)/|\nabla_\perp J^h|, c_2 h(x,x_\perp)^{3/2}\right), \tag{6.1}$$

where h is a total mesh-size and $c_i,\ i = 1, 2$ are sufficiently small constants. For the original degenerate problem $\hat{\varepsilon}$ is defined by replacing ε in (6.1) by σ. With a simplified form of the artificial transport cross-section as

$$\hat{\varepsilon} = \max(\varepsilon, c_1 h), \tag{6.2}$$

the SD-modification b) may be omitted. The a *posteriori* error estimate underlying the adaptive algorithm is, in the case of discretizing in the transversal variable $(y, z) = x_\perp$ only, basically as follows:

$$\|\hat{e}_h\|_Q \leq C^s C^i \|\hat{\varepsilon}^{-1} h^2 \mathcal{R}(J^h)\|_Q, \tag{6.3}$$

where $\hat{e}_h = \hat{J} - J^h$, with $\hat{J}$ being the solution of (1.14) with ε replaced by $\hat{\varepsilon}$ and

$$e = J - J^h = (J - \hat{J}) + (\hat{J} - J^h) := \hat{e} + \hat{e}_h. \tag{6.4}$$

Note that $J - \hat{J}$ is a perturbation error caused by changing ε to $\hat{\varepsilon}$ in the continuous problem (1.14). Further C^s is a stability constant, C^i is an interpolation constant and $\|\cdot\|_Q$ is the usual $\|\cdot\|_{L_2(Q)}$-norm. In the simplified case (6.2) the error estimate (6.3) takes the form

$$|\|\hat{e}_h|\|_Q \leq C^s C^i \|h\mathcal{R}(J^h)\|_Q. \tag{6.5}$$

The adaptive algorithm is based on (6.3) and seeks to find a mesh with as few degrees of freedom as possible such that for a given tolerance TOL> 0,

$$C^s C^i \|\hat{\varepsilon}^{-1} h^2 \mathcal{R}(J^h)\|_Q \leq TOL, \tag{6.6}$$

which, through (6.3), would L_2-bound $\hat{e}^h$. To control the remaining part of the error; i.e., $\hat{e} = J - \hat{J}$, we may adaptively refine the mesh until $\hat{\varepsilon} = \varepsilon$, giving $J = \hat{J}$, or alternatively approximate $\hat{e}$ in terms of $\hat{\varepsilon} - \varepsilon$. To approximately minimize the total number of degrees of freedom of a mesh with mesh size $(x, x_\perp)$ satisfying (6.6), typically a simple iterative procedure is used where a new mesh-size is computed by equidistribution of element contributions in the quantity $C^s C^i \|\hat{\varepsilon}^{-1} h^2 \mathcal{R}(J^h)\|_Q$ with the values of $\hat{\varepsilon}$ and $\mathcal{R}(J^h)$ taken from the previous mesh.

The structure of the proof of the a posteriori error estimate (6.5) is as follows:

i) Representation of the error $\hat{e}_h$ in terms of the residual $\mathcal{R}(J^h)$ and the solution ψ of a dual problem with $\hat{e}_h$ as right hand side.

ii) Use of the Galerkin orthogonality to replace ψ by $\psi - \Psi$, where Ψ is a finite element interpolant of ψ.

iii) Interpolation error estimates for $\psi - \Psi$ in terms of certain derivative $\mathcal{D}\psi$ of ψ and the mesh-size h.

iv) Strong stability estimate for the dual solution ψ estimating $\mathcal{D}\psi$ in terms of the data $\hat{e}_h$ of the dual problem.

