INTERNATIONAL JOURNAL OF
NUMERICAL ANALYSIS AND MODELING
Volume 10, Number 4, Pages 860–875

ON CONVERGENCE OF THE STREAMLINE DIFFUSION AND DISCONTINUOUS GALERKIN METHODS FOR THE MULTI-DIMENSIONAL FERMI PENCIL BEAM EQUATION

MOHAMMAD ASADZADEH AND EHSAN KAZEMI

Abstract. We derive error estimates in the L_2 norms, for the streamline diffusion (SD) and discontinuous Galerkin (DG) finite element methods for steady state, energy dependent, Fermi equation in three space dimensions. These estimates yield optimal convergence rates due to the maximal available regularity of the exact solution. Here our focus is on theoretical aspects of the h and hp approximations in both SD and DG settings.

Key words. Fermi equation, particle beam, streamline diffusion, discontinuous Galerkin, stability, convergence

1. Introduction

We study approximate solutions for the three-dimensional Fermi equation using streamline diffusion (SD) and discontinuous Galerkin (DG) finite element methods. We prove stability estimates and derive optimal convergence rates for the current function. This work extends the results in [2]-[3] to the multidimensional case, and includes the *hp* approach. The physical problem has diverse applications in, e.g. astrophysics, material science, electron microscopy, radiation therapy, etc. We shall consider a pencil beam of particles normally incident on a slab of finite thickness, entering the slab at a single point, e.g. $(0, 0, 0)$, in the direction of positive x-axis.

Fermi equation is a convection-diffusion equation, obtained as an asymptotic limit of the Fokker-Planck equation as the *transport cross-section* (σ_{tr}) gets smaller, see [7]. The equation is *degenerate* in both convection and diffusion in the sense that drift and diffusion are taking place in, physically, different domains, and the problem is *convection dominated*. Further, the associated boundary conditions are in the form of product of δ functions, which are not suitable for L_2-estimates. Therefore, we consider model problems with data smoother than Dirac δ-function.

Fermi equation has closed form solutions for σ_{tr} being a constant or a function of only x. In the present setting the direction of penetration of the beam, x, may also be interpreted as the direction of a *hypothetic* time variable.

The SD-method is obtained modifying the weak form by adding a multiple of the "drift-terms" in the equation to the test function. This yields artificial diffusion added only in the streamlines direction (motivating for the name: *the streamline diffusion method*) which improves stability in the characteristic direction so that internal layers are not smeared out while the added diffusion removes oscillations near boundary layers. The oscillations merge from the lack of stability of standard Galerkin for convection dominated problems, see, e.g. [14]. While SD may have discontinuities in x-direction only, the DG method allows jump discontinuities across interelement boundaries in order to count for the local effects. We study both h and hp versions of SD and DG methods. A semi-streamline diffusion for Fermi

Received by the editors June 5, 2012 and, in revised form, January 23, 2013.
1991 *Mathematics Subject Classification.* 65M15, 65M60.

equation has been implemented in [3]. The *hp* version is considered in a general setting for a Vlasov-Poisson-Fokker-Planck system in [5].

An outline of this paper is as follows: In Section 2, we introduce the model problem. Section 3 is devoted to the stability estimates and convergence analysis for the h and hp streamline diffusion approximations of the Fermi equation. Section 4 is the discontinuous Galerkin counterpart of Section 3, counting for local properties.

2. Model Problem

We consider a model problem for three dimensional Fermi equation on a bounded polygonal domains $\Omega_{\mathbf{x}} \subset \mathbb{R}^3$, $\mathbf{x} = (x, y, z) =: (x, x_\perp)$, with velocities $v \in \Omega_v \subset \mathbb{R}^2$:

$$(2.1)\quad \begin{cases} \frac{\partial f}{\partial x} + v \cdot \nabla_\perp f = \frac{\sigma_{tr}}{2}(\Delta_v f), & \text{in } (0, L] \times \Omega =: Q_L, \\ f(0, x_\perp, v) = f_0(x_\perp, v), & \text{in } \Omega = \Omega_{x_\perp} \times \Omega_v, \\ f(x, x_\perp, v) = 0, & \text{in } (0, L] \times ([\Gamma_v^- \times \Omega_v] \cup [\Omega_{x_\perp} \times \partial\Omega_v]), \end{cases}$$

where $f_0 \in L_2(\Omega)$, and for each $v \in \Omega_v$, the outflow boundary is given by

$$(2.2)\qquad \Gamma_v^- = \{x_\perp \in \partial\Omega_{x_\perp} : \mathbf{n}(x_\perp).v < 0\}.$$

Here $\Omega_\perp = \{(y, z)\}$, $\mathbf{n}(x_\perp)$ is the outward unit normal to $\partial\Omega_{x_\perp}$ at the point $x_\perp = (y, z) \in \partial\Omega_{x_\perp}$, $v = (v_1, v_2)$, $\nabla_\perp = (\frac{\partial}{\partial y}, \frac{\partial}{\partial z})$ and $\sigma_{tr} = \sigma_{tr}(x, y, z)$.

2.1. Notations and preliminaries. Let $T_h^{x_\perp} = \{\tau_{x_\perp}\}$ and $T_h^v = \{\tau_v\}$ be finite element subdivisions of $\Omega_{x_\perp}$ and Ω_v, into the elements $\tau^{x_\perp}$ and τ^v, respectively. Thus, $T_h = T_h^{x_\perp} \times T_h^v$ will be a subdivision of $\Omega = \Omega_{x_\perp} \times \Omega_v$ with elements $\{\tau_{x_\perp} \times \tau_v\} = \{\tau\}$. Consider a partition $\mathcal{T}_h : 0 = x_0 < x_1 < \ldots < x_M = L$ of the interval $I = (0, L]$ into subintervals $I_m = (x_{m-1}, x_m]$, $m = 1, ..., M$, and let $\mathcal{C}_h$ be the corresponding subdivision of $Q_L := (0, L] \times \Omega$ into elements $K = I_m \times \tau$ with the mesh size $h_K = \operatorname{diam} K$. We assume that each $K \in \mathcal{C}_h$ is the image under a family of bijective affine maps $\{F_K\}$ of a fixed standard element $\hat{K}$ into K, where $\hat{K}$ is either the open unit simplex or the open unit hypercube in $\mathbb{R}^5$ (in the *hp*-analysis, $\hat{K}$ is the open unit hypercube in $\mathbb{R}^5$). Let $P_p(K)$ be the set of all polynomials of degree $\le p$ on K; in $x, x_\perp$ and v, and define the finite element space

$$(2.3)\qquad V_h = \{g \in \widetilde{\mathcal{H}}_0 : g \circ F_K \in P_p(\hat{K});\ \forall K \in \mathcal{C}_h\}, \qquad \text{where}$$

$$(2.4)\qquad \widetilde{\mathcal{H}}_0 = \prod_{m=1}^{M} H_0^1(S_m), \qquad S_k = I_k \times \Omega, \qquad k = 1, \cdots, M, \qquad \text{with}$$

$$(2.5)\qquad H_0^1(S_m) = \{g \in H^1(S_m) : g \equiv 0 \quad \text{on } \partial\Omega_v\}.$$

For piecewise polynomials w_i defined on the triangulation $\mathcal{C}_h' = \{K\}$ with $\mathcal{C}_h' \subset \mathcal{C}_h$ and for D_i being some differential operators, we use the notation,

$$(2.6)\qquad (D_1 w_1, D_2 w_2)_{Q'} = \sum_{K \in \mathcal{C}_h'} (D_1 w_1, D_2 w_2)_K, \qquad Q' = \bigcup_{K \in \mathcal{C}_h'} K,$$

where $(.,.)_Q$ is the $L_2(Q)$ scalar product and $\|.\|_Q$ is the corresponding $L_2(Q)$-norm. Further, for $m = 1, 2, \ldots, M$, $\beta = (v, \mathbf{0})$, $\mathbf{n} = (\mathbf{n}_{x_\perp}, \mathbf{n}_v)$ and with $\Gamma = \partial(\Omega_{x_\perp} \times \Omega_v)$,

$$(2.7)\qquad \begin{array}{ll} (f, g)_m = (f, g)_{S_m}, & \|g\|_m^2 = (g, g)_m, \\ \langle f, g\rangle_m = (f(x_m, ., .), g(x_m, ., .))_\Omega, & |g|_m^2 = \langle g, g\rangle_m, \\ \langle f, g\rangle_{\Gamma^-} = \int_{\Gamma^-} fg(\beta \cdot \mathbf{n}) ds, & \langle f, g\rangle_{\Gamma_m^-} = \int_{I_m} \langle f, g\rangle_{\Gamma^-} ds, \\ \langle f, g\rangle_{\Gamma_I^-} = \int_I \langle f, g\rangle_{\Gamma^-} ds, & \Gamma^- = \{(x_\perp, v) \in \Gamma : \beta \cdot \mathbf{n} < 0\}, \end{array}$$

where $\mathbf{n}_{x_\perp}$ and $\mathbf{n}_v$ are outward unit normals to $\partial\Omega_{x_\perp}$ and $\partial\Omega_v$, respectively. Below C will denote a constant not necessarily the same at each occurrence and independent of the parameters in the problem, unless otherwise specifically specified.

