CHALMERS | UNIVERSITY OF GOTHENBURG

PREPRINT 2011:28

On Convergence of the Streamline Diffusion and Discontinuous Galerkin Methods for the Multi-dimensional Fermi Pencil Beam Equation

MOHAMMAD ASADZADEH
EHSAN KAZEMI

Department of Mathematical Sciences
Division of Mathematics
CHALMERS UNIVERSITY OF TECHNOLOGY
UNIVERSITY OF GOTHENBURG
Gothenburg Sweden 2011

Preprint 2011:28

On Convergence of the Streamline Diffusion and Discontinuous Galerkin Methods for the Multi-dimensional Fermi Pencil Beam Equation

Mohammad Asadzadeh and Ehsan Kazemi

Department of Mathematical Sciences
Division of Mathematics
Chalmers University of Technology and University of Gothenburg
SE-412 96 Gothenburg, Sweden
Gothenburg, December 2011

Preprint 2011:28
ISSN 1652-9715

Matematiska vetenskaper
Göteborg 2011

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. We also introduce a penalty approach having computational advantageous in dealing with the diffusive part of the weak form.

1. Introduction

In this paper we study the approximate solution for the three-dimensional Fermi pencil beam equation using the streamline diffusion and discontinuous Galerkin finite element methods. We prove stability estimates and derive optimal convergence rates for the weighted current function, as in the convection dominated convection diffusion problems. This work extends the results introduced in [3] to the case of the multidimensional Fermi equation. The physical problem has diverse applications in, e.g. astrophysics, material science, electron microscopy, radiation cancer therapy, etc. We shall consider a pencil beam of particles normally incident on a slab of finite thickness. The particles enter at a single point, say at $\mathbf{x_0} := (0, 0, 0)$, in the direction of positive x-axis. The Fermi equation is obtained either as an asymptotic limit of the Fokker-Planck equation as the transport cross-section (σ_{tr}) gets smaller or as an asymptotic limit of the transport (linear Boltzmann) equation for vanishing transport cross-section and high (tends to ∞) total cross-section (σ_t) (the mean scattering angle is assumed to be small, and the large scattering is negligible). For details in derivation of Fermi equation we refer to [12]. (The physical quantities σ_{tr} and σ_t are defined below).

There are several points of concern with this type of problems: The Fermi equation considered in this paper is degenerate in both convection and diffusion in the sense that drift and diffusion are taking place in, physically, different domains . Besides the problem is convection dominated since the diffusion term has a very small coefficient compared to the coefficient of the convection term. Furthermore, the problem is associated with a boundary condition in form of product of certain δ functions, which are not suitable for numerical consideration involving L_2 norms. We have therefore considered model problems with somewhat smoother data approaching Dirac δ function. Finally, in spite of the assumption of no back-scattering, i.e. the scattering angle $-\pi/2 \le \theta \le \pi/2$, we still need to restrict the range of θ, through focusing or filtering, and avoid small intervals in vicinity of the endpoints $\pm\pi/2$, in

1991 *Mathematics Subject Classification.* 65M15, 65M60.

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

order to get, after scaling, bounded computational domains relevant in numerical considerations.

Fermi equation has closed form solutions for σ_{tr} being a constant or a function of only x. The subject of this paper is error estimates for the stationary (steady state), energy dependent, three space dimensional Fermi equation. In the present setting we have transformed and scaled the variables so that the x-direction, the direction of penetration of the beam, being perpendicular to the slab, may also be interpreted as the direction of the time variable. After scaling, the present technique treats all the variables as components of a multi-dimensional space variable.

The streamline diffusion method (SD-method) is a generalized form of the standard Galrekin method designed for the finite element studies of the hyperbolic problems, giving good stability and high accuracy. The SD-method which is used for our purpose in this paper is obtained by modifying the test function through adding a multiple of the "drift-terms" involved in the equation to the usual test function. This yields a weighted least square control of the residual of the finite element solution. See, e.g. [20] and [22] and the references therein for further details in the SD method. The discontinuous Galerkin method allows jump discontinuities across interelement boundaries in order to count for the local effects. Here we have considered both h and hp versions of SD and DG methods. As for numerical implementation, a characteristic method, as well as a semi-streamline diffusion for Fermi pencil beam equation have been studied in [5] and [8], respectively.

An outline of this paper is as follows: In Section 2, we introduce the model problem and present some notations. Section 3 is devoted to the study of stability estimates and proof of the convergence rates for the, h and hp, streamline diffusion approximation of the Fermi equation. Section 4 is the discontinuous Galerkin counterpart of Section 3, where we have also studied a penalty approach.

2. Notations and preliminaries

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

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

where $f_0 \in L_2(\Omega_0)$, with $\Omega_0 := \{x = 0\} \times \Omega_{x_\perp} \times \Omega_v$ and the outflow boundary is given by

$$\Gamma_v^- = \{x_\perp \in \partial\Omega_{x_\perp} : \mathbf{n}(x_\perp).v < 0\}, \quad \text{for } v \in \Omega_v. \tag{2.2}$$

Here $\mathbf{n}(x_\perp)$ is the outward unit normal to $\partial\Omega_{x_\perp}$ at the point $x_\perp \in \partial\Omega_{x_\perp}$, $x_\perp = (y, z)$, $v = (v_1, v_2)$, $\nabla_\perp = (\frac{\partial}{\partial y}, \frac{\partial}{\partial z})$ and, $\sigma_{tr} = \sigma_{tr}(x, y, z)$ is the transport cross-section (actually $\sigma_{tr} = \sigma_{tr}[E(x, y, z)]$ is energy dependent).
We shall use a finite element structure on $\Omega_{x_\perp} \times \Omega_v$: by letting $T_h^{x_\perp} = \{\tau_{x_\perp}\}$ 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, $T_h = T_h^{x_\perp} \times T_h^v = \{\tau_{x_\perp} \times \tau_v\} = \{\tau\}$ will be a subdivision of $\Omega = \Omega_{x_\perp} \times \Omega_v$ with elements $\{\tau_{x_\perp} \times \tau_v\} = \{\tau\}$. We also use the partition $0 = x_0 < x_1 < \ldots < x_M = L$ of the interval $I = (0, L]$ into subintervals $I_m = (x_{m-1}, x_m)$, $m = 1, ..., M$. Now, 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 parameter $h = \text{diam } K$. 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 master 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 purely the open unit hypercube in $\mathbb{R}^5$). Let $P_p(K)$ be the set of all polynomials of degree at

most p on K; in $x, x_\perp$ and v, and define the finite element space

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

with

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

and

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

Moreover

$$\begin{array}{ll} (f,g)_m = (f,g)_{S_m}, & \|g\|_m^2 = (g,g)_m, \\ \langle f,g\rangle_m = (f(x_m,.,.), g(x_m,.,.))_\Omega, & |g|_m^2 = \langle g,g\rangle_m, \\ \langle f,g\rangle_{\Gamma^-} = \int_{\Gamma^-} fg(\beta\cdot\mathbf{n})ds, & \langle f,g\rangle_{\Gamma_m^-} = \int_{I_m}\langle f,g\rangle_{\Gamma^-}ds, \\ \langle f,g\rangle_{\Gamma_I^-} = \int_I \langle f,g\rangle_{\Gamma^-}ds, & \end{array} \tag{2.6}$$

where

$$\Gamma^- = \{(x_\perp, v) \in \Gamma = \partial(\Omega_{x_\perp} \times \Omega_v) : \beta\cdot\mathbf{n} < 0\},$$

$\beta = (v, \mathbf{0})$ and $\mathbf{n} = (\mathbf{n}_{x_\perp}, \mathbf{n}_v)$ with $\mathbf{n}_{x_\perp}$ and $\mathbf{n}_v$ being outward unit normals to $\partial\Omega_{x_\perp}$ and $\partial\Omega_v$, respectively. Throughout the paper C will denote a constant not necessarily the same at each occurrence and independent of the parameters, and functions involved in the problem, unless otherwise specifically specified. Finally 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,

$$(D_1w_1, D_2w_2)_{Q'} = \sum_{K\in\mathcal{C}_h'} (D_1w_1, D_2w_2)_K, \qquad Q' = \bigcup_{K\in\mathcal{C}_h'} K, \tag{2.7}$$

where $(.,.)_Q$ is the usual $L_2(Q)$ scalar product and $\|.\|_Q$ is the corresponding $L_2(Q)$-norm.