Below we specify the steps i)-iv). Recall that $\hat{J}$ satisfies

$$\left\{\begin{array}{ll} \hat{J}_x + \beta\cdot\nabla_\perp \hat{J} - \hat{\varepsilon}\Delta_\perp \hat{J} = 0, & \text{in} \quad Q,\\ \hat{J}(0,x_\perp) = f(x_\perp), & \text{for} \quad x_\perp \in I_y\times I_z,\\ \hat{J}_z(x,y,\pm z_0) = 0, & \text{for} \quad (x,y)\in [0,L]\times I_y,\\ \hat{J}(0,\pm y_0,z) = 0, & \text{for} \quad z\in \Gamma_0^-, \end{array}\right. \tag{6.7}$$

with $\Gamma_0^- = \Gamma^-\cap\{x=0\}$, where $\Gamma^{-(+)} = \{\mathbf{x}\in\Gamma = \partial Q : \tilde{\beta}\cdot\mathbf{n}(\mathbf{x}) < 0 (>0), \quad \tilde{\beta} = (1,\beta)\}$, and Γ^0 is defined analogously, so that $\Gamma^0 = \{(x,y,\pm z_0)\}\cup\{(x,\pm y_0,0)\}$.

Suppose now that $J^h\in\mathcal{V}_h$, where $\mathcal{V}_h\subset L_2(Q)$ is a finite element space, is a Galerkin type approximate solution satisfying

$$\left\{\begin{array}{ll} J^h_x + \beta\cdot\nabla_\perp J^h - \hat{\varepsilon}\Delta_\perp J^h = \mathcal{R}, & \text{in} \quad Q,\\ J^h(0,\cdot) = f_h, & \text{in} \quad I_y\times I_z,\\ J^h = 0, \quad \textit{on } \Gamma_0^-, \quad \text{and} \quad \hat{J}^h_z = 0, & \textit{on } \Gamma^0, \end{array}\right. \tag{6.8}$$

where f_h is a Galerkin approximation of f and the residual $\mathcal{R}$ satisfies Galerkin orthogonality relation

$$\int_Q \mathcal{R}v\,dx\,dx_\perp = 0, \quad \forall v\in\mathcal{V}_h. \tag{6.9}$$

We shall also use the following *semi-consistency* assumption:

$$\int_{\Gamma_s^-} J^h|\mathbf{n}\cdot\tilde{\beta}|\,d\Gamma = \int_{\Gamma_s^-} J|\mathbf{n}\cdot\tilde{\beta}|\,d\Gamma, \tag{6.10}$$

where $\Gamma_s^- := \Gamma^-\setminus\{x=0\}$, is the *side-inflow boundary*. Observe that both in our continuous and discrete model problems (6.7) and (6.8), primarily, we may assume

$$J|_{\Gamma_s^-} = J^h|_{\Gamma_s^-} = 0, \tag{6.11}$$

however, there is no guarantee that "after-collision" particles would obey the same boundary condition as (6.11). Therefore, assumption (6.10) is to ensure that: in the approximation procedure the total inflow of particles is preserved.

In the sequel and to avoid multiple-indices, we shall refer to all approximated functions with alternate sub or super-index h. Subtracting (6.8) form (6.7) gives the following equation for the error $\hat{e}_h = \hat{J} - J^h$:

$$\left\{\begin{array}{ll} \mathcal{L}\hat{e}^h \equiv \hat{e}^h_x - \beta\cdot\nabla_\perp\hat{e}^h - \hat{\varepsilon}\Delta_\perp\hat{e}^h = -\mathcal{R}, & \text{in} \quad Q,\\ \hat{e}^h(0,\cdot) = f - f_h, & \text{in} \quad I_y\times I_z,\\ \hat{e}^h = 0, \quad \text{on } \Gamma_0^-, \quad \textit{and} \quad \hat{e}^h_z = 0, & \text{on } \Gamma^0. \end{array}\right. \tag{6.12}$$

We now introduce a dual for the non-degenerate problems (6.7), (6.8) or (6.12) as

$$\left\{\begin{array}{ll} \mathcal{L}^*\psi = -\psi_x - \beta\cdot\nabla_\perp\psi - \hat{\varepsilon}\Delta_\perp\psi = \hat{e}^h, & \text{in} \quad Q,\\ \psi = 0, \quad \text{on} \quad \Gamma^+, \quad \text{and} \quad \psi_z = 0, & \text{on} \quad \Gamma^0. \end{array}\right. \tag{6.13}$$