3. Streamline diffusion method

3.1. Streamline diffusion method with discontinuity in x. In this section we study the h and hp-versions of SD-method for the three dimensional Fermi equation (2.1) with $\sigma = \frac{1}{2}\sigma_{tr}(x,y,z)$. We use continuous trial functions in $x_\perp$ and v with possible jump discontinuities in x on the nodes of a partition $\mathcal{T}_h$ of $[0,L]$ with the jumps in x as

$$[g] = g_+ - g_-, \qquad \text{where} \tag{3.1}$$

$$\begin{array}{ll} g_\pm = \lim_{s\to 0\pm} g(x+s, x_\perp, v), & \text{for } (x_\perp, v) \in \mathrm{Int}(\Omega_{x_\perp}) \times \Omega_v,\ x \in I, \\ g_\pm = \lim_{s\to 0\pm} g(x+s, x_\perp + sv, v), & \text{for } (x_\perp, v) \in \partial\Omega_{x_\perp} \times \Omega_v,\ x \in I. \end{array} \tag{3.2}$$

Equation (2.1), associated with L_2 boundary conditions, gives rise to the variational formulation: find $f^h \in V_h$ such that for $m = 0, 1, \cdots, M-1$, and for all $g \in V_h$,

$$\begin{aligned} &\sum_{K\in I_m\times T_h} \big[(f^h_x + v\cdot\nabla_\perp f^h, g + \delta(g_x + v\cdot\nabla_\perp g))_K + \sigma(\nabla_v f^h, \nabla_v g)_K \\ &-\delta\sigma(\Delta_v f^h, g_x + v\cdot\nabla_\perp g)_K\big] + \langle f^h_+, g_+\rangle_m - \langle f^h_+, g_+\rangle_{\Gamma^-_m} = \langle f^h_-, g_+\rangle_m. \end{aligned} \tag{3.3}$$

In the h-version for (2.1), using test functions of the form $g + \delta(g_x + v\cdot\nabla_\perp g)$, with $\delta \sim h^\alpha$, $\alpha \geq 1$, would supply us with an extra diffusion term of order h^α in the streamline direction: $(1, v, \mathbf{0})$. Then, we will be able to control an extra term of the form $h\|g_x + v\cdot\nabla_\perp g\|$. In the hp-version, however, the choice of δ is more involved and depends on optimal choice of the parameters h and p *locally*. Therefore in hp-analysis, δ would appear as an elementwise (local) parameter δ_K.

3.1.1. The h-version of the SD-method. We formulate the SD-approximation of the Fermi equation (2.1), with jump discontinuities in x. Introducing the bilinear form

$$\tilde{B}(f,g) = B(f,g) + \sum_{m=1}^{M-1} \langle [f], g_+\rangle_m + \langle f_+, g_+\rangle_0 - \langle f_+, g_+\rangle_{\Gamma_I^-}, \tag{3.4}$$

$$\begin{aligned} B(f,g) = \sum_{K\in\mathcal{C}_h} &\big[(f_x + v\cdot\nabla_\perp f, g + \delta(g_x + v\cdot\nabla_\perp g))_{Q_L} + \sigma(\nabla_v f, \nabla_v g)_K \\ &- \delta\sigma(\Delta_v f, g_x + v\cdot\nabla_\perp g)_K\big] + \langle f, g\rangle_0 - \langle f, g\rangle_{\Gamma^-}, \end{aligned} \tag{3.5}$$

and the linear form, viz

$$\tilde{L}(g) = \langle f_0, g_+\rangle_0,$$

we may rewrite (3.3) in global form as

$$\tilde{B}(f^h, g) = \tilde{L}(g), \qquad \forall g \in V_h. \tag{3.6}$$

It is easy to see that the adequate triple norm in this case is:

$$[|||g|||]^2 = \tfrac{1}{2}\left[|||g|||^2 + \delta \parallel g_x + v.\nabla_\perp g \parallel^2_{Q_L} + \sum_{m=1}^{M-1} |[g]|^2_m\right] \qquad \text{with} \tag{3.7}$$

$$|||g|||^2 = \left[\sigma\|\nabla_v g\|^2_{Q_L} + |g|^2_M + |g|^2_0 + \textstyle\int_{I\times\partial\Omega} g^2 \mid \beta\cdot\mathbf{n}\mid dvds\right]. \tag{3.8}$$

We shall frequently use the following interpolation error estimates, see, e.g. [10] or [15]: Let $f \in H^{r+1}(\Omega)$ then there exists an interpolant $\tilde{f}^h \in V_h$ of f such that

$$\|f - \tilde{f}^h\|_{s,Q_L} \leq Ch^{r+1-s}\|f\|_{r+1,Q_L}, \quad s = 0, 1, \tag{3.9}$$

$$\|f - \tilde{f}^h\|_{\partial Q_L} \leq Ch^{r+1/2}\|f\|_{r+1,Q_L}. \tag{3.10}$$

Below we state the main results of the SD-approach (the proofs are as in [2]-[5]).

Lemma 3.1. *The bilinear form $\tilde{B}$ satisfies the coercivity estimate*

$$\tilde{B}(g, g) \geq [|||g|||]^2 \qquad \forall g \in V_h.$$

Theorem 3.1. *Let f and f^h satisfy* (2.1) *and* (3.6)*, respectively, then*

$$[|||f - f^h|||] \leq Ch^{k+1/2}\|f\|_{k+1,Q_L}. \tag{3.11}$$

3.1.2. The *hp*-version of the SD-method. In this part we derive error bounds which are simultaneously optimal, both in the mesh size h and the spectral order p in a stabilization parameter $\delta \sim \left(\frac{h^2}{\sigma p^4}\right)$. Below we extend the results of h-version (global) to *hp*-version for local case. To this end we consider the bilinear form

$$\hat{B}_\delta(f, g) = \sum_{K \in \mathcal{C}_h} [(f_x + v \cdot \nabla_\perp f, g + \delta(g_x + v \cdot \nabla_\perp g))_K + \sigma(\nabla_v f, \nabla_v g)_K$$
$$-\delta\sigma(\Delta_v f, g_x + v \cdot \nabla_\perp g)_K] + \sum_{m=1}^{M-1} \langle [f], g_+ \rangle_m + \langle f, g \rangle_0 - \langle f, g \rangle_{\Gamma^-}$$

and the linear functional

$$\hat{L}_\delta(g) = \langle f_0, g_+ \rangle_0,$$

where the non-negative piecewise constant function δ is defined by

$$\delta|_K = \delta_K \qquad \delta_K = \text{constant for } K \in \mathcal{C}_h.$$

The precise choice of δ will be discussed below. We now define the local version of (3.6): find $f^h \in V_h^p$, the space of all polynomials of degree $\leq p$, such that

$$\hat{B}_\delta(f^h, g) = \hat{L}(g) \qquad \forall g \in V_h^p, \tag{3.12}$$

Note that in the h version of the SD-approach we interpret $(.,.)_{Q_L}$ as $\sum_{m=1}^{M}(.,.)_m$ and, assuming discontinuities in x, we include jump terms it the x direction. Thus we estimate the sum of the norms over slabs S_m, as well as the contributions from the jumps over $x_m : m = 1, \ldots, M-1$. In the *hp*-version we have, in addition to slab-wise estimates, a further step of identifying $(.,.)_m$ by $\sum_{K \in I_m \times T_h}(.,.)_K$ counting for the local character of the parameter δ_K. We also define the norm $[|||.|||]_\delta$, obtained from (3.7), replacing δ (h) by δ_K and considering its local effects:

$$[|||g|||]_\delta^2 =: \frac{1}{2}\left[|||g|||^2 + \sum_{K \in \mathcal{C}_h} \delta_K \parallel g_x + v.\nabla_\perp g \parallel_K^2 + \sum_{m=1}^{M-1} |[g]|_m^2\right]. \tag{3.13}$$

Further, we assume that the family of partitions $\{\mathcal{C}_h\}_{h>0}$ is shape regular, in the sense that there is a positive constant C_0, independent of h, such that

$$C_0 h_K^5 \leq \rho(K), \qquad \forall K \in \bigcup_{h>0} \{\mathcal{C}_h\}, \tag{3.14}$$

where $\rho(K)$ is the diameter of the five dimensional sphere inscribed in K.