3. Streamline diffusion method

3.1. **Streamline diffusion method with discontinuity in x.** For σ_{tr} constant or $\sigma_{tr} = \sigma_{tr}(x)$ one can obtain closed form analytic solution for the Fermi equation. We prove stability lemmas for the discrete problem in general three dimensional case, i.e. with $\sigma = \frac{1}{2}\sigma_{tr} = \frac{1}{2}\sigma_{tr}(x,y,z)$, using also the corresponding variational formulation we derive a priori error estimates. Through out the paper, the parameter σ is, basically, of the order of mesh size or smaller. In order to study the distribution of the particle beams in a certain depth, e.g. $x = x_d$, a reasonable initial guess would be obtained using the information in some previous distinct depths $x = x_i,\ i = 1, 2, \dots, n$, with $x_i < x_{i+1}$, one may assume various filters installed in different depths to control or adjust the beam intensity. This corresponds to considering discontinuities in x-direction. In this section we study the SD-method for the Fermi equation given by (2.1) with the trial functions being continuous in $x_\perp$ and v but may have jump discontinuities in x. In applications, normally, these discontinuities in x are in a quasi-uniform partition $\mathcal{T}_h : \bar{x}_0 = 0 < \bar{x}_1 < \ldots < \bar{x}_N = L$ of $[0, L]$ that contains all $x_i : s,\ 1 \le i \le n$, and also possibly more additional discretization points in x.

We present the jump in x-direction

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

where

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

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

$$\begin{aligned}(f_x^h + v\cdot\nabla_\perp f^h, g + \delta(g_x + v\cdot\nabla_\perp g))_m + \sigma(\nabla_v f^h, \nabla_v g)_m \\ -\delta\sigma(\Delta_v f^h, g_x + v\cdot\nabla_\perp g)_m + \langle f_+^h, g_+\rangle_m - \langle f_+^h, g_+\rangle_{\Gamma_m^-} = \langle f_-^h, g_+\rangle_m.\end{aligned} \tag{3.3}$$

Below we study this streamline diffusion method for Fermi equation (2.1) in two different approaches: h-version and hp-version. In the h- version of the SD-method, assuming f^h to be the approximate solution and using test functions of the form $g + \delta(g_x + v\cdot\nabla_\perp g)$ where δ is a small parameter of order h (or h^α, $\alpha > 1$) , would supply us with a necessary (missing) diffusion term of order h in the direction of streamlines: $(1, v, \mathbf{0})$. More specifically, in the stability estimates we will be able to control an extra term of the form $h\|g_x + v\cdot\nabla_\perp g\|$. In the hp- studies, however, the choice of δ is somewhat involved and in addition to the equation type, it also depends on the choice of the parameters h and p which are chosen locally (elementwise) in an optimal manner. Therefore in hp-analysis, δ would appropriately appear as an elementwise (local) parameter.

3.1.1. *The h-version of the SD-method.* To proceed, we formulate the finite element approximation of (2.1), using SD-method with jump discontinuities in x. Introducing the bilinear form

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

where

$$\begin{aligned}B(f,g) = (f_x + v\cdot\nabla_\perp f, g + h(g_x + v\cdot\nabla_\perp g))_{Q_L} + \sigma(\nabla_v f, \nabla_v g)_{Q_L} \\ - h\sigma(\Delta_v f, g_x + v\cdot\nabla_\perp g)_{Q_L} + \langle f, g\rangle_0 - \langle f, g\rangle_{\Gamma^-}.\end{aligned} \tag{3.5}$$

and the linear form

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

we may rewrite (3.3) in global form as

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

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

$$[||g||]^2 = \tfrac{1}{2}\left[|||g|||^2 + h \parallel g_x + v.\nabla_\perp g \parallel_{Q_L}^2 + \sum_{m=1}^{M-1} |[g]|_m^2\right], \tag{3.7}$$

where

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

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

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

Proof. We use the definition of $\tilde{B}$ in (3.4) and write

$$\begin{aligned}\tilde{B}(g,g) &= B(g,g) + \sum_{m=1}^{M-1}\langle [g], g_+\rangle_m + \langle g_+, g_+\rangle_0 - \langle g_+, g_+\rangle_{\Gamma_I^-} \\ &= h\|g_x + v\cdot\nabla_\perp g\|_{Q_L}^2 + \sigma\|\nabla_v g\|_{Q_L}^2 - h\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_{Q_L} \\ &\quad - \langle g_+, g_+\rangle_{\Gamma_I^-} + (g_x, g)_{Q_L} + \sum_{m=1}^{M-1}\langle [g], g_+\rangle_m + \langle g_+, g_+\rangle_0 + (v\cdot\nabla_\perp g, g).\end{aligned} \tag{3.9}$$

Integrating by parts we have

$$(g_x, g) = \tfrac{1}{2}\langle g, g\rangle \mid_{x_{n-1}^+}^{x_n^-} = \tfrac{1}{2}\int_\Omega [g^2(x_n^-) - g^2(x_{n-1}^+)]. \tag{3.10}$$

Using Green's formula we have also

$$(v\cdot\nabla_\perp g, g) - \langle g_+, g_+\rangle_{\Gamma_I^-} = \tfrac{1}{2}\int_{I\times\partial\Omega} g^2(\beta\cdot\mathbf{n})dv - \int_{I\times\Gamma^-} g^2(\beta\cdot\mathbf{n})dv. \tag{3.11}$$

Thus, rearranging the terms we may write

$$\begin{aligned}(g_x, g)_{Q_L} + &\sum_{m=1}^{M-1}\langle [g], g_+\rangle_m + \langle g_+, g_+\rangle_0 - \langle g_+, g_+\rangle_{\Gamma_I^-} + (v\cdot\nabla_\perp g, g)_{Q_L}\\ &= \frac{1}{2}\left(\sum_{m=1}^{M-1} \mid [g]\mid_m^2 + \mid g_-\mid_M^2 + \mid g_+\mid_0^2 + \int_{I\times\partial\Omega} g^2|\beta\cdot\mathbf{n}|dv\right).\end{aligned} \tag{3.12}$$

To estimate the term involving $\Delta_v g$ we use inverse estimate and assumption on σ to obtain

$$h\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_{Q_L} \le \tfrac{1}{2}(\sigma\|\nabla_v g\|_{Q_L}^2 + h\|g_x + v\cdot\nabla_\perp g\|_{Q_L}^2). \tag{3.13}$$

Combining (3.9)-(3.13) will give the desired result. □

We shall also need the following interpolation error estimates, see Ciarlet [16]: Let $f\in H^{r+1}(\Omega)$ then there exists an interpolant $\tilde f^h \in V_h$ of f such that

$$\|f-\tilde f^h\|_{Q_L} \quad \le Ch^{r+1}\|f\|_{r+1,Q_L}, \tag{3.14}$$

$$\|f-\tilde f^h\|_{1,Q_L} \quad \le Ch^{r}\|f\|_{r+1,Q_L}, \tag{3.15}$$

$$\|f-\tilde f^h\|_{\partial Q_L} \quad \le Ch^{r+1/2}\|f\|_{r+1,Q_L}. \tag{3.16}$$

Our main result in this section is as follows:

Theorem 3.1. *There is a constant C such that for f and f^h satisfying in* (2.1) *and* (3.6)*, respectively, we have*

$$[||f-f^h||] \le Ch^{k+1/2}\|f\|_{k+1,Q_L}. \tag{3.17}$$

Proof. Let $\tilde f^h\in V_h$ be an interpolant of the exact solution f and $\eta = f-\tilde f^h$. The error term can be split as

$$e := f - f^h = (f-\tilde f^h) - (f^h - \tilde f^h) = \eta - \xi. \tag{3.18}$$

Now since $\xi\in V_h$, we have the Galerkin orthogonality property $\tilde B(e,\xi) = 0$ which follows from (3.6) with $g=\xi$ and the definition of boundary value problem (2.1). Thus, we have using Lemma 3.1 and (3.4), that

$$\begin{aligned}[||\xi||]^2 \le \tilde B(\xi,\xi) =& \tilde B(\eta - e,\xi) = \tilde B(\eta,\xi)\\ =&(\eta_x + v\cdot\nabla_\perp\eta, \xi + h(\xi_x + v\cdot\nabla_\perp\xi))_{Q_L}\\ &+\sigma(\nabla_v\eta,\nabla_v\xi)_{Q_L} - h\sigma(\Delta_v\eta, \xi_x + v\cdot\nabla_\perp\xi)_{Q_L}\\ &+\sum_{m=1}^{M-1}\langle[\eta],\xi_+\rangle_m + \langle\eta_+,\xi_+\rangle_0 - \langle\eta,\xi_+\rangle_{\Gamma_I^-}.\end{aligned} \tag{3.19}$$

Integrating by parts we have

$$\begin{aligned}(\eta_x + v\cdot\nabla_\perp\eta,\xi)_{Q_L} + &\sum_{m=1}^{M-1}\langle[\eta],\xi_+\rangle_m + \langle\eta_+,\xi_+\rangle_0 - \langle\eta,\xi_+\rangle_{\Gamma_I^-}\\ &= -(\eta, \xi_x + v\cdot\nabla_\perp\eta\xi)_{Q_L} + \langle\eta_-,\xi_-\rangle_M\\ &\quad -\sum_{m=1}^{M-1}\langle\eta_-,[\xi]\rangle_m + \int_{I\times\partial\Omega}\eta\xi|\beta\cdot\mathbf{n}|.\end{aligned} \tag{3.20}$$

Then using Cauchy-Schwarz inequality we get

$$\sigma(\nabla_v\eta,\nabla_v\xi)_{Q_L}\leq\sigma\|\nabla_v\eta\|^2_{Q_L}+\frac{\sigma}{4}\|\nabla_v\xi\|^2_{Q_L}, \tag{3.21}$$

and

$$\sum_{m=1}^{M-1}\langle\eta_-,[\xi]\rangle_m\leq\sum_{m=1}^{M-1}|\eta_-|^2_m+\frac{1}{4}\sum_{m=1}^{M-1}|[\xi]|^2_m. \tag{3.22}$$

By inverse inequality we can also write

$$h\sigma(\Delta_v\eta,\xi_x+v\cdot\nabla_\perp\xi)_{Q_L}\leq Ch^{-1}\|\eta\|^2_{Q_L}+\tfrac{h}{4}\|\xi_x+v\cdot\nabla_\perp\xi\|_{Q_L}. \tag{3.23}$$

Then, combining the estimates (3.19)-(3.23) gives

$$\begin{aligned}|\tilde{B}(\eta,\xi)|\leq\frac{1}{4}[||\xi||]^2+C\Bigg[&h^{-1}\|\eta\|^2_{Q_L}+\sum_{m=1}^{M-1}|\eta_-|^2_m+h\|\eta_x+v\cdot\nabla_\perp\eta\|^2_{Q_L}\\&+\sigma\|\nabla_v\eta\|^2_{Q_L}+\int_{I\times\partial\Omega}\eta^2|\beta\cdot\mathbf{n}|dvds\Bigg].\end{aligned} \tag{3.24}$$

Now by standard interpolation theory we have (see Ciarlet [16], p.123)

$$\begin{aligned}\Bigg[\|\eta\|^2_{Q_L}+h\sum_{m=1}^{M-1}|\eta_-|^2_m+h^2\|\eta\|^2_{1,Q_L}+h\int_{I\times\partial\Omega}\eta^2|\beta\cdot\mathbf{n}|dvds\Bigg]^{\frac{1}{2}}\\\leq Ch^{k+1}\|f\|_{k+1,Q_L}.\end{aligned} \tag{3.25}$$

Thus

$$[||\xi||]^2\leq Ch^{2k+1}, \tag{3.26}$$

and since $[||\eta||]$, the interpolation error, is of the same order as $[||\xi||]$, we have the desired result. □

Remark 3.1. Here are some features of problem (2.1): (i) The lack of pure current term for the beam problem, i.e. no absorption on the left hand side of the equation, will lead to stability with no explicit L_2-norm control. Besides, in all the above estimates the semi-norms, (L_2-norms of partial derivatives), appear with a small coefficients of order $\mathcal{O}(\sqrt{h})$. Since the test functions are zero on part of $\partial\Omega$ with positive Lebesgue measure, we could again use a version of the Poincare-Friedricks inequality and obtain an estimate for the L_2-norm with the same coefficients as for the semi-norms involved in the weighted stability norm, i.e. we add a L_2-norm with a coefficient of order $\mathcal{O}(\sqrt{h})$ to the $[||.||]$ norm in Lemma 3.1. However, a better approach would be through Lemma 3.2 (cf. [3]) below, in a situation where jump discontinuities are introduced and included in the stability norm $[||.||]$. This approach improves the L_2-norm estimate regaining the factor $h^{1/2}$.

Lemma 3.2. *For any constant $C_1>0$, we have for $g\in V_h$,*

$$\|g\|_{Q_L}\leq\Bigg[\tfrac{1}{C_1}\|g_x+v.\nabla_\perp g\|^2_{Q_L}+\sum_{m=1}^{M}|g_-|^2_m+\int_{I\times\partial\Omega}g^2|\beta\cdot\mathbf{n}|\Bigg]he^{C_1h} \tag{3.27}$$

Proof. See the argument in the proof of Lemma 4.2 in [3]. □

3.1.2. *The hp-version of the SD-method.* In this section, we analyze the hp-version of the streamline diffusion method for Fermi equation (2.1). We derive error bounds which are simultaneously optimal in both mesh size h and the spectral order p in a stabilization parameter $\delta \sim \left(\frac{h^2}{\sigma p^4}\right)$. Below we formulate the local SD-method for problem 2.1 and extend the analysis of h-version to hp-version for local case. To this end we consider the bilinear form $\hat{B}_\delta(.,.)$ defined by

$$\hat{B}_\delta(f,g) = \sum_{K\in\mathcal{C}_h} [(f_x + v\cdot\nabla_\perp f, g + \delta_K(g_x + v\cdot\nabla_\perp g))_K + \sigma(\nabla_v f, \nabla_v g)_K$$

$$-\delta_K \sigma(\Delta_v f, g_x + v\cdot\nabla_\perp g)_K] + \sum_{m=1}^{M-1} \langle [f], g_+\rangle_m + \langle f, g\rangle_0 - \langle f, g\rangle_{\Gamma^-}$$

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 depends on the nature of the coefficients in the partial differential equation and will be discussed in more details later. We now define the local version of (3.3) as follows: Find $f^h \in V_h^p$ such that

$$\hat{B}_\delta(f^h, g) = \hat{L}(g) \qquad \forall g \in V_h^p, \tag{3.28}$$

where V_h^p stands for V_h when the bases functions are polynomials of degree at most p in all variables x, $x_\perp$ and v. Note that in the h version of the SD-approach we interpret $(.,.)_{Q_L}$ as $\sum_{m=1}^M (.,.)_m$ and, assuming discontinuities in the x variable, 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 $x_m : s,\ m = 1, \dots, M-1$. 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 parameter δ_K. We also define the norm $[||.||]_\delta$, in a natural way obtained from (3.7) and (3.8) by replacing h by δ_K and considering its local effects as

$$[||g||]_\delta^2 = \frac{1}{2}\left[|||g|||^2 + \sum_{K\in\mathcal{C}_h} \delta_K \parallel g_x + v.\nabla_\perp g \parallel_K^2 + \sum_{m=1}^{M-1} |[g]|_m^2\right]. \tag{3.29}$$

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

$$C_0 h_K^5 \le \text{meas}(K), \qquad \forall K \in \bigcup_{h>0}\{\mathcal{C}_h\}, \tag{3.30}$$

where meas(K) is the diameter of five dimensional sphere inscribed in K.