Let us, for simplicity, start to consider the original Fermi case by replacing, in (6.7)-(6.13), $\beta\cdot\nabla_\perp$, $\Delta_\perp$, and ψ by $z\partial_y$, ∂_{zz}, and φ, respectively. Then we have the following version of the dual problem (6.13):

$$\left\{\begin{array}{ll} \mathcal{L}^*\varphi = -\varphi_x - z\varphi_y - \hat{\varepsilon}\varphi_{zz} = \hat{e}^h, & \text{in} \quad Q,\\ \varphi = 0, \quad \text{on} \quad \Gamma^+, \quad \text{and} \quad \varphi_z = 0, & \text{on} \quad \Gamma^0. \end{array}\right. \tag{6.14}$$

Recall that, in (6.14), $\hat{\varepsilon}$ is obtained from (6.1) by replacing ε by σ. We shall use the notation $\hat{\varepsilon}$ for both degenerate and non-degenerate cases, the meaning would

be obvious from the context. Further, for simplicity, we need a *weighted angular symmetry* viz

$$\int_{I_x\times I_y}(\varphi w)(z_0)\,dxdy=\int_{I_x\times I_y}(\varphi w)(-z_0)\,dxdy,\qquad \forall w\in L_2(Q). \tag{6.15}$$

Integrating by parts and using (6.15) with $w=(\hat{\varepsilon}\hat{e}^h)_z$ and $w=\hat{\varepsilon}_z\hat{e}^h$, we have

$$-(\hat{e}^h,\hat{\varepsilon}\varphi_{zz})_Q=-(\hat{\varepsilon}_z\hat{e}^h_z,\varphi)_Q+(\hat{\varepsilon}_z\hat{e}^h,\varphi_z)_Q-(\hat{\varepsilon}\hat{e}^h_{zz},\varphi)_Q, \tag{6.16}$$

where we have used the boundary condition $\varphi_z=\hat{e}^h_z=0$, on Γ^0. Using (6.14)-(6.16), we get the following error representation formula:

$$\begin{aligned}\|\hat{e}^h\|^2&=(\hat{e}^h,\mathcal{L}^*\varphi)_Q=\int_Q\hat{e}^h(-\varphi_x-z\varphi_y-\varepsilon\varphi_{zz})\,dx\,dx_\perp\\&=(\mathcal{L}\hat{e}^h,\varphi)_Q-\int_{I_\perp}\hat{e}^h\varphi\Big|_{x=0}^{x=L}dy\,dz-\int_{I_x\times I_z}z\hat{e}^h\varphi\Big|_{y=-y_0}^{y=y_0}dx\,dz\\&\quad-(\hat{\varepsilon}_z\hat{e}^h_z,\varphi)_Q+(\hat{\varepsilon}_z\hat{e}^h,\varphi_z)_Q:=\sum_{i=1}^5 I_i.\end{aligned} \tag{6.17}$$

Below we identify the terms $I_i,\ i=1,\dots 5$, more closely. We have that

$$I_1=(\mathcal{L}\hat{e}^h,\varphi)=-\int_Q\mathcal{R}\varphi\,dx\,dx_\perp.$$

The incidental boundary conditions give

$$I_2=-\int_{I_\perp}\hat{e}^h(L,\cdot)\varphi(L,\cdot)\,dx_\perp+\int_{I_\perp}\hat{e}^h(0,\cdot)\varphi(0,\cdot)\,dx_\perp=\int_{\Gamma^0}(f-f_h)\varphi\,dx_\perp,$$

while the outflow boundary conditions, i.e., $\varphi=0$, on Γ^+ imply that

$$\begin{aligned}I_3&=-\int_{I_x}\left\{\int_0^{z_0}z\hat{e}^h\varphi\Big|_{y=-y_0}^{y=y_0}dz+\int_{-z_0}^0 z\hat{e}^h\varphi\Big|_{y=-y_0}^{y=y_0}dz\right\}dx\\&=\int_{I_x}\int_0^{z_0}z\hat{e}^h(-y_0)\varphi(-y_0)\,dzdx-\int_{I_x}\int_{-z_0}^0 z\hat{e}^h(y_0)\varphi(y_0)\,dzdx\\&=\int_{\Gamma_s^-}\hat{e}^h\varphi|\mathbf{n}\cdot\tilde{\beta}|\,d\Gamma,\end{aligned}$$

where, $\mathbf{n}$ is the outward unit normal defined at the boundary and, for the sake of generality, we have not used the assumption (6.10), yet. Thus

$$I_2+I_3=\int_{\Gamma^-}\hat{e}^h\varphi|\mathbf{n}\cdot\tilde{\beta}|\,d\Gamma. \tag{6.18}$$