Lemma 3.2. *Assume that the local SD-parameter δ_K is selected in the range*

$$0 < \delta_K \leq \frac{h_K^2}{\sigma C_I^2 p^4}, \qquad \forall K \in \mathcal{C}_h, \tag{3.15}$$

where C_I is the constant from the standard inverse estimate (see [8]*, Lemma 4.5.3 and Theorem 4.5.11). Then the bilinear form $\hat{B}_\delta(.,.)$ is coercive on $V_h^p \times V_h^p$, i.e.*

$$\hat{B}_\delta(g,g) \geq \frac{1}{2}[|||g|||]_\delta^2, \qquad \forall g \in V_h^p. \tag{3.16}$$

Proof. The proof is a standard argument followed by the estimate of the $\delta_K\sigma$-term:

$$\begin{aligned}\delta_K\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_K &\leq \frac{1}{2}C_I h_K^{-1}p^2\sqrt{\sigma\delta_K}\left[\sigma\|\nabla_v g\|_K^2 + \delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2\right]\\ &\leq \frac{1}{2}\left[\sigma\|\nabla_v g\|_K^2 + \delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2\right],\end{aligned}$$

where we use Cauchy-Schwarz and inverse inequality and the assumption on δ_K. □

We shall use the following approximation property: Let $g \in H^s(K)$ and $\|.\|_{s,K}$ be the Sobolev norm on K; there exists a constant C depending on s and r but independent of g, h_K and p, and a polynomial $\Pi_p g$ of degree p such that (see [6]),

$$\|g - \Pi_p g\|_{r,K} \leq C\frac{h_K^{\mu-r}}{p^{s-r}}\|g\|_{s,K}, \quad \text{for} \quad 0 \leq r \leq s, \qquad \mu = \min(p+1, s). \tag{3.17}$$

We shall also require a global counterpart of (3.17) for the finite element space V_h^p,

Lemma 3.3. *Let $g \in H_0^1(Q_L) \cap L^2(I, H^r(\Omega))$, $r > 2$ such that $g\,|_K \in H^s(K)$, with a positive integer $s \geq r$ and $K \in \mathcal{C}_h$. Then, there exists an interpolant $\Pi_p g \in V_h^p$ of g which is continuous on Ω such that*

$$\|g - \Pi_p g\|_{1,K} \leq C\frac{h_K^{\mu-1}}{p^{s-1}}\|g\|_{s,K}, \tag{3.18}$$

where $C > 0$ is a constant independent of h and p and $\mu = \min(p+1, s)$.

See, e.g. [12] where a proof is outlined assuming certain regularity. More elaborated proofs can be found in [16] and [8]. We shall also need the trace inequality:

$$\|\eta\|_{\partial K}^2 \leq C(\|\nabla\eta\|_K\|\eta\|_K + h_K^{-1}\|\eta\|_K^2), \qquad \forall K \in \mathcal{C}_h. \tag{3.19}$$

Theorem 3.2. *Let $\mathcal{C}_h$ be a shape regular mesh on Q_L and f be the exact solution of* (2.1) *that satisfies the assumptions of Lemma 3.3. Let f^h be the solution of* (3.12) *and assume that $0 < \delta_K$ satisfies $0 < \delta_K \leq \frac{h_K^2}{\sigma C_I^2 p^4}$ for each $K \in \mathcal{C}_h$. Then,*

$$[|||f - f^h|||]_\delta^2 \leq C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\Big(\frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1} + \frac{h_K}{\delta_K p^2}\Big)\|f\|_{s,K}^2. \tag{3.20}$$

Proof. We start with the triangle inequality

$$[|||f - f^h|||]_\delta \leq [||\eta|||]_\delta + [||\xi|||]_\delta, \tag{3.21}$$

where $\eta = f - \Pi_p f$ and $\xi = f^h - \Pi_p f$. Here $\Pi_p f \in V_h^p$ is the conforming interpolant in Lemma 3.3. Using Lemma 3.2 and Galerkin orthogonality $\hat{B}_\delta(e,\xi) = 0$, we have

$$\begin{aligned}\frac{1}{2}[|||\xi|||]_\delta^2 &\le \hat{B}_\delta(\xi,\xi) = \hat{B}_\delta(\eta,\xi) - \hat{B}_\delta(e,\xi) = \hat{B}_\delta(\eta,\xi)\\ &= \sigma(\nabla_v\eta, \nabla_v\xi)_{Q_L} - \sigma\sum_{K\in\mathcal{C}_h}\delta_K(\Delta_v\eta, \xi_x + v\cdot\nabla_\perp\xi)_K\\ &\quad + (\eta_x + v\cdot\nabla_\perp\eta, \xi)_{Q_L} + \sum_{K\in\mathcal{C}_h}\delta_K(\eta_x + v\cdot\nabla_\perp\eta, \xi_x + v\cdot\nabla_\perp\xi)_K\\ &\quad + \sum_{m=1}^{M-1}\langle[\eta],\xi_+\rangle_m + \langle\eta_+,\xi_+\rangle_0 - \langle\eta,\xi_+\rangle_{\Gamma_I^-} = \sum_{i=1}^{7} T_i.\end{aligned} \tag{3.22}$$

The terms T_1 and T_3-T_7 are easily estimated by standard techniques (see [2]-[5]). As for the T_2 term, using the inverse inequality and assumptions on σ, and δ_K,

$$|T_2| \le C_I\delta_K\sigma p^2 h_K^{-1}\|\nabla_v\eta\|_K\|\xi_x + v\cdot\nabla_\perp\xi\|_K \le 2\sigma\|\eta\|_K^2 + \frac{\delta_K}{8}\|\xi_x + v\cdot\nabla_\perp\xi\|_K^2.$$

Then, we end up rewriteing the estimate (3.22) concisely (we skip the details) as

$$[|||\xi|||]_\delta \le C(I_1 + I_2), \tag{3.23}$$

where I_1 and I_2 are given by

$$\begin{aligned}I_1 &= \textstyle\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|_K^2 + \delta_K\|\eta_x + v\cdot\nabla_\perp\eta\|_K^2 + \sigma\|\nabla_v\eta\|^2\right),\\ I_2 &= \sum_{m=1}^{M-1}|\eta_-|_m^2 + \int_{I\times\partial\Omega}\eta^2|\beta\cdot\mathbf{n}|dvds.\end{aligned}$$

To estimate I_1 we have, using Lemma 3.3 and assumption on δ_K, that

$$I_1 \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-2}}{p^{2s-2}}(\delta_K^{-1}\frac{h_K^2}{p^2} + \delta_K + \sigma)\|f\|_{s,K}^2. \tag{3.24}$$

As, for the term I_2, using the trace estimate (3.19), yields

$$I_2 \le \sum_{K\in\mathcal{C}_h}(\frac{h_K^{\mu-1}}{p^{s-1}}\frac{h_K^\mu}{p^s} + h_K^{-1}\frac{h_K^{2\mu}}{p^{2s}})\|f\|_{s,K}^2 = \sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-1}}(1+\frac{1}{p})\|f\|_{s,K}^2. \tag{3.25}$$

Hence from (3.23)-(3.25) we get

$$[|||\xi|||]_\delta^2 \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}(\frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1} + \frac{h_K}{\delta_K p^2})\|f\|_{s,K}^2. \tag{3.26}$$

Finally, the term $[|||\eta|||]_\delta$ can be estimated in the same way, for which we get,

$$[|||\eta|||]_\delta^2 \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}(\frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1})\|f\|_{s,K}^2. \tag{3.27}$$

Substituting (3.26)-(3.27) into (3.21), we obtain the desired result. □

Remark 3.1. In Theorem 3.2, we chose δ_K for all $K \in \mathcal{C}_h$ when σ is small compared to h_k and $1/p$. The parameters are selected in a way that δ_K satisfies the hypothesis of Theorem 3.2. This particular choice of δ_K is motivated by our analysis in the discretization error (3.20) in $[|||.|||]_\delta$ norm, in order to give *hp*-error bound as,

$$[|||f - f^h|||]_\delta^2 \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-1}}\|f\|_{s,K}^2. \tag{3.28}$$

The assumption on σ is crucial for, simultaneous, optimal error bound in h and p.