Lemma 3.3. *Assume that the local SD-parameter δ_K is selected in the range*

$$0 < \delta_K \le \frac{h_K^2}{\sigma C_I^2 p^4}, \qquad \forall K \in \mathcal{C}_h, \tag{3.31}$$

where C_I is the constant in an inverse estimate. Then the bilinear form $\hat{B}_\delta(.,.)$ is coercive on $V_h^p \times V_h^p$, i.e.

$$\hat{B}_\delta(g,g) \ge \frac{1}{2}[||g||]_\delta^2, \qquad \forall g \in V_h^p. \tag{3.32}$$

Proof. The proof is based on the same argument as in the proof of Lemma 3.1, except the estimate of the term involving $\delta_K\sigma$, where we apply Cauchy-Schwarz and inverse inequalities together with the assumption on δ_K, to get

$$\begin{aligned}
&\delta_K\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_K\\
&\le \frac{1}{2}C_I h_K^{-1}p^2\sqrt{\sigma\delta_K}\left[\sigma\|\nabla_v g\|_K^2 + \delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2\right]\\
&\le \frac{1}{2}\left[\sigma\|\nabla_v g\|_K^2 + \delta_K\|g_x + v\cdot\nabla_\perp g\|_K^2\right].
\end{aligned} \tag{3.33}$$

□

In what follows we shall use the following approximation property: Let $g \in H^s(K)$ and $\|.\|_{s,K}$ be the usual 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 for any $0 \le r \le s$ the following estimate holds true (see [10]),

$$\|g - \Pi_p g\|_{r,K} \le C\frac{h_K^{\mu-r}}{p^{s-r}}\|g\|_{s,K}, \tag{3.34}$$

where $s \ge 0$, and $\mu = \min(p+1, s)$. We shall also require a global counterpart of the above approximation result for the finite element space V_h^p, so in the sequel we adopt the following:

Lemma 3.4. *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 $s \ge r$ and $K \in \mathcal{C}_h$; there exists an interpolant $\Pi_p g \in V_h^p$ of g which is continuous on Ω such that*

$$\|g - \Pi_p g\|_{1,K} \le C\frac{h_K^{\mu-1}}{p^{s-1}}\|g\|_{s,K}, \tag{3.35}$$

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

Proof. See, e.g. [18] where a proof is outlined, assuming certain regularity degree. More elaborated proof can be found in [26], [13] and the references therein. □

We shall also need the following trace inequality:

$$\|\eta\|_{\partial K}^2 \le C(\|\nabla\eta\|_K\|\eta\|_K + h_K^{-1}\|\eta\|_K^2), \qquad \forall K \in \mathcal{C}_h. \tag{3.36}$$

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 3.4. Let f^h be the solution of* (3.28) *and assume that the SD-parameter δ_K satisfies $0 < \delta_K \le \frac{h_K^2}{\sigma C_I^2 p^4}$ for each $K \in \mathcal{C}_h$. Then the following error bound holds true*

$$[||f - f^h||]_\delta^2 \le C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1} + \frac{h_K}{\delta_K p^2}\right)\|f\|_{s,K}^2. \tag{3.37}$$

Proof. Using triangle inequality we get

$$[||f - f^h||]_\delta \le [||\eta||]_\delta + [||\xi||]_\delta, \tag{3.38}$$

where $\eta = f - \Pi_p f$ and $\xi = f^h - \Pi_p f$. $\Pi_p f \in V_h^p$ is the conforming interpolant in Lemma 3.4. Using Lemma 3.3 and Galerkin orthogonality $\hat{B}_\delta(e, \xi) = 0$ we get

$$\begin{aligned}
\frac{1}{2}[||\xi||]_\delta^2 &\le \hat{B}_\delta(\xi, \xi) = \hat{B}_\delta(\eta, \xi) - \hat{B}_\delta(e, \xi) = \hat{B}_\delta(\eta, \xi) \\
&= \sigma(\nabla_v \eta, \nabla_v \xi)_{Q_L} - \sigma \sum_{K \in \mathcal{C}_h} \delta_K (\Delta_v \eta, \xi_x + v \cdot \nabla_\perp \xi)_K \\
&\quad + (\eta_x + v \cdot \nabla_\perp \eta, \xi)_{Q_L} + \sum_{K \in \mathcal{C}_h} \delta_K (\eta_x + v \cdot \nabla_\perp \eta, \xi_x + v \cdot \nabla_\perp \xi)_K \\
&\quad + \sum_{m=1}^{M-1} \langle [\eta], \xi_+ \rangle_m + \langle \eta_+, \xi_+ \rangle_0 - \langle \eta, \xi_+ \rangle_{\Gamma_I^-} = \sum_{i=1}^{7} T_i.
\end{aligned} \tag{3.39}$$

The terms T_1 and T_3 to T_7 can be estimated by the same techniques as in the proof of Theorem 3.1. Further, using the inverse inequality and assumptions on σ and δ_K we get

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

We shall rewrite the estimates above concisely as

$$|||\xi|||_\delta \le C(I_1 + I_2), \tag{3.40}$$

where

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

Below we estimate I_1 and I_2 separately. As for I_1 we have using Lemma 3.4 and assumption on δ_K,

$$I_1 \le C \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-2}}{p^{2s-2}} (\delta_K^{-1} \frac{h_K^2}{p^2} + \delta_K + \sigma) \|f\|_{s,K}^2. \tag{3.41}$$

As, for the term I_2, we have from trace estimate (3.36),

$$I_2 \le \sum_{K \in \mathcal{C}_h} (\frac{h_K^{\mu-1}}{p^{s-1}} \frac{h_K^\mu}{p^s} + h_K^{-1} \frac{h_K^{2\mu}}{p^{2s}}) \|f\|_{s,K}^2 \le \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2s-1}} (1 + \frac{1}{p}) \|f\|_{s,K}^2. \tag{3.42}$$

Hence from (3.40)-(3.42) we get that

$$[||\xi||]_\delta^2 \le C \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2s-2}} (\frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1} + \frac{h_K}{\delta_K p^2}) \|f\|_{s,K}^2. \tag{3.43}$$

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

$$[||\eta||]_\delta^2 \le C \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2s-2}} (\frac{1}{p} + \sigma h_K^{-1} + \delta_K h_K^{-1}) \|f\|_{s,K}^2. \tag{3.44}$$

Substituting the estimates (3.43)-(3.44) into (3.38), we get the desired result and the proof is complete. □

Remark 3.2. In Theorem 3.2, we chose δ_K for all $K \in \mathcal{C}_h$ when σ is small compared to h_k and $\frac{1}{p}$. The parameter C_δ is selected in a way that δ_K satisfies hypothesis of Theorem 3.2. This particular choice of δ_K is motivated by our analysis in the discretization error (3.37) in the norm $[||.||]_\delta$, in order to give hp-error bound as,

$$[||f - f^h||]_\delta^2 \leq C \sum_{K \in \mathcal{C}_h} \frac{h_K^{2\mu-1}}{p^{2s-1}} \|f\|_{s,K}^2. \tag{3.45}$$

We note that our assumption on σ has a key role on obtaining the optimality of the error bound simultaneously in h and p.