Hence, summing up we have

$$\|\hat{e}^h\|^2=-\int_Q\mathcal{R}\varphi\,d\mathbf{x}+\int_{\Gamma^-}\hat{e}^h\varphi|\mathbf{n}\cdot\tilde{\beta}|\,d\Gamma-\int_Q\hat{\varepsilon}_z\hat{e}^h_z\varphi\,d\mathbf{x}+\int_Q\hat{\varepsilon}_z\hat{e}^h\varphi_z\,d\mathbf{x}. \tag{6.19}$$

We use Galerkin orthogonality (6.9) and write

$$\int_Q\mathcal{R}\varphi\,dx\,dx_\perp=\int_Q\mathcal{R}(\varphi-\mathcal{P}_h\varphi)\,dx\,dx_\perp=\int_Q(\mathcal{R}-\mathcal{P}_h\mathcal{R})(\varphi-\mathcal{P}_h\varphi)\,dx\,dx_\perp,$$

where $\mathcal{P}_h : L_2(Q) \to \mathcal{V}_h$ is the $L_2(Q)$-projection. By Cauchy-Schwartz inequality we may estimate the boundary integral term in (6.18) as

$$\int_{\Gamma_-} \hat{e}^h \varphi |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma \le \left(\int_{\Gamma_-} |\hat{e}^h|^2 |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma\right)^{1/2} \times \left(\int_{\Gamma_-} \varphi^2 |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma\right)^{1/2}.$$

Now using an interpolation error, with a symmetry assumption of the form $\varphi_{yy} = \varphi_{zz}$, we get the estimate

$$\|\hat{\varepsilon} h^{-2}(\varphi - \mathcal{P}_h\varphi)\|_Q \le C^i \|\hat{\varepsilon}\Delta_\perp \varphi\|_Q \approx C^i \|\hat{\varepsilon}\varphi_{zz}\|_Q, \tag{6.20}$$

together with a strong stability estimate for the dual problem (6.14) of the form

$$\|\hat{\varepsilon}\varphi_{zz}\|_Q \le C^s \|\hat{e}^h\|_Q, \tag{6.21}$$

we get that

$$-\int_Q \mathcal{R}\varphi\, dx dx_\perp \le C^s C^i \|h^2\hat{\varepsilon}^{-1}(\mathcal{R} - \mathcal{P}_h\mathcal{R})\|_Q \|\hat{e}^h\|_Q. \tag{6.22}$$

To estimate the boundary integrals we recall the L_2 trace theorem, see [7],

$$\|u\|^2_{L_2(\partial\Omega)} \le C\|u\|^2_{L_2(\Omega)} \|u\|^2_{W_2^1(\Omega)}, \tag{6.23}$$

and also the inverse estimate

$$\|v\|^2_{W_2^1(\Omega)} \le C\|h^{-1}v\|^2_{L_2(\Omega)}, \tag{6.24}$$

where W_p^r is the usual Sobolev space consisting of functions having their derivatives up to order r in L_p, u and v are sufficiently smooth functions and Ω has a Lipschitz boundary, see [1] for the details. So that applying (6.23)-(6.24) to φ and Q we get using (6.20)

$$\begin{aligned}\int_{\Gamma_-} |\varphi|^2 |\mathbf{n}\cdot\beta|\, d\Gamma &\le C\|\varphi\|_Q \|\varphi\|_{W_2^1(Q)} \le C\|\varphi - \mathcal{P}_h\varphi\|_Q \|\varphi - \mathcal{P}_h\varphi\|_{W_2^1(Q)} \\ &\le C\|\hat{\varepsilon}h^{-2}(\varphi - \mathcal{P}_h\varphi)\|_Q \|\hat{\varepsilon}^{-1}h^2(\varphi - \mathcal{P}_h\varphi)\|_{W_2^1(Q)} \\ &\le CC^s(C^i)^2 \|\hat{e}^h\|_Q \|\hat{\varepsilon}^{-1}h^3\Delta_\perp\varphi\|_Q,\end{aligned}$$