Remark 3.2. The assumptions of Lemma 3.3, for the global regularity of the solution, are somehow restrictive, but since we assume our test functions are continuous in $(x_\perp, v)$, so in this framework it is difficult to relax these assumptions. For the DG counterpart of current analysis we shall substantially ease these requirements.

Remark 3.3. We have not allowed element-by-element local parameters p, or s for the exact solution f. Our analysis can be extended easily to this case replacing s by s_K and $\|f\|_s$ by $\|f\|_{s,K}$, $K \in C_h$. However, to replace p by p_K, (although straightforward for the DG studies below) is an uneasy procedure in the SD case. Going through this cumbersome procedure for SD, subsequently, in the local approximation (3.17), $\mu = \min(p+1, s)$ will be replaced by $\mu_K = \min(p_K + 1, s_K)$.

4. Discontinuous Galerkin

4.1. Description of discontinuous Galerkin (DG)-method. Here we assume trial functions as being polynomials of degree $k \geq 1$ on each element K which may be discontinuous across inter-element boundaries in all variables. We define

$$\partial K_\pm(\tilde{\beta}) = \{(x, x_\perp, v) \in \partial K : \tilde{\beta} \cdot \mathbf{n} = \mathbf{n}_x(x, x_\perp, v) + \mathbf{n}_{x_\perp}(x, x_\perp, v) \cdot v \gtrless 0\}, \;\; K \in \mathcal{C}_h,$$

where $\tilde{\beta} = (1, v, \mathbf{0})$ and $\mathbf{n} = (\mathbf{n}_x, \mathbf{n}_{x_\perp}, \mathbf{n}_v)$ is the outward unit normal to ∂K. To treat the diffusive part of (2.1), using discontinuous trial functions, we introduce an operator R as defined in, e.g. [4] and [9]. To this end, we first define the spaces

$$\begin{aligned} \tilde{V} &= \prod_{K \in \mathcal{C}_h} H^1(K), \\ V_h &= \{w \in L_2(Q_L) : w\,|_K \in P_k(K) : \; \forall K \in \mathcal{C}_h; \; w = 0 \text{ on } \partial\Omega_v\}, \\ \mathbf{W}_h &= \{\mathbf{w} \in [L_2(Q_L)]^2 : \mathbf{w}\,|_K \in [P_k(K)]^2; \; \forall K \in \mathcal{C}_h\}. \end{aligned} \tag{4.1}$$

Then, given $g \in \tilde{V}$ we define $R : \; \tilde{V} \to \mathbf{W}_h$ by the following weak formulation

$$(R(g), \mathbf{w}) = -\sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \sum_{e \in \mathcal{E}_v} \int_e [[g]] \mathbf{n}_v \cdot (\mathbf{w})^0 dv, \qquad \forall \mathbf{w} \in \mathbf{W}_h.$$

Here $\mathcal{E}_v$ denotes the set of all interior edges of the triangulation T_h^v of the domain Ω_v^h and $\mathbf{n}_v$ is the outward unit normal from element τ_i to τ_j, sharing the edge e with $i > j$, $\tau_i, \tau_j \in T_h^v$. Further, for an appropriately chosen function χ let

$$(\chi)^0 := \frac{\chi + \chi^{ext}}{2}, \qquad [[\chi]] := \chi - \chi^{ext}, \tag{4.2}$$

where χ^{ext} denotes the value of χ in the element τ_v^{ext} having $e \in \mathcal{E}_v$ as the common edge with τ_v. Hence, roughly speaking, $[[\chi]]$ corresponds to the jump and $(\chi)^0$ is the average value of χ in the velocity variable. Next for $e \in \mathcal{E}_v$ we define the operator r_e to be the restriction of R to the elements sharing the edge $e \in \mathcal{E}_v$, i.e.

$$(r_e(g), \mathbf{w})_{Q_L} = -\sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \int_e [[g]] \mathbf{n}_v \cdot (\mathbf{w})^0 dv, \qquad \forall \mathbf{w} \in \mathbf{W}_h.$$

One can easily verify that, for any element τ_v of the triangulation of Ω_v,

$$\sum_{e \subset \partial\tau_v \cap \mathcal{E}_v} r_e = R \qquad \text{on } \tau_v. \tag{4.3}$$

As a consequence of this we have the following estimate

$$\|R(g)\|_K^2 \leq \kappa \sum_{e \subset \partial\tau_v \cap \mathcal{E}_v} \|r_e(g)\|_K^2, \tag{4.4}$$

where τ_v corresponds to the element K and $\kappa > 0$ is a constant. Now, since the support of each r_e is the union of elements sharing the edge e, we evidently have

$$\sum_{e\in\mathcal{E}_v} \|r_e(g)\|^2_{Q_L} = \sum_{K\in\mathcal{C}_h} \sum_{e\subset\partial\tau_v\cap\mathcal{E}_v} \|r_e(g)\|^2_K. \tag{4.5}$$

Hence, the DG method for (2.1) is now formulated as: find $f^h \in V_h$ such that

$$B_{\delta,\theta}(f^h, g) = \langle f_0, g_+\rangle_0, \quad \forall g \in V_h, \qquad \text{where} \tag{4.6}$$

$$B_{\delta,\theta}(f, g) = A_\delta(f, g) + D_\theta(f, g). \tag{4.7}$$

The bilinear forms A_δ and D_θ correspond to the convective and diffusive parts viz:

$$\begin{aligned} A_\delta(f^h, g) =& \sum_{K\in\mathcal{C}_h} (f^h_x + v\cdot\nabla_\perp f^h, g + \delta_K(g_x + v\cdot\nabla_\perp g))_K + \langle f_+, g_+\rangle_0 \\ &+ \sum_{K\in\mathcal{C}_h} \int_{\partial K_-(\tilde\beta)'} [f] g_+ |\tilde\beta\cdot\mathbf{n}|, \quad \partial K_-(\tilde\beta)' = \partial K_-(\tilde\beta)\backslash\{0\}\times\Omega, \end{aligned} \tag{4.8}$$

$$\begin{aligned} D_\theta(f^h, g) =& \sigma(\nabla_v f^h, \nabla_v g)_{Q_L} + \sigma(\nabla_v f^h, R(g))_{Q_L} + \sigma(R(f^h), \nabla_v g)_{Q_L} \\ &+ \lambda\sigma \sum_{e\in\mathcal{E}_v} (r_e(f^h), r_e(g))_{Q_L} - \sum_{K\in\mathcal{C}_h} \theta_K \sigma(\Delta_v f^h, g_x + v\cdot\nabla_\perp g)_K. \end{aligned} \tag{4.9}$$

Here, $[f^h] = f^h_+ - f^h_-$ where $f^h_\pm$ is defined as in (3.2), $\delta_K > 0$ is a positive constant on element K, $0 \le \theta_K \le \delta_K$ and $\lambda > 0$ is a given constant. We also define the norms corresponding to (4.8) and (4.9) by

$$\begin{aligned} |||g|||^2_{A_\delta} = \frac{1}{2}\Bigg[&\sum_{K\in\mathcal{C}_h} \delta_K \|g_x + v.\nabla_\perp g\|^2_K + |g|^2_M + |g|^2_0 + \int_{I\times\partial\Omega_+} g^2 |v\cdot\mathbf{n}_{x_\perp}| \\ &+ \sum_{K\in\mathcal{C}_h} \int_{\partial K_-(\tilde\beta)'} [g]^2 |\tilde\beta\cdot\mathbf{n}| \Bigg], \end{aligned}$$

and

$$|||g|||^2_{D_\theta} = \frac{1}{2}\left[\sigma\|\nabla_v g\|^2_{Q_L} + 2\sigma \sum_{e\in\mathcal{E}_v} \|r_e(g)\|^2_{Q_L}\right].$$

Finally, we define

$$|||g|||^2_{\delta,\theta} = |||g|||^2_{A_\delta} + |||g|||^2_{D_\theta}. \tag{4.10}$$

Note that, in general $[g]$ is distinct from the jump $[[g]]$, defined by (4.2), in the sense that the latter depends on element numbering as well. Recall that since the characteristic $\tilde\beta = (1, v, \mathbf{0})$ is divergent free, $(\tilde\beta\cdot\mathbf{n})$ is continuous across the inter-element boundaries of $\mathcal{C}_h$ and thus $\partial K_\pm$ is well defined. If we chose $\delta_K := h$, and $\theta_K := h$ for all $K \in \mathcal{C}_h$, then the problem (4.6) can be formulated as

$$B_*(f^h, g) = \langle f_0, g_+\rangle_0, \qquad \forall g \in V_h, \tag{4.11}$$

$$B_*(f^h, g) = A(f^h, g) + D(f^h, g). \tag{4.12}$$

We shall suppress the indexes δ from A_δ and θ from D_θ, when we set $\delta_K := h$ and $\theta_K := h$ for all $K \in \mathcal{C}_h$. Then, the stability lemma for bilinear forms A_δ and D_θ is:

Lemma 4.1 (Extended coercivity Lemma)**.** *Suppose that δ_K satisfies* (3.15) *for all $K \in \mathcal{C}_h$ and $\lambda > \max(2, 2\kappa)$, then there is a constant $0 < \alpha < 1/2$ such that*

$$A_\delta(g, g) + D_\theta(g, g) \ge \alpha(|||g|||^2_{A_\delta} + |||g|||^2_{D_\theta}), \qquad \forall g \in V_h.$$

Proof. By the definition of A_δ in (4.8) we have that

$$\begin{aligned} A_\delta(g,g) =&(g_x + v\cdot\nabla_\perp g, g)_{Q_L} + \sum_{K\in\mathcal{C}_h}\delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2 + |g|_0^2 \\ &+ \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'}[g]g_+|\tilde\beta\cdot\mathbf{n}|. \end{aligned} \tag{4.13}$$

Further, using Green's formula we may write

$$\begin{aligned} (g_x+&v\cdot\nabla_\perp g, g)_{Q_L} = \frac{1}{2}\sum_{K\in\mathcal{C}_h}\int_{\partial K} g^2\tilde\beta\cdot\mathbf{n} \\ &= \frac{1}{2}\left[-\sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'} g_+^2|\tilde\beta\cdot\mathbf{n}| + \sum_{K\in\mathcal{C}_h}\int_{\partial K_+(\tilde\beta)'} g_-^2|\tilde\beta\cdot\mathbf{n}|\right]. \end{aligned} \tag{4.14}$$

Hence,

$$\begin{aligned} &(g_x + v\cdot\nabla_\perp g, g)_{Q_L} + \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'}[g]g_+|\tilde\beta\cdot\mathbf{n}| + |g|_0^2 \\ &= \frac{1}{2}\left[\sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'}[g]^2|\tilde\beta\cdot\mathbf{n}| + \int_{I\times\partial\Omega_+} g^2|v\cdot\mathbf{n}_{x_\perp}| + |g|_0^2 + |g|_M^2\right]. \end{aligned} \tag{4.15}$$

Similarly, by the definition of D_θ and using (4.7), we have also

$$\begin{aligned} D_\theta(g,g) =&\sigma\|\nabla_v g\|_{Q_L}^2 + 2\sigma(\nabla_v g, R(g))_{Q_L} + \lambda\sigma\sum_{K\in\mathcal{C}_h}\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|_K^2 \\ &- \sum_{K\in\mathcal{C}_h}\theta_K\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_K. \end{aligned} \tag{4.16}$$

Finally, the estimate (4.4), for some $0<\varepsilon<\frac{1}{2}$, yields

$$2\sigma(\nabla_v g, R(g))_{Q_L} \le \sigma\sum_{K\in\mathcal{C}_h}\left[\varepsilon\|\nabla_v g\|_K^2 + \frac{\kappa}{\varepsilon}\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|_K^2\right]. \tag{4.17}$$

Thus

$$\begin{aligned} &2\sigma(\nabla_v g, R(g))_{Q_L} + \lambda\sigma\sum_{K\in\mathcal{C}_h}\sum_{e\in\mathcal{E}_h\cap\partial\tau_v}\|r_e(g)\|_K^2 \\ &\ge \sigma\sum_{K\in\mathcal{C}_h}\left[-\varepsilon\|\nabla_v g\|_K^2 + (\lambda - \frac{\kappa}{\varepsilon})\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|_K^2\right]. \end{aligned} \tag{4.18}$$

Hence, by an inverse estimate, using $\theta_K\le\delta_K$ and assumptions on σ and δ_K,

$$\sum_{K\in\mathcal{C}_h}\sigma\theta_K(\Delta_v g, g_x + v\cdot\nabla_\perp g)_{Q_L} \le \frac{1}{2}\left(\sigma\|\nabla_v g\|_{Q_L}^2 + \sum_{K\in\mathcal{C}_h}\delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2\right).$$

Taking $\alpha = \min[\frac{1}{2}-\varepsilon, \lambda - \frac{\kappa}{\varepsilon}]$, ($>0$ for $\frac{\kappa}{\lambda}<\varepsilon<\frac{1}{2}$) we conclude the desired result. □

Corollary 4.1. *For B_* defined as in* (4.12) *we have the coercivity estimate*

$$B_*(g,g)\ge\alpha|||g|||_*^2, \qquad \forall g\in V_h, \tag{4.19}$$

where $|||g|||_*^2 =: |||g|||_A^2 + |||g|||_D^2$.

Suppose now that $f^h \in W^h$ and f are the solutions of (4.6) and (2.1), respectively, and let $\tilde{f}^h \in V_h$ be the interpolant of the exact solution f. Then, we write

$$e := f - f^h = (f - \tilde{f}^h) - (f^h - \tilde{f}^h) \equiv \eta - \xi. \tag{4.20}$$

Lemma 4.2. *There exists a constant C independent of the mesh size h such that for δ_K chosen as in* (3.15) *we have the following estimates*

$$\begin{aligned} A_\delta(\eta,\xi) \leq & \frac{1}{8}|||\xi|||^2_{A_\delta} + C\sum_{K\in\mathcal{C}_h}(\delta_K^{-1}\|\eta\|_K + \delta_K\|\nabla\eta\|_K) \\ & + \sum_{K\in\mathcal{C}_h}|[\eta]|_{\partial K_-(\tilde{\beta})'} + |\eta|_{\Gamma_+} + |\eta|_0 + |\eta|_M, \\ D_\theta(\eta,\xi) \leq & \frac{1}{8}|||\xi|||^2_{A_\delta} + \frac{1}{8}|||\xi|||^2_{D_\theta} + C\sigma\|\nabla_v\eta\|^2_{Q_L}. \end{aligned} \tag{4.21}$$

Proof. We proceed as in the SD approach in [5]. Here, we need to control some additional jump and boundary terms. We have, using the definition of A_δ, that

$$\begin{aligned} A_\delta(\eta,\xi) = & \sum_{K\in\mathcal{C}_h}(\eta_x + v\cdot\nabla_\perp\eta, \xi + \delta_K(\xi_x + v\cdot\nabla_\perp\xi))_K \\ & + \langle\eta_+,\xi_+\rangle_0 + \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde{\beta})'}[\eta]\xi_+|\tilde{\beta}\cdot\mathbf{n}|. \end{aligned} \tag{4.22}$$

Integrating by parts we end up with

$$\begin{aligned} & (\eta_x + v\cdot\nabla_\perp\eta,\xi)_{Q_L} + \langle\eta_+,\xi_+\rangle_0 + \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde{\beta})'}[\eta]\xi_+|\tilde{\beta}\cdot\mathbf{n}| \\ & = -(\eta, \xi_x + v\cdot\nabla_\perp\xi)_{Q_L} - \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde{\beta})'}\eta_-[\xi]|\tilde{\beta}\cdot\mathbf{n}| \\ & \quad + \langle\eta_-,\xi_-\rangle_M + \int_{I\times\partial\Omega_+}\eta_-\xi_-|\tilde{\beta}\cdot\mathbf{n}|. \end{aligned} \tag{4.23}$$

Inserting (4.23) in (4.22) and applying Cauchy-schwarz inequality we obtain

$$\begin{aligned} A_\delta(\eta,\xi) \leq & \frac{1}{8}|||\xi|||^2_{A_\delta} + C\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|^2_K + \delta_K\|\nabla\eta\|^2_K\right) \\ & + \sum_{K\in\mathcal{C}_h}|[\eta]|^2_{\partial K_-(\tilde{\beta})'} + |\eta|^2_{\Gamma_+} + |\eta|^2_0 + |\eta|^2_M. \end{aligned} \tag{4.24}$$

For D_θ we have by the definition,

$$\begin{aligned} D_\theta(\eta,\xi) = & \sigma(\nabla_v\eta,\nabla_v\xi)_{Q_L} + \sigma(\nabla_v\eta, R(\xi))_{Q_L} + \sigma(R(\eta),\nabla_v\xi)_{Q_L} \\ & + \lambda\sigma\sum_{e\in\mathcal{E}_v}(r_e(\eta), r_e(\xi))_{Q_L} - \sum_{K\in\mathcal{C}_h}\theta_K\sigma(\Delta_v\eta, \xi_x + v\cdot\nabla_\perp\xi)_K := \sum_{i=1}^5 T_i. \end{aligned}$$