Remark 3.3. The assumptions of Lemma 3.4, 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. Later, for the discontinuous Galerkin counterpart of current analysis, we will ease the requirement of Lemma 3.4.

Remark 3.4. For notational simplicity we have not chosen to allow an element-by-element variation of the polynomial degree p and the local Sobolev smoothness parameter s of the analytical solution f; however our analysis can be extended easily to this case by replacing p by p_K, s by s_K and $\|f\|_s$ by $\|f\|_{s,K}$ for $K \in C_h$. Subsequently, in the local approximation (3.34), $\mu = \min(p+1, s)$ is replaced by $\mu_K = \min(p_K + 1, s_K)$.

4. Discontinuous Galerkin

4.1. **Description of discontinuous Galerkin (DG)-method.** In the DG-method we assume trial functions with discontinuities in both space and velocity variables. So trial functions are polynomials of degree $k \geq 1$ on each element K of triangulation and may be discontinuous across inter-element boundaries in all variables. We may define for $K \in \mathcal{C}_h$

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

where $\tilde{\beta} = (1, v, \mathbf{0})$ and $\mathbf{n} = (\mathbf{n}_x, \mathbf{n}_{x_\perp}, \mathbf{n}_v)$ denotes the outward unit normal to $\partial K \subset Q_L$.

To derive a variational formulation, for the diffusive part of (2.1), based on discontinuous trial functions, we shall introduce an operator R as defined in, e.g. [14] and [15]. To this approach, we first define the spaces $\tilde{V}$, V_h and $\mathbf{W}_h$ as

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

More precisely, given $g \in \tilde{V}$ we define $R : \tilde{V} \to \mathbf{W}_h$ by the following relation

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

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

$$\begin{aligned} (\chi)^0 &:= \tfrac{\chi + \chi^{ext}}{2}, \\ [[\chi]] &:= \chi - \chi^{ext}, \end{aligned} \tag{4.2}$$

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

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

for any element τ_v of the triangulation of Ω_v. As a consequence of this we have the following estimate:

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

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

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

Using these notations, the discontinuous Galerkin finite element method for Fermi equation can now be formulated as follows: Find $f^h \in V_h$ such that for all $g \in V_h$,

$$A_\delta(f^h, g) + D_\delta(f^h, g) = \langle f_0, g_+ \rangle_0, \tag{4.6}$$

where the bilinear forms A_δ and D_δ correspond to the convective and diffusive parts of the equation (2.1) and are defined as follows:

$$\begin{aligned} A_\delta(f^h, g) = &\sum_{K \in \mathcal{C}_h} (f_x^h + v \cdot \nabla_\perp f^h, g + \delta_K(g_x + v \cdot \nabla_\perp g))_K \\ &+ \langle f_+, g_+ \rangle_0 + \sum_{K \in \mathcal{C}_h} \int_{\partial K_-(\tilde{\beta})'} [f] g_+ |\tilde{\beta} \cdot \mathbf{n}|, \end{aligned} \tag{4.7}$$

with $\partial K_-(\tilde{\beta})' = \partial K_-(\tilde{\beta}) \backslash \{0\} \times \Omega$ and

$$\begin{aligned} D_\delta(f^h, g) = &\sigma(\nabla_v f^h, \nabla_v g)_{Q_L} + \sigma(\nabla_v f^h, R(g))_{Q_L} + \sigma(R(f^h), \nabla_v g)_{Q_L} \\ &+ \lambda\sigma \sum_{e \in \mathcal{E}_v} (r_e(f^h), r_e(g))_{Q_L} - \sum_{K \in \mathcal{C}_h} \delta_K \sigma(\Delta_v f^h, g_x + v \cdot \nabla_\perp g)_K. \end{aligned} \tag{4.8}$$

Here, $[f^h] = f_+^h - f_-^h$ where $f_\pm^h$ is defined as in (3.2), $\delta_K > 0$ is a positive constant on element K and $\lambda > 0$ is a given constant. We also define the norms corresponding to (4.7) and (4.8) by

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

and

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

We note that, in general $[g]$ is distinct from the jump $[[g]]$ defined in (4.2), in the sense that, the latter depends on element numbering as well. Recall that since $\tilde{\beta} = (1, v, 0)$ 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 $\delta_K := h$, for all $K \in \mathcal{C}_h$, then the problem (4.6) can be formulated as

$$B_*(f^h, g) = \langle f_0, g_+\rangle_0, \qquad \forall g \in V_h, \tag{4.9}$$

where

$$B_*(f^h, g) = A(f^h, g) + D(f^h, g). \tag{4.10}$$

For notational convenience we shall suppress the index δ from A_δ and D_δ, when we set $\delta_K := h$ for all $K \in \mathcal{C}_h$. Then, the stability lemma for bilinear forms A_δ and D_δ is:

Lemma 4.1 (Extended coercivity Lemma)**.** *Suppose that δ_K satisfies* (3.31) *for all $K \in \mathcal{C}_h$, then, there is a constant $\alpha > 0$ such that*

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

Proof. Using the definition of A_δ in (4.7) we deduce that

$$\begin{aligned} A_\delta(g, g) =&(g_x + v\cdot\nabla_\perp g, g)_{Q_L} + \sum_{K\in\mathcal{C}_h} \delta_K \|g_x + v\cdot\nabla_\perp g\|^2_K + |g|^2_0 \\ &+ \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'} [g]g_+|\tilde\beta\cdot\mathbf{n}|. \end{aligned} \tag{4.11}$$

Integrating by parts we have

$$\begin{aligned} (g_x + v\cdot\nabla_\perp g, g)_{Q_L} &= \frac{1}{2}\sum_{K\in\mathcal{C}_h}\int_{\partial K} g^2\tilde\beta\cdot\mathbf{n} \\ &= \frac{1}{2}\left[-\sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'} g_+^2|\tilde\beta\cdot\mathbf{n}| + \sum_{K\in\mathcal{C}_h}\int_{\partial K_+(\tilde\beta)'} g_-^2|\tilde\beta\cdot\mathbf{n}|\right], \end{aligned} \tag{4.12}$$

and so, we obtain

$$\begin{aligned} &(g_x + v\cdot\nabla_\perp g, g)_{Q_L} + \sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'} [g]g_+|\tilde\beta\cdot\mathbf{n}| + |g|^2_0 \\ &= \frac{1}{2}\left[\sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde\beta)'} [g]^2|\tilde\beta\cdot\mathbf{n}| + \int_{I\times\partial\Omega_+} g^2|v\cdot\mathbf{n}_{x_\perp}| + |g|^2_0 + |g|^2_M\right]. \end{aligned} \tag{4.13}$$

Similarly, by the definition of D_δ and using (4.5) we get,

$$\begin{aligned} D_\delta(g, g) =&\sigma\|\nabla_v g\|^2_{Q_L} + 2\sigma(\nabla_v g, R(g))_{Q_L} + \lambda\sigma\sum_{K\in\mathcal{C}_h}\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|^2_K \\ &- \sum_{K\in\mathcal{C}_h}\delta_K\sigma(\Delta_v g, g_x + v\cdot\nabla_\perp g)_K. \end{aligned} \tag{4.14}$$

Using (4.4) for some $0 < \varepsilon < \frac{1}{2}$ we obtain

$$\begin{aligned} 2\sigma(\nabla_v g, R(g))_{Q_L} &\leq \sigma\sum_{K\in\mathcal{C}_h}\left[\varepsilon\|\nabla_v g\|^2_K + \frac{1}{\varepsilon}\|R(g)\|^2_K\right] \\ &\leq \sigma\sum_{K\in\mathcal{C}_h}\left[\varepsilon\|\nabla_v g\|^2_K + \frac{\gamma}{\varepsilon}\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|^2_K\right]. \end{aligned} \tag{4.15}$$