where C depends on the trace theorem and inverse inequality constants. By (6.1) we have that $\hat{\varepsilon} > h^{3/2}$ and therefore $\hat{\varepsilon}^{-1}h^3 \le h^{3/2} \le \hat{\varepsilon}$. Thus

$$\|\hat{\varepsilon}^{-1}h^3\varphi_{zz}\|_Q \le \|\hat{\varepsilon}^{-1}h^3\Delta_\perp\varphi\|_Q \approx \|\hat{\varepsilon}\varphi_{zz}\|_Q \le C^s\|\hat{e}^h\|_Q.$$

Hence

$$\int_{\Gamma_-} |\varphi|^2 |\mathbf{n}\cdot\beta|\, d\Gamma \le C_T(C^sC^i)^2 \|\hat{e}^h\|^2_Q. \tag{6.25}$$

At this moment we need to invoke (6.10), (note that if there is a feasible information on behaviour of the secondary particles at the inflow boundary we would be able to continue without using (6.10)), identifying the boundary integral

$$\int_{\Gamma_-} |\hat{e}^h|^2 |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma = \int_{\Gamma^0} |f - f_h|^2 |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma. \tag{6.26}$$

It remains to estimate I_4 and I_5, where there is no orthogonality relation, such as (6.9), available. Now we assume, for a sufficiently small constant $c \ll 1$, that

$$|\nabla_\perp \hat{\varepsilon}| \le c\hat{\varepsilon}h_\perp^{-1}, \tag{6.27}$$

and let $\tilde{C} = \sup(\hat{\varepsilon}h_\perp^{-1})/\inf(\hat{\varepsilon}h_\perp^{-1})$, (works for the case corresponding to $\hat{\varepsilon} \approx \mathcal{O}(h)$ in (6.2), as well), then by (6.27) and the inverse estimate we have

$$|(\hat{\varepsilon}_z \hat{e}_z^h, \varphi)_Q| \le c \sup_{\mathbf{x}\in Q} \left(\hat{\varepsilon}h_\perp^{-1}\right) \|h_\perp^{-1}\hat{e}^h\| \|h_\perp^2 \Delta_\perp \varphi\| \le c\tilde{C}\|\hat{e}^h\| \|\hat{\varepsilon}\varphi_{zz}\| \le c\tilde{C}C^s \|\hat{e}^h\|^2,$$

the choice of $\tilde{C}$ is for moving $\hat{\varepsilon}h_\perp^{-1}$ in and outside the norms, and c is chosen so that $c\tilde{C}C^s < 1/8$. Estimating I_5, in a similar way we finally get

$$|(\hat{\varepsilon}_z \hat{e}_z^h, \varphi)_Q| + |(\hat{\varepsilon}_z \hat{e}^h, \varphi_z)_Q| \le \frac{1}{4}\|\hat{e}^h\|^2. \tag{6.28}$$

Inserting (6.22)-(6.28) in (6.19) and using a kick-back argument we end up with

$$|||\hat{e}^h|||_Q \le \bar{C}\left[\|h^2\hat{\varepsilon}^{-1}(\mathcal{R} - \mathcal{P}_h\mathcal{R})\|_Q + \left(\int_{\Gamma^0} |f - f_h|^2 |\mathbf{n}\cdot\tilde{\beta}|\, d\Gamma\right)^{1/2}\right]. \tag{6.29}$$

where $\bar{C} = \bar{C}(c, \tilde{C}, C_T, C^s, C^i)$. Thus we have estimated the error in terms of the residual and the incident boundary error and we have a complete control over all the involved constants (note that C_T being a theoretical constant is not affected by our approximation procedure). The estimate (6.29), which is an analogue of (6.3), is appropriate in the present contest with $\mathcal{R}$ satisfying the Galerkin orthogonality relation (6.9) and f being a sufficiently smooth approximation for the product of incident Dirac δ functions at the boundary.