Here, T_1 and T_5 can be estimated by standard techniques. Below we estimate the terms T_2, T_3 and T_4. Since η is continuous, the definition of operators R and r_e yield that $T_3 = T_4 = 0$. To estimate T_2 we use (4.4) and (4.5) to obtain

$$|T_2| \leq \sum_{K\in\mathcal{C}_h}\sigma\|\nabla_v\eta\|_K\|R(\xi)\|_K \leq \sum_{K\in\mathcal{C}_h}\left(C\sigma\|\nabla_v\eta\|^2_K + \frac{\sigma}{C_1}\|R(\xi)\|^2_K\right). \tag{4.25}$$

Hence, by Cauchy-Schwarz inequality and assumption on σ we finally get

$$D_\theta(\eta,\xi) \leq \frac{1}{8}|||\xi|||^2_{A_\delta} + \frac{1}{8}|||\xi|||^2_{D_\theta} + C\sigma\|\nabla_v\eta\|^2_{Q_L}, \tag{4.26}$$

and the proof is complete. □

In the sequel we shall use the following lemma (see, e.g. [4]),

Lemma 4.3. *Let $u \in L^2(I \times \Omega_{x_\perp}, H^1(\Omega_v))$ with $\Delta_v u \in L^2(Q_L)$, and let $w \in V_h$. Then*

$$\sum_{K\in\mathcal{C}_h} \int_{I_m\times\tau_{x_\perp}} \int_{\partial\tau_v} w\frac{\partial u}{\partial \mathbf{n}_v} = \sum_{I_m\times\tau_{x_\perp}} \int_{I_m\times\tau_{x_\perp}} \sum_{e\in\mathcal{E}_v} \int_e [[w]]\mathbf{n}_v.(\nabla_v u)^0. \tag{4.27}$$

Theorem 4.1 (Convergence Theorem)**.** *Suppose $f^h \in V^h$ and f are the solutions of* (4.11) *and* (2.1) *respectively, then there exists a constant C independent of the mesh size h such that we have the following error estimate*

$$|||f - f^h|||_* \le Ch^{k+1/2}\|f\|_{k+1,Q_L}. \tag{4.28}$$

Proof. Using Lemma 4.1 and (4.20), we have

$$\alpha|||\xi|||_*^2 \le B_*(\xi,\xi) = B_*(\eta - e, \xi) = B_*(\eta,\xi) - B_*(e,\xi). \tag{4.29}$$

We may split $B_*(e,\xi)$ as

$$B_*(e,\xi) = A(e,\xi) + D(e,\xi). \tag{4.30}$$

Recall that

$$D(e,\xi) = D(f,\xi) - D(f^h,\xi). \tag{4.31}$$

Hence, by the definition of D and since $R(f) = r_e(f) = 0$ we have that

$$\begin{aligned} D(f,\xi) &= \sum_{K\in\mathcal{C}_h} \int_{I_m\times\tau_{x_\perp}} \int_{\tau_v} \sigma\nabla_v f\nabla_v\xi - \sigma \sum_{I_m\times\tau_{x_\perp}} \int_{I_m\times\tau_{x_\perp}} \sum_{e\in\mathcal{E}_v} \int_e [[\xi]]\mathbf{n}_v.(\nabla_v f)^0 \\ &= \sum_{K\in\mathcal{C}_h} \int_K -\sigma(\Delta_v f)\xi + \sigma \sum_{K\in\mathcal{C}_h} \int_{I_m\times\tau_{x_\perp}} \int_{\partial\tau_v} \xi\frac{\partial f}{\partial\mathbf{n}_v} \\ &\quad - \sigma \sum_{I_m\times\tau_{x_\perp}} \int_{I_m\times\tau_{x_\perp}} \sum_{e\in\mathcal{E}_v} \int_e [[\xi]]\mathbf{n}_v.(\nabla_v f)^0 = \sum_{K\in\mathcal{C}_h} \int_K -\sigma(\Delta_v f)\xi, \end{aligned}$$

where in the last equality we have used Lemma 4.3. Thus, the problem (4.11) is fully consistent and $B_*(e,\xi) = 0$. Further, we get from (4.29) that

$$\alpha|||\xi|||_*^2 \le B_*(\eta,\xi) = A(\eta,\xi) + D(\eta,\xi). \tag{4.32}$$

We have now using Lemma 4.2, the multiplicative trace inequality (3.19), and the local interpolation error estimates (3.9)-(3.10) that

$$A(\eta,\xi) \le \frac{1}{8}|||\xi|||_A^2 + Ch^{2k+1}\|f\|_{k+1,Q_L}^2, \tag{4.33}$$

and

$$D(\eta,\xi) \le \frac{1}{8}|||\xi|||_*^2 + Ch^{2k+1}\|f\|_{k+1,Q_L}^2. \tag{4.34}$$

Inserting (4.33) and (4.34) into (4.32) we obtain

$$|||\xi|||_*^2 \le Ch^{2k+1}\|f\|_{k+1,Q_L}^2. \tag{4.35}$$

Using the interpolation estimates as above we also have

$$|||\eta|||_*^2 \le Ch^{2k+1}\|f\|_{k+1,Q_L}^2. \tag{4.36}$$

Then (4.28) is a consequence of (4.35), (4.36) and the triangle inequality. □

Remark 4.1. Choosing $0 \leq \theta_K < \delta_K$ in (4.9), specially $\theta_K = 0$, i.e., with extra diffusion in $\mathbf{x}$ only for the convective terms, the Galerkin orthogonality would not hold any longer and this renders the scheme (4.6) as an inconsistent one. The consistency error introduces an additional term of $\mathcal{O}(h^3)$ in the convergence analysis of the scheme, since

$$B_{\delta,\theta}(f - f^h, \xi) = \sum_{K \in \mathcal{C}_h} (\delta_K - \theta_K)\sigma(\Delta_v f, \xi_x + v \cdot \nabla_\perp \xi)_K = T_6, \tag{4.37}$$

and the consistency error bound follows from

$$|T_6| \leq \; C\delta_K \sigma^2 \|\Delta_v f\|^2_{Q_L} + \frac{1}{8}|||\xi|||^2_{\delta,\theta}. \tag{4.38}$$

Hence, the scheme cannot be better than third order accurate, no matter how high the spectral degree k is, and the stabilizing term is therefore non-compatible with the optimal order guaranteed by the polynomial approximation.

4.2. ***hp*-Discontinuous Galerkin method.** In this section we employ the approach in [17] and derive error bound that is optimal in both h and p. We assume that the family $\{\mathcal{C}_h\}$ is shape regular in the sense of (3.14) and that every $K \in \mathcal{C}_h$ is affine equivalent to the unit hypercube in $\mathbb{R}^5$. We allow the meshes to be 1-irregular, i.e. elements may contain hanging nodes. Let us first consider the bilinear form

$$\tilde{D}_\delta = D_\delta(f, g) + D_s(f, g), \tag{4.39}$$

where D_δ is as in (4.9) and the stabilizer D_s is defined by

$$D_s(f, g) = \sigma \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \int_{\mathcal{E}_v} \gamma(h_e)[[f]][[g]]. \tag{4.40}$$

$\gamma(h_e)$ is the discontinuity scaling function below. We introduce the bilinear form

$$\tilde{B}_\delta = A_\delta + \tilde{D}_\delta. \tag{4.41}$$

Then, the *hp*-DG for the equation (2.1) reads as follows: find $f^h \in V_h^p$ such that

$$\tilde{B}_\delta(f^h, g) = \langle f_0, g_+ \rangle_0 \qquad \forall g \in V_h^p. \tag{4.42}$$

As in subsection 3.1.2, we use V_h^p to emphasize the polynomials degree $p := k$ in (4.1). Note that when $\gamma(h_e)$ is set to zero and the SD-parameter $\delta_K \approx h$, for all $K \in \mathcal{C}_h$, then (4.42) is identical to the method introduced in (4.11). In the sequel we assume that the solution f of the equation (2.1) is sufficiently smooth on Ω_v: namely $f \in L^2(I, \Omega_{x_\perp}, H_0^1(\Omega_v)) \cap L^2(I, \Omega_{x_\perp}, H^2(\Omega_v))$, therefore, f is continuous across interelement boundaries in Ω_v and hence $D_s(f, g) = 0$ for all $g \in V_h^p$. Consequently, the Galerkin orthogonality $\tilde{B}_\delta(f - f^h, g) = 0$ holds for all $g \in V_h^p$.