So

$$\begin{aligned} &2\sigma(\nabla_v g, R(g))_{Q_L} + \lambda\sigma\sum_{K\in\mathcal{C}_h}\sum_{e\in\mathcal{E}_h\cap\partial\tau_v}\|r_e(g)\|^2_K \\ &\geq \sigma\sum_{K\in\mathcal{C}_h}\left[-\varepsilon\|\nabla_v g\|^2_K + (\lambda - \frac{\gamma}{\varepsilon})\sum_{e\in\mathcal{E}_v\cap\partial\tau_v}\|r_e(g)\|^2_K\right]. \end{aligned} \tag{4.16}$$

By inverse estimate and assumptions on σ and δ_K, we obtain

$$\sum_{K\in\mathcal{C}_h}\sigma\delta_K(\Delta_v g, g_x+v\cdot\nabla_\perp g)_{Q_L}\leq\frac{1}{2}\left(\sigma\|\nabla_v g\|^2_{Q_L}+\sum_{K\in\mathcal{C}_h}\delta_K\|g_x+v\cdot\nabla_\perp g\|^2_K\right). \tag{4.17}$$

Now taking $\alpha=\min[\frac{1}{2}-\varepsilon,\lambda-\frac{\gamma}{\varepsilon}]$, which is positive for $\frac{\gamma}{\lambda}<\varepsilon<\frac{1}{2}$, we conclude the desired result. $\square$

Corollary 4.1. *For B_t defined as in* (4.10) *we have the coercivity*

$$B_*(g,g)\geq\alpha|||g|||^2_*,\qquad\forall g\in V_h, \tag{4.18}$$

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

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, for $\eta=f-\tilde{f}^h$ the error term can be written as

$$e:=f-f^h=(f-\tilde{f}^h)-(f^h-\tilde{f}^h)\equiv\eta-\xi. \tag{4.19}$$

Lemma 4.2. *There exists a constant C independent of the mesh size h such that for δ_K chosen as in* (3.31) *we have the following estimates*

$$\begin{aligned}A_\delta(\eta,\xi)\leq&\frac{1}{8}|||\xi|||^2_{A_\delta}+C\sum_{K\in\mathcal{C}_h}(\delta_K^{-1}\|\eta\|_K+\delta_K\|\nabla\eta\|_K)\\&+\sum_{K\in\mathcal{C}_h}|[\eta]|_{\partial K_-(\tilde{\beta})'}+|\eta|_{\Gamma_+}+|\eta|_0+|\eta|_M,\end{aligned} \tag{4.20}$$

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

Proof. The proof is similar to that of Theorem 3.1. Here, we need to control some additional jump and boundary terms. We have, using the definition of A_δ, that

$$\begin{aligned}A_\delta(\eta,\xi)=&\sum_{K\in\mathcal{C}_h}(\eta_x+v\cdot\nabla_\perp\eta,\xi+\delta_K(\xi_x+v\cdot\nabla_\perp\xi))_K\\&+\langle\eta_+,\xi_+\rangle_0+\sum_{K\in\mathcal{C}_h}\int_{\partial K_-(\tilde{\beta})'}[\eta]\xi_+|\tilde{\beta}\cdot\mathbf{n}|.\end{aligned} \tag{4.21}$$

Integrating by parts we have

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

Inserting (4.22) in (4.21) and applying Cauchy-schwarz inequality we obtain

$$\begin{aligned}A_\delta(\eta,\xi)\leq&\frac{1}{8}|||\xi|||^2_{A_\delta}+C\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|^2_K+\delta_K\|\eta_x+v\cdot\nabla_\perp\eta\|^2_K\right)\\&+|\eta|^2_0+|\eta|^2_M+\sum_{K\in\mathcal{C}_h}|[\eta]|^2_{\partial K_-(\tilde{\beta})'}+|\eta|^2_{I\times\partial\Omega_+}\\\leq&\frac{1}{8}|||\xi|||^2_{A_\delta}+C\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|^2_K+\delta_K\|\nabla\eta\|^2_K\right)\\&+\sum_{K\in\mathcal{C}_h}|[\eta]|^2_{\partial K_-(\tilde{\beta})'}+|\eta|^2_{\Gamma_+}+|\eta|^2_0+|\eta|^2_M.\end{aligned} \tag{4.23}$$

For D_δ we have by definition,

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

Using similar techniques as in the proof of Theorem 3.1, we need only to estimate the terms T_2, T_3 and T_4. Since η is continuous, from the definition of operators R and r_e we deduce that $T_3 = T_4 = 0$. It remains to estimate the term T_2. To this end we use (4.4) and (4.5) to obtain

$$|T_2| \le \sum_{K\in\mathcal{C}_h}\sigma\|\nabla_v\eta\|_K\|R(\xi)\|_K \le \sum_{K\in\mathcal{C}_h}\left(C\sigma\|\nabla_v\eta\|_K^2 + \frac{\sigma}{C_1}\|R(\xi)\|_K^2\right). \tag{4.24}$$

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

$$D_\delta(\eta,\xi) \le \frac{1}{8}|||\xi|||_{A_\delta}^2 + \frac{1}{8}|||\xi|||_{D_\delta}^2 + C\sigma\|\nabla_v\eta\|_{Q_L}^2, \tag{4.25}$$

and we conclude the proof. □

In what follows we shall use the following lemma (see [7]),

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*

$$\sum_{K\in\mathcal{C}_h}\int_{I_m\times\tau_{x_\perp}}\int_{\partial\tau_v} w\frac{\partial u}{\partial\mathbf{n}_v} = \sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\sum_{e\in\mathcal{E}_v}\int_e [[w]]\mathbf{n}_v.(\nabla_v u)^0. \tag{4.26}$$

Theorem 4.2 (Convergence Theorem). *Suppose $f^h \in V^h$ and f are the solutions of (4.9) and (2.1) respectively, then there exists a constant C independent of the mesh size h such that we have the following error estimate*

$$|||f - f^h|||_* \le Ch^{k+1/2}\|f\|_{k+1,Q_L}. \tag{4.27}$$

Proof. Using Corollary 4.1 and (4.19), we have

$$\alpha|||\xi|||_*^2 \le B_*(\xi,\xi) = B_*(\eta - e,\xi) = B_*(\eta,\xi) - B_*(e,\xi). \tag{4.28}$$

For the term $B_*(e,\xi)$ we have

$$B_*(e,\xi) = A(e,\xi) + D(e,\xi). \tag{4.29}$$

Since

$$D(e,\xi) = D(f,\xi) - D(f^h,\xi), \tag{4.30}$$

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

$$\begin{aligned}
D(f,\xi) = &\sum_{K\in\mathcal{C}_h}\int_{I_m\times\tau_{x_\perp}}\int_{\tau_v}\sigma\nabla_v f\nabla_v\xi \\
&- \sigma\sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\sum_{e\in\mathcal{E}_v}\int_e [[\xi]]\mathbf{n}_v.(\nabla_v f)^0 \\
= &\sum_{K\in\mathcal{C}_h}\int_K -\sigma(\Delta_v f)\xi + \sigma\sum_{K\in\mathcal{C}_h}\int_{I_m\times\tau_{x_\perp}}\int_{\partial\tau_v}\xi\frac{\partial f}{\partial\mathbf{n}_v} \\
&- \sigma\sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\sum_{e\in\mathcal{E}_v}\int_e [[\xi]]\mathbf{n}_v.(\nabla_v f)^0 \\
= &\sum_{K\in\mathcal{C}_h}\int_K -\sigma(\Delta_v f)\xi,
\end{aligned} \tag{4.31}$$

where in the last equality we have used Lemma 4.3. So the problem (4.9) is fully consistent and $B_*(e,\xi)=0$. Further, we get from (4.28) that