We have now outlined the basic ideas in the proof of the a posteriori error estimate (6.3) which relay on the Galerkin orthogonality relation (6.9) and the strong stability estimate (6.21) of the dual problems (6.13) and (6.14).

Remark 6.1. The strong stability estimate (6.21) should be compared with the non-validity of a *weak stability* estimate for (6.13) and (6.14) of the form

$$\|\rho\|_Q \le C\|\hat{e}^h\|_Q, \quad \text{with } \rho = \psi, \text{ or } \varphi, \tag{6.30}$$

corresponding to the L_2-instability phenomenon, related to the lack of absorption. However, since ψ and $\varphi = 0$ on a part of the boundary (Γ^+), with positive measure, we may derive a weak variant of (6.30) (with ρ replaced by $\hat{\varepsilon}\rho$) using Poincare inequality, (see [3], Lemma 2.2). We note that in (6.21) the derivative $\Delta_\perp\varphi \sim \varphi_{zz}$ of the dual solution is L_2-controlled (with the factor $\hat{\varepsilon}$) in terms of $\|\hat{e}^h\|_Q$, whereas L_2-control of φ itself as in the estimate (6.30) is not possible to achieve in general. For the a posteriori error control, using the strong stability estimates of the type (6.21) (with derivative control only), it is necessary to use Galerkin orthogonalities.

To motivate for removing degeneracy through introducing ε and also the role played by the artificial viscosity $\hat{\varepsilon}$ in the error estimate (6.3) we notice that the corresponding sharp a posteriori error estimate for elliptic problems is

$$\|\hat{e}^h\|_Q \le C\|h^2\mathcal{R}(J^h)\|_Q. \tag{6.31}$$

The estimates (6.3) and (6.29) may be viewed as a variant of (6.31) where the ellipticity introduced, by $\hat{\varepsilon}$, in the hyperbolic problem is compensated by the multiplicative factor $\hat{\varepsilon}^{-1}$ in (6.3) and (6.29).

In conclusion: A posteriori error estimates for numerical schemes may be viewed as special cases of a general stability theory controlling the effect on the solutions resulting from non-vanishing residuals. The perturbations in the finite element method corresponding to certain orthogonality relations make the a posteriori error estimates possible in cases where a general perturbation argument would fail.

7. Numerical Examples

To justify the advantageous behaviour of the characteristic streamline diffusion (CSD) versus characteristic Galerkin (CG) we present, through some simple numerical examples, the smoothing effect of a modified SD method compared to the standard Galerkin (SG): (i) We consider a *semi-streamline diffusion* (SSD) approach by interpreting x as being a time variable and perform different time discretizations of I_x by backward Euler (BE), Crank-Nicolson (CN) and discontinuous Galerkin (DG). (ii) We combine these time discretizations with both SG and SD methods for the discretization of $I_\perp$. The implementations are performed over two different initial conditions: Maxwellian and modified Dirac approximating our data: the Dirac δ function.

We split the problem into two steps. First we discretize the two dimensional domain $I_\perp = I_y \times I_z$ by means of piecewise linear approximation cG(1), and establish a mesh there in order to obtain a semidiscrete solution and subsequently we apply one of the three schemes, BE, CN or DG to step advance in the x direction. Our cG(1) basis functions have the form, $\phi_i = a_1 y + a_2 z + a_3$.

In some special cases (for instance, for $\varepsilon = \varepsilon(x) := \sigma(x)/2$, see [8], [9] and [10]) the closed form exact solution of (1.11) is given by,

$$J(x,y,z) = \frac{\sqrt{3}}{\pi\varepsilon x^2} e^{-2\left[3(y/x)^2 - 3(y/2)z + z^2\right]/(\varepsilon x)}. \tag{7.1}$$

This allows us to draw some limited comparisons in terms of the actual error. In addition to being a limited case, (7.1) also displays singularities near the origin which, (although removable), makes it difficult to numerically implement as is. Obviously the final solution depends on initial conditions and therefore it is not correct to compare (7.1) with the solutions we obtain numerically since the underlying initial conditions were not the same to start with. For instance we can not numerically provide an initial data of the form of a Dirac δ function. We therefore use two different types of *computable initial conditions*, each approximating the Dirac δ function, in the L_1 sense, for comparison purposes. Through these examples we also ascertain how strongly can differences in initial conditions affect our estimates on convergence established for CG and CSD, see [5] for further details on implementations.