We shall derive the stability of the method (4.42) in the following norm

$$|||g|||^2_{\gamma,\delta} = |||g|||^2_{A_\delta} + |||g|||^2_{D_\delta} + \sigma \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \int_{\mathcal{E}_v} \gamma(h_e)[[g]]^2. \tag{4.43}$$

Lemma 4.4. *There is a constant $C > 0$ such that*

$$\tilde{B}_\delta(g, g) \geq C|||g|||^2_{\gamma,\delta}, \qquad \forall g \in V_h^p. \tag{4.44}$$

Proof. This is a simple observation namely, by (4.41) and (4.39), we have

$$\tilde{B}_\delta(g, g) = A_\delta(g, g) + D_\delta(g, g) + D_s(g, g), \tag{4.45}$$

with

$$D_s(g,g) = \sigma \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \int_{\mathcal{E}_v} \gamma(h_e)[[g]]^2. \tag{4.46}$$

Inserting (4.46) in (4.45), and using Lemma 4.1, we obtain the desired result. □

Before proceeding we state an approximation result for the space V_h^p (see, [13]). We consider $Q_k(K)$, the set of all polynomials of degree $\leq k$ in each variable on K.

Lemma 4.5. *Let $K \in \mathcal{C}_h$ and assume that $g \in H^s(K)$ for some integer $s \geq 1$. Then, for any integer $\mu = \min(p+1, s)$, and $p \geq 0$, we have that*

$$\|g - \mathrm{P}g\|_{L^2(\partial K)} \leq C \left(\frac{h_K}{p+1}\right)^{\mu-\frac{1}{2}} \|g\|_{\mu,K}, \tag{4.47}$$

where $\mathrm{P} : L^2(K) \to Q_p(K)$ is the usual L^2- projection of degree p on K.

We denote by P_v the univariate elementwise $L^2(\tau_v)$-projection onto the polynomials of degree p in the variable v for every $\tau_v \in T_h^v$. Local error estimates for $f - \mathrm{P}_v f$ can now be obtained from Lemma 4.5. Actually for $K \in \mathcal{C}_h$ we have

$$\|f - \mathrm{P}_v f\|_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)} \leq C \left(\frac{h_K}{p+1}\right)^{\mu-\frac{1}{2}} \|f\|_{L^2(I_m,\tau_{x_\perp},H^\mu(\tau_v))}. \tag{4.48}$$

where $K := I_m \times \tau_{x_\perp} \times \tau_v$. We also recall a restatement of Lemma 3.3: suppose

$$f \in L^2(I, \Omega_{x_\perp}, H_0^1(\Omega_v)) \cap L^2(I, \Omega_{x_\perp}, H^2(\Omega_v)), \tag{4.49}$$

and assume that for $s \geq 2$,

$$f|_K \in H^s(K), \qquad \forall K \in \mathcal{C}_h, \tag{4.50}$$

then, there is an interpolant $\Pi_p f \in L^2(I, \Omega_{x_\perp}, H_0^1(\Omega_v))$ which is continuous on Ω_v. Thus, by local interpolation error estimate (3.17), with $r = 1$, we have

$$\|f - \Pi_p f\|_{1,K} \leq C \frac{h_K^{\mu-1}}{p^{s-1}} \|f\|_{s,K}, \qquad \mu = \min(p+1, s). \tag{4.51}$$

Theorem 4.2. *For $h_e \in \mathcal{E}_v$ we define the scaling discontinuity function γ by*

$$\gamma(h_e) = \frac{p^2}{h_e}. \tag{4.52}$$

Assume that δ_K satisfies (3.15) *and the solution f satisfies* (4.49)-(4.50). *Then, there is a constant $C > 0$ independent of h and p such that for $\mu = \min(p+1, s)$,*

$$\begin{aligned} |||f - f^h|||_{\gamma,\delta}^2 &\leq C \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2\mu-1}} \|f\|_{\mu,K}^2 \\ &+ \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2s-2}} \left(\frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1} + \frac{h_K}{p^2 \delta_K}\right) \|f\|_{s,K}^2. \end{aligned} \tag{4.53}$$

Proof. We follow the proof of Theorem 4.1, except now we decompose the error as

$$e := f - f^h = (f - \tilde{f}^h) + (\tilde{f}^h - f^h) \equiv \eta + \xi, \tag{4.54}$$

where $\tilde{f}^h \in V_h^p$ is hp-interpolant of f satisfying (4.51), i.e. $\tilde{f}^h := \Pi_p f$. By virtue of Lemma 4.4, we have

$$C_I |||\xi|||_{\gamma,\delta} \leq \tilde{B}_\delta(\xi, \xi) = \tilde{B}_\delta(e - \eta, \xi) = \tilde{B}_\delta(-\eta, \xi), \tag{4.55}$$

where we use Galerkin orthogonality: $\tilde{B}_\delta(e,\xi)=0$ which follows form (4.42) with $g=\xi$ and the definition of the problem, given the assumed smoothness of f. Thus,

$$C_I|||\xi|||_{\gamma,\delta} \le |\tilde{B}_\delta(\eta,\xi)| \le |A_\delta(\eta,\xi)| + |\tilde{D}_\delta(\eta,\xi)|. \tag{4.56}$$

Since $\eta \in L^2(I,\Omega_{x_\perp},H_0^1(\Omega_v))$, we have

$$[[\eta]] = 0 \qquad \text{on } \mathcal{E}_v, \tag{4.57}$$

also

$$R(\eta)=0 \quad \text{on } \Omega, \qquad \text{and} \qquad r_e(\eta)=0 \quad \text{on } \Omega \quad \forall e\in\mathcal{E}_v. \tag{4.58}$$

Hence,

$$|\tilde{D}_\delta(\eta,\xi)| \le I + II + III, \tag{4.59}$$

where

$$\begin{aligned} I &= \sigma|(\nabla_v\eta,\nabla_v\xi)_{Q_L}|, \qquad II = \sigma|(\nabla_v\eta,R(\xi))_{Q_L}|,\\ III &= \sum_{K\in\mathcal{C}_h}\sigma\delta_K|(\Delta_v\eta,\xi_x+v\cdot\nabla_\perp\xi)_K|. \end{aligned} \tag{4.60}$$

I is estimated as in the Lemma 4.2. The orthogonal projector to $L^2(Q_L)$ yields

$$\begin{aligned} \sigma(\nabla_v\eta,R(\xi))_{Q_L} &= \sigma(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L} + \sigma(\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L}\\ &= \sigma(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L} + \sigma(\nabla_v\eta,R(\xi))_{Q_L} = T_1+T_2. \end{aligned} \tag{4.61}$$

By the definition of R and the shape regularity of $\mathcal{C}_h$, relating h_e to h_K,

$$\begin{aligned} T_1 &= \sigma\sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\sum_{e\in\mathcal{E}_v}\int_e [[\xi]]\mathbf{n}_v\cdot(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta)^0\\ &\le \sigma\|\sqrt{\gamma}[[\xi]]\|_{\mathcal{E}_v}\|\gamma^{-\frac12}(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta)^0\|_{\mathcal{E}_v}\\ &\le C\sigma\|\sqrt{\gamma}[[\xi]]\|_{\mathcal{E}_v}\left(\sum_{I_m\times\tau_{x_\perp}}\sum_{\tau_v\in T_h^v}p^{-2}h_{\tau_K}\|\nabla_v\eta-\mathrm{P}_v\nabla_v\eta\|^2_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)}\right)^{1/2}, \end{aligned}$$

where, in the first inequality, we used the notation

$$\|g\|_{\mathcal{E}_v} = \sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\sum_{e\in\mathcal{E}_v}\int_e g dv.$$

Further, since $\nabla_v(\Pi_p f)\in V_h^p\times V_h^p$ and the L_2-projection preserves polynomials,

$$\nabla_v\eta-\mathrm{P}_v\nabla_v\eta = \nabla_v f-\nabla_v\Pi_p f-\mathrm{P}_v\nabla_v f+\mathrm{P}_v\nabla_v\Pi_p f = \nabla_v f-\mathrm{P}_v\nabla_v f.$$

Hence,

$$T_1 \le C\sigma\|\sqrt{\gamma}[[\xi]]\|_{\mathcal{E}_v}\left(\sum_{I_m\times\tau_{x_\perp}}\sum_{\tau_v\in T_h^v}p^{-2}h_K\|\nabla_v f-\mathrm{P}_v\nabla_v f\|^2_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)}\right)^{1/2}.$$

Moreover, using (4.4) and (4.5), we estimate the T_2 term as

$$T_2 \le \sqrt{\sigma}\|(\nabla_v\eta)\|_{Q_L}\Big(\sigma\sum_{e\in\mathcal{E}_v}\|r_e(\xi)\|^2_{Q_L}\Big)^{1/2}.$$