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

We have now using Lemma 4.2, multiplicative trace inequality (3.36) and the local interpolation error estimates (3.14)-(3.16),

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

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) in (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 (4.27) is a consequence of (4.35), (4.36) and the triangle inequality. □

4.2. **Penalty method.** In this subsection we present a variant of the scheme introduced in the previous subsection for the diffusive part of (4.10). As mentioned in [15], this approach presents some computational advantages, as it reduces the number of integrals to be computed when building the elementary matrices. On the negative side, very large coefficients might be introduced in the matrices since this scheme is a penalty method. Here we replace D in (4.10) by $\hat{D}$ and consider the following problem: Find $f^h \in V_h$ such that

$$\hat{B}_*(f^h,g) = \langle f_0, g_+\rangle_0, \qquad \forall g \in V_h, \tag{4.37}$$

where

$$\hat{B}_*(f,g) = A(f,g) + \hat{D}(f,g), \tag{4.38}$$

and

$$\begin{aligned}\hat{D}(f,g) = {} & \sigma(\nabla_v f, \nabla_v g)_{Q_L} + \sigma \sum_{e\in\mathcal{E}_v} \varrho(h_e)(r_e(f), r_e(g))_{Q_L} \\ & - \sigma h(\Delta_v f, g_x + v\cdot\nabla_\perp g)_{Q_L},\end{aligned} \tag{4.39}$$

with $\varrho(h_e)$ a positive constant which tends to $+\infty$ as h_e, the length of the edge e, tends to zero. For future use, we choose

$$\varrho(h_e) = \frac{1}{h_e^{2k+1}}, \tag{4.40}$$

where k is the order of polynomial used in the approximation. We define the diffusion norm

$$|||g|||_{\hat{D}}^2 = \frac{1}{2}\left[\sigma\|\nabla_v g\|_{Q_L}^2 + 2\sigma\varrho(h_e)\sum_{e\in\mathcal{E}_v}\|r_e(g)\|_{Q_L}^2\right]. \tag{4.41}$$

and estimate the error in the following norm

$$[||g||]_*^2 = |||g|||_A^2 + |||g|||_{\hat{D}}^2. \tag{4.42}$$

For this scheme we derive coercivity and stability estimates similar to Lemmas 4.1 and 4.2.

Lemma 4.4. *We have,*

$$A(g,g) + \hat{D}(g,g) \ge \alpha[||g||]_*^2, \qquad \forall g \in V_h. \tag{4.43}$$

Proof. The proof is similar to that of Lemma 4.1. □

Lemma 4.5. *There is a constant $C > 0$ independent of the mesh size h such that*

$$\hat{D}(\eta, \xi) \leq \frac{1}{8}[||\xi||]_*^2 + Ch\|\nabla_v \eta\|_{Q_L}^2. \tag{4.44}$$

Proof. Since $r_e(\eta) = 0$, using the definition of $\hat{D}$ we obtain

$$\hat{D}(\eta, \xi) = \sigma(\nabla_v \eta, \nabla_v \xi)_{Q_L} - \sigma h(\Delta_v \eta, \xi_x + v \cdot \nabla_\perp \xi)_{Q_L}. \tag{4.45}$$

Now by a similar argument as in the proof of Lemma 4.2 we obtain the desired result. □

We shall also use the following result due to Brezzi et al. [15].

Proposition 4.1. *Let $u|_K \in L^2(I_m, \tau_{x_\perp}, H^2(\tau_v))$ for all $K := I_m \times \tau_{x_\perp} \times \tau_v$ and $v_h \in V_h$. Then*

$$\begin{aligned} &\sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \sum_{e \in \mathcal{E}_v} \int_e [[v_h]]\mathbf{n}_v.(\nabla_v u)^0 \\ &\leq C|||v_h|||_{\hat{D}} \left(\sum_{K \in \mathcal{C}_h} \int_{I_m \times \tau_{x_\perp}} \sum_{e \subset \partial\tau_v} \varrho(h_e)^{-1} \|u\|_{H^2(\tau_v)}^2 \right)^{1/2}. \end{aligned} \tag{4.46}$$

Proof. It can be proved by the same argument as in the proof of Lemma 4 in [15]. □

Theorem 4.3. *Let f^h and f be the solutions of* (4.37) *and* (2.1)*, respectively. Then, the following estimate holds true*

$$[||f - f^h||]_* \leq Ch^{k+1/2}\|f\|_{k+1,Q_L}. \tag{4.47}$$

Proof. Proceeding exactly as in the proof of Theorem 4.2 and Lemmas 4.1 and 4.4 yields

$$\alpha[||\xi||]_*^2 \leq \hat{B}_*(\xi, \xi) = \hat{B}_*(\eta - e, \xi) = \hat{B}_*(\eta, \xi) - \hat{B}_*(e, \xi). \tag{4.48}$$

For the term $\hat{B}_*(e, \xi)$ we have

$$\hat{B}_*(e, \xi) = A(e, \xi) + \hat{D}(e, \xi). \tag{4.49}$$

But

$$\hat{D}(e, \xi) = \hat{D}(f, \xi) - \hat{D}(f^h, \xi). \tag{4.50}$$

Using (4.39), integrating by parts and applying Lemma 4.3 we have that

$$\begin{aligned} \hat{D}(f, \xi) &= \sum_{K \in \mathcal{C}_h} \int_{I_m \times \tau_{x_\perp}} \int_{\tau_v} \sigma \nabla_v f \cdot \nabla_v \xi \\ &= \sum_{K \in \mathcal{C}_h} \int_K -\sigma(\Delta_v f)\xi + \sum_{K \in \mathcal{C}_h} \int_{I_m \times \tau_{x_\perp}} \int_{\partial\tau_v} (\sigma\xi)\frac{\partial f}{\partial \mathbf{n}_v} \\ &= \sum_{K \in \mathcal{C}_h} \int_K -\sigma(\Delta_v f)\xi + \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \sum_{e \in \mathcal{E}_v} \int_e \sigma[[\xi]]\mathbf{n}_v \cdot (\nabla_v f)^0. \end{aligned} \tag{4.51}$$

So the scheme (4.37) is not consistent and we have to estimate the last term above. Using Lemma 4.1, we get

$$\begin{aligned} \hat{B}_*(e, \xi) &= \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \sum_{e \in \mathcal{E}_v} \int_e \sigma[[\xi]]\mathbf{n}_v.(\nabla_v f)^0 \\ &\leq C|||\xi|||_{\hat{D}} \left(\sum_{K \in \mathcal{C}_h} \int_{I_m \times \tau_{x_\perp}} \sum_{e \subset \partial\tau_v} \varrho(h_e)^{-1} \|f\|_{H^2(\tau_v)}^2 \right)^{1/2}, \end{aligned} \tag{4.52}$$

and inserting inequality (4.52) into (4.48), we end up with

$$\begin{aligned}\alpha[||\xi||]_*^2 &\le \hat{B}_*(\eta,\xi) - \hat{B}_*(e,\xi)\\ &\le A(\eta,\xi) + \hat{D}(\eta,\xi)\\ &+ C|||\xi|||_{\hat{D}}\left(\sum_{K\in\mathcal{C}_h}\int_{I_m\times\tau_{x_\perp}}\sum_{e\subset\partial\tau_v}\varrho(h_e)^{-1}\|f\|^2_{H^2(\tau_v)}\right)^{1/2}.\end{aligned} \tag{4.53}$$