More specifically in Figures 1 and 2 below we consider Dirac and Maxwellian data and look at slices of the domain Q and the differences between the *"exact"* and approximate solutions over all three cases of time discretization schemes. The computational parameters that are used relay on the theoretical results presented in previous sections. For instance ε must be chosen to be small and given such a choice we take $h^2 \leq \varepsilon \leq h$, also $\delta \sim h$. Specifically our Figures 1 and 2, were produced for values of $\varepsilon = .05$ and $h = .1$. In these examples the value of δ is taken as $\delta = h/2$, and the time increment was chosen as $k := \hbar = h^2$. For further studies in this direction we refer to [5]. In Figure 1, the superiority of the SSD is more obvious when the data is non-smooth. Even for a highly smooth "initial data", such as Maxwellian (viz Figure 2), the smoothing effects of SSD over SG are clearly pronounced, especially for the discontinuous Galerkin approximation of I_x.

We conclude that: For this type of problems using modified SD approaches, such as SSD and CSD, the oscillatory behaviour of the Galerkin schemes and L_2 projections are substantially improved.

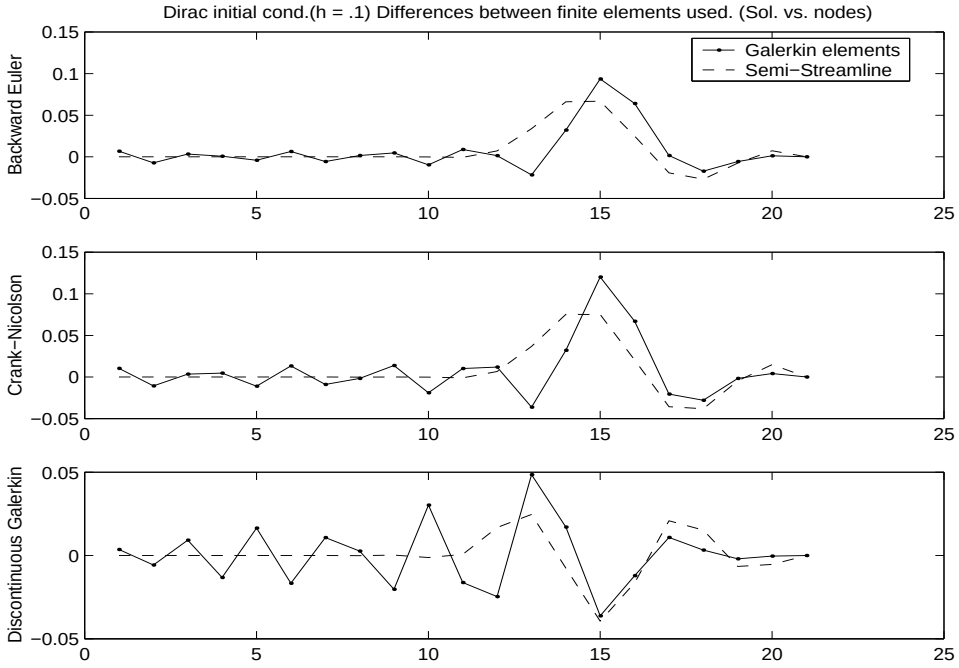

FIGURE 1. Galerkin vs. Semi-Streamline elements for Dirac initial condition at $h = .1$ for the slice, $-1 \leq y \leq 1$ at $z = -.9$.

REFERENCES

[1] Adams, R.A., *Sobolev spaces*, Academic Press, New York, 1975.

[2] Asadzadeh, M., *Streamline diffusion methods for the Vlasov-Poisson equation*, Math. Model. Numer. Anal., **24** (1990), no 2, 177-196.

[3] ———, *Streamline Diffusion Methods for Fermi and Fokker-Planck Equations*, TTSP, **26** (1997), 319-340.