It remains to estimate the term III. By inverse inequality and assumption (3.15),

$$\sigma\delta_K|(\Delta_v\eta,\xi_x+v\cdot\nabla_\perp\xi)_K| \le \sqrt{\sigma\delta_K}\|\nabla_v\eta\|_K\|\xi_x+v\cdot\nabla_\perp\xi\|_K. \tag{4.62}$$

Substituting the above estimates, for T_1 and T_2, into (4.61) and then using (4.62) and inserting the estimates for (4.60) into (4.59), Cauchy-Schwarz inequality yields

$$|\tilde{D}_\delta(\eta,\xi)| \le C_1|||\xi|||^2_{\gamma,\delta}+C\sigma\Bigg(\|(\nabla_v\eta)\|^2_{Q_L} + \sum_{I_m\times\tau_{x_\perp}}\sum_{\tau_v\in T_h^v} p^{-2}h_K\|\nabla_v f-\mathrm{P}_v\nabla_v f\|^2_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)}\Bigg), \tag{4.63}$$

where, $C_1 \le \frac{1}{3}C_I$. As for the term $|A_\delta(\eta,\xi)|$, using Lemma 4.2 and with $C_2 \le \frac{1}{3}C_I$,

$$|A_\delta(\eta,\xi)| \le C_2|||\xi|||^2_{\gamma,\delta}+C\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|^2_K+\delta_K\|\eta_x+v.\nabla_\perp\eta\|^2_K+\|\eta\|^2_{\partial K}\right). \tag{4.64}$$

Substituting (4.63) and (4.64) into (4.56), using a kick back argument and applying the error estimates (4.51) and (4.48) and trace inequality (3.19) we deduce that

$$|||\xi|||^2_{\gamma,\delta} \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\delta_K h_K^{-1}+\sigma h_K^{-1}+\frac{h_K}{p^2\delta_K}\right)\|f\|^2_{s,K}+\frac{h_K^{2\mu-1}}{p^{2\mu-1}}\|f\|^2_{\mu,K}. \tag{4.65}$$

Similarly, due to (4.57) and (4.58), for the interpolation error we get

$$|||\eta|||^2_{\gamma,\delta} \le C\sum_{K\in\mathcal{C}_h}\left(h_K^{-1}\|\eta\|^2_K+\sigma\|\nabla_v\eta\|^2_K+\delta_K\|\eta_x+v.\nabla_\perp\eta\|^2_K\right). \tag{4.66}$$

Hence, using (4.51) and trace inequality (3.19) we end up with

$$|||\eta|||^2_{\gamma,\delta} \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\delta_K h_K^{-1}+\frac{1}{p^2}+\frac{1}{p}+\sigma h_K^{-1}\right)\|f\|^2_{s,K}. \tag{4.67}$$

Inserting the bound for $|||\eta|||_{\gamma,\delta}$ and (4.65) in (4.56) we obtain the desired result. □

Remark 4.2. Let, in Theorem 4.2, $2 \le s \le p+1$ and choose $\delta_K = \frac{h_K^2}{\sigma C_I^2 p^4}$ for all $K \in \mathcal{C}_h$, then assuming $\mathcal{O}(\frac{\sigma}{h_K}) \sim 1$ for all $K \in \mathcal{C}_h$, we deduce from Theorem 4.2 that the discretization error in the norm $|||.|||_{\gamma,\delta}$, converges as $\mathcal{O}(\frac{h^{(\mu-\frac{1}{2})}}{p^{(\mu-1)}})$. Hence, the error bound is optimal in both h and p. The parameter δ_K may be selected as

$$\delta_K = C_\delta\frac{h_K}{p}, \qquad \forall K \in \mathcal{C}_h. \tag{4.68}$$

The constant C_δ is chosen subject to the constraint on δ_K in Theorem 4.2. In this case the parameter $\delta_K h_K^{-1}$ in (4.53) is equal to $\frac{1}{p}$, and the error of the method in DG-norm is of order $\mathcal{O}(\frac{h^{(\mu-\frac{1}{2})}}{p^{(\mu-\frac{1}{2})}})$, which is again simultaneously optimal in h and p.

Remark 4.3. The choices made for δ_K in Remark 4.2, are closely connected to the degeneracy of diffusion term in Fermi equation (2.1). The use of continuous interpolant in velocity space and the homogeneity of boundary condition on Ω_v cause the suboptimal stabilization terms in the method (4.42) to vanish.

Conclusion: We extend the result in [2]- [3] to the three dimensional problem. We present an h- and hp-a priori error analysis of both SD- and DG- schemes for Fermi equation. We show that both schemes are optimally convergent with respect to the mesh size h and the spectral degree p of approximating polynomial. For DG method we admit general 1-irregular meshes, it allows less restrictive smoothness assumptions compared to the SD method. Of course, for given h and p, the number of degrees of freedom in DG method is higher than in the SD case. The analytic

solution is non-negative with L_1 and L_∞ stability properties, see, e.g. [11]. Same stabilities are achieved for the approximate solutions by the construction. However, the positivity is guaranteed only for the limit. For positivity of final step approximate solutions excessive stability assuptions and uniform convergence are required. As for the computational aspects, the discretized problem being 5-dimensional is difficult to handle. One remedy would be considering the discrete velocity model of the Fermi equation combined with backward Euler method for penetration variable.

References

[1] H. Adams, *Sobolev Spaces*, Academic Press, New York, (1978).

[2] M. Asadzadeh, *Streamline diffusion methods for Fermi and Fokker-Planck equations*, Transport Theory Statist. Phys. 26 (1997), no. 3, 319-340.

[3] M. Asadzadeh, and A. Sopasakis, *On Fully Discrete Schemes for the Fermi Pencil-Beam Equations*. Comput. Methods Appl. Mech. Engrg. 191 (2002), 4641-4659.

[4] M. Asadzadeh and P. Kowalczyk, *Convergence of Streamline Diffusion Methods for the Vlasov-Poisson-Fokker-Planck System*, Numer. Meth. Part. Diff. Eqs., 21 (2005), 472-495.

[5] M. Asadzadeh and A. Sopasakis, *Convergence of a hp-Streamline Diffusion Scheme for Vlasov-Fokker-Planck system*, Math. Mod. Meth. Appl. Sci., 17 (2007), 1159-1182.

[6] I. Babuŝka and M. Suri, *The hp-version of the finite element method with quasiuniform meshes*, Math. Model. Numer. Anal. 21 (1987), 199-238.

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

[8] S. C. Brenner and L. R. Scott, *The Mathematical Theory of Finite Element Methods*, Springer-Verlag, 3:rd ed., (2008).

[9] F. Brezzi, G. Manzini, D. Marini, P. Pietra and A. Russo, *Discontinuous Galerkin approximations for elliptic problems*, Numer. Meth. Partial Diff. Equs., 16 (2000), no. 4, 365-378.

[10] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.

[11] P. Degond,*Global existence of smooth solutions for the Vlasov-Fokker-Planck equation in 1 and 2 space dimentions*, Ann Scient Èc Norm Sup, 4^e série, 19 (1986), 519-542.

[12] P. Houston and E. Süli, *Stabilized hp-finite element approximation of partial differential equations with nonnegative characteristic form*, Computing 66 (2001), 99-119.

[13] P. Houston, C. Schwab and E. Süli, *Discontinuous hp-finite element methods for advection-diffusion-reaction problems*, SIAM J. Numer. Anal. 39 (2002), 2133-2163.

[14] T. J. R. Hughes and M. Mallet, *A new finite element formulation for computational fluid dynamics. III. The generalized streamline operator for multidimensional advective-diffusive systems*, Comput. Methods Appl. Mech. Engrg. 58 (1986), no. 3, 305-328.

[15] H. Roos, M. Stynes, and L. Tobiska, *Robust Numerical Methods for Singularly Perturbed Differential Equations*, 2nd ed., Comput. Math. 24, Springer-Verlag, Berlin, 2008.

[16] L. R. Scott and S. Zhang, *Finite element interpolation of non-smooth functions satisfying boundary conditions*, Math. Comp. 54, (1990), 484–493.

[17] B. Stamm and T. P. Wihler, *hp-optimal discontinuous Galerkin methods for linear elliptic problems*, Math. Comp. 79 (2010), no. 272, 2117–2133.

Department of Mathematics, Chalmers University of Technology and Göteborg University, SE–412 96, Göteborg, Sweden

E-mail: mohammad@chalmers.se

Department of Mathematics, Isfahan University of Technology, Isfahan, Iran