[4] ———, *A posteriori error estimates for the Fokker-Planck and Fermi pencil beam equations*, Math. Models Meth. Appl. Sci. **10** (2000), no. 5, 737-769.

[5] Asadzadeh, M. and Sopasakis, A., *On Fully Discrete Echemes for the Fermi Pencil-Beam Equation*, Preprint NO **2000-48**, ISSN 0347-2809, Department of Mathematics, Chalmers University of Technology, Goteborg, Sweden.

[6] Börgers, C. and Larsen, E. W., *Asymptotic derivation of the Fermi pencil beam approximation*, Nucl. Sci. Eng. **123** (1996), 343-357.

[7] Brenner, C. S. and Scott, L. R., *The Mathematical Theory of Finite Element Methods*, Springer, New York, 1994.

[8] Eyges, L., *Multiple scattering with energy loss*, Phys. Rev. **74** (1948), 1534-.

[9] Fermi, E., quoted in Rossi B. and Greisen K., *Cosmic ray theory*, Rev. Mod. Phys. **13** (1941), 240-.

[10] Jette, D., *Electron dose calculation using multiple-scattering theory. A. Gaussian multiple-scattering theory*, Med. Phys. **15** (1988), 123-137.

[11] Johnson, C., *A new approach to algorithms for convection problems based on exact transport+projection*, Comput. Meyhods Appl. Mech. Engrg., **100** (1992), 45-62.

[12] Pomraning, G. C., *The Fokker-Planck operator as an asymptotic limit*, Math. Mod. Meth. Appl. Sci. **2** (1992), 21-36.

DEPARTMENT OF MATHEMATICS, CHALMERS UNIVERSITY OF TECHNOLOGY AND GÖTEBORG UNIVERSITY, SE–412 96, GÖTEBORG, SWEDEN

E-mail address: mohammad@math.chalmers.se

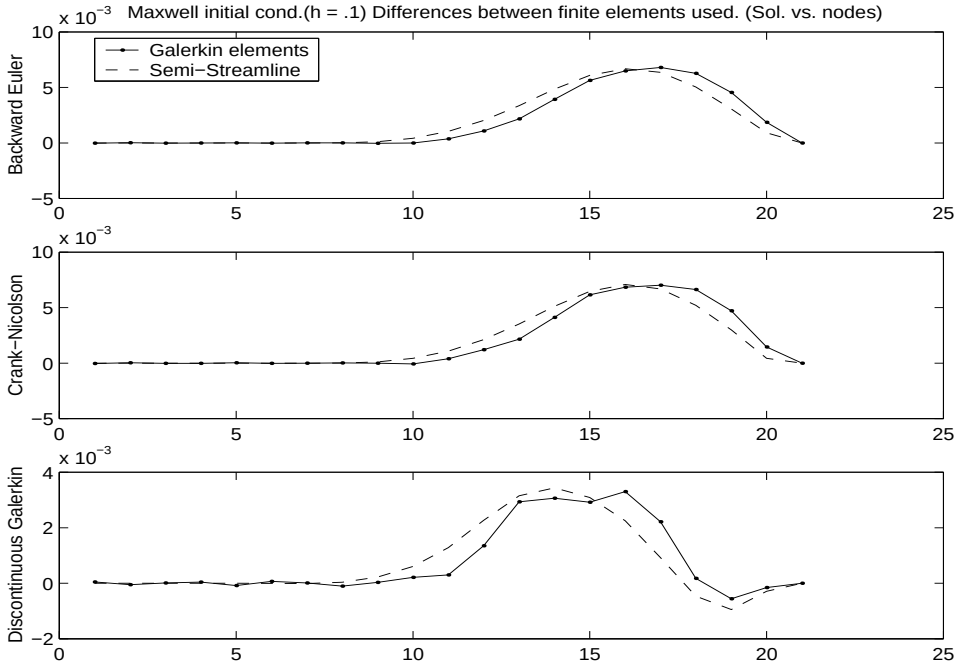

FIGURE 2. Galerkin vs. Semi-Streamline elements for Maxwell initial condition at $h = .1$ for the slice, $-1 \leq y \leq 1$ at $z = -.9$.