Applying Lemmas 4.2 and 4.5, and choosing $\varrho(h_e)$ as in (4.40) we get

$$[||\xi||]_t^2 \le Ch^{2k+1}\|f\|^2_{k+1,Q_L}. \tag{4.54}$$

On the other hand, to estimate the interpolation error, we note that due to the fact that $r_e(\eta)=0$, (4.41) yields

$$|||\eta|||^2_{\hat{D}} = \frac{1}{2}\sigma\|\nabla_v\eta\|^2_{Q_L}. \tag{4.55}$$

Hence, applying interpolation error estimates (3.14)-(3.16) and assumption on σ, we get

$$\begin{aligned}[||\eta||]_*^2 &= |||\eta|||_A^2 + |||\eta|||^2_{\hat{D}}\\ &\le C(h^{-1}\|\eta\|^2_{Q_L} + h\|\nabla\eta\|^2_{Q_L} + \sigma\|\nabla_v\eta\|^2_{Q_L})\\ &\le Ch^{2k+1}\|f\|^2_{k+1,Q_L}.\end{aligned} \tag{4.56}$$

Then (4.47) is a consequence of (4.54), (4.56) and the triangle inequality. □

4.3. ***hp*-Discontinuous Galerkin method.** The aim of this section is to establish error bounds, using discontinuous Galerkin method. We shall employ the approach in [27], and derive error bound that is optimal in both h and p. We assume that the family of partitions $\{\mathcal{C}_h\}$ is shape regular in the sense of (3.30) and that every $K\in\mathcal{C}_h$ is affine equivalent to unit hypercube in $\mathbb{R}^5$. Let us first consider the bilinear form

$$\tilde{D}_\delta = D_\delta(f,g) + D_s(f,g), \tag{4.57}$$

where D_δ is as in (4.8) and the stabilizer D_s is defined by

$$D_s(f,g) = \sigma\sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\int_{\mathcal{E}_v}\gamma(h_e)[[f]][[g]]. \tag{4.58}$$

Here $\gamma(h_e)$ is the discontinuity scaling function and the precise choice of it will be discussed later. We now introduce the bilinear form

$$\tilde{B}_\delta = A_\delta + \tilde{D}_\delta. \tag{4.59}$$

The hp-DG for Fermi equation (2.1) is: find $f^h\in V_h^p$ such that

$$\tilde{B}_\delta(f^h,g) = \langle f_0, g_+\rangle_0 \qquad \forall g\in V_h^p. \tag{4.60}$$

Again, as in Subsection 3.1.2, we use V_h^p to emphasize the polynomials degree $p:=k$ in (4.1). We note that when the discontinuity scaling function $\gamma(h_e)$ is set to zero and the SD-parameter δ_K is considered to be h for all $K\in\mathcal{C}_h$, then the hp-DG (4.60) is identical to the method introduced in (4.9). Throughout the paper we shall assume that the solution f to the Fermi 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 $g\in V_h^p$. It causes the Galerkin orthogonality $\tilde{B}_\delta(f-f^h,g)=0$ to be hold for all $g\in V_h^p$. We shall derive the stability of the method (4.60) in the following norm

$$|||g|||^2_{\gamma,\delta} = |||g|||^2_{A_\delta} + |||g|||^2_{D_\delta} + \sigma\sum_{I_m\times\tau_{x_\perp}}\int_{I_m\times\tau_{x_\perp}}\int_{\mathcal{E}_v}\gamma(h_e)[[g]]^2. \tag{4.61}$$

Lemma 4.6. *There is a constant $C > 0$ such that*

$$\tilde{B}_\delta(g, g) \geq C|||g|||^2_{\gamma,\delta}, \qquad \forall g \in V_h^p. \tag{4.62}$$

Proof. By (4.59) and (4.57), we have

$$\tilde{B}_\delta(g, g) = A_\delta(g, g) + D_\delta(g, g) + D_s(g, g), \tag{4.63}$$

we notice that

$$D_s(g, g) = \sigma \sum_{I_m \times \tau_{x_\perp}} \int_{I_m \times \tau_{x_\perp}} \int_{\mathcal{E}_v} \gamma(h_e)[[g]]^2, \tag{4.64}$$

Inserting (4.64) in (4.63), and using Lemma 4.1, we obtain the desired result. □

Before continuing with the a priori error analysis of the *hp*-DG method (3.28), we state an approximation result for the finite element space V_h^p. We consider $Q_k(K)$, the set of all polynomials of degree at most k in each variable on K.

Lemma 4.7. *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*

$$\|g - \mathrm{P}g\|_{L^2(\partial K)} \leq C \left(\frac{h_K}{p+1}\right)^{\mu - \frac{1}{2}} \|g\|_{\mu,K}, \tag{4.65}$$

where $C > 0$ is a constant independent of h_K and p, and $\mathrm{P} : L^2(K) \to Q_p(K)$ is the usual L^2- projector of degree p on K.

Proof. see, e.g. [19]. □

We denote by P_v the univariate elementwise $L^2(\tau_v)$-projector 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 4.7. Actually for an element $K \in \mathcal{C}_h$ we have

$$\|f - \mathrm{P}_v f\|_{L^2(I_m, \tau_{x_\perp}, \partial\tau_v)} \leq C \left(\frac{h_K}{p+1}\right)^{\mu - \frac{1}{2}} \|f\|_{L^2(I_m, \tau_{x_\perp}, H^\mu(\tau_v))}. \tag{4.66}$$

where $K := I_m \times \tau_{x_\perp} \times \tau_v$. We also recall a restatement of Lemma 3.4: Suppose

$$f \in L^2(I, \Omega_{x_\perp}, H_0^1(\Omega_v)) \cap L^2(I, \Omega_{x_\perp}, H^2(\Omega_v)), \tag{4.67}$$

and

$$f|_K \in H^s(K), \qquad \forall K \in \mathcal{C}_h, \tag{4.68}$$

with $s \geq 2$. 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 (cf. [25, Theorem 4.72]). Then, by local interpolation error estimates (3.34), with $r = 1$, we have that

$$\|f - \Pi_p f\|_{1,K} \leq C \frac{h_K^{\mu-1}}{p^{s-1}} \|f\|_{s,K}, \tag{4.69}$$

with $\mu = \min(p + 1, s)$.

Theorem 4.4. *Suppose that for each $h_e \in \mathcal{E}_v$ the scaling discontinuity function is defined by*

$$\gamma(h_e) = \frac{p^2}{h_e}, \tag{4.70}$$

and the SD-parameter satisfies (3.31). *Let further the exact solution f of* (2.1) *to satisfy the assumptions* (4.67)-(4.68). *Then, there is a constant $C > 0$ independent* *of h and p such that the following hp-error bound holds true*

$$
\begin{aligned}
|||f-f^h|||^2_{\gamma,\delta} \le & C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2\mu-1}}\|f\|^2_{\mu,K}\\
&+\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\frac{1}{p^2}+\frac{1}{p}+\sigma h_K^{-1}+\delta_K h_K^{-1}+\frac{h_K}{p^2\delta_K}\right)\|f\|^2_{s,K},
\end{aligned}
\tag{4.71}
$$

with $\mu=\min(p+1,s)$.

Proof. The structure of the proof is the same as that of Theorem 4.2, except now we decompose the global error as

$$
e:=f-f^h=(f-\tilde f^h)+(\tilde f^h-f^h)\equiv\eta+\xi, \tag{4.72}
$$

where $\tilde f^h\in V_h^p$ is hp-interpolant of f satisfying (4.69), i.e. $\tilde f^h:=\Pi_p f$. By virtue of Lemma 4.6, we have

$$
C_I|||\xi|||_{\gamma,\delta}\le\tilde B_\delta(\xi,\xi)=\tilde B_\delta(e-\eta,\xi)=\tilde B_\delta(-\eta,\xi), \tag{4.73}
$$

where we have used the Galerkin orthogonality property $\tilde B_\delta(e,\xi)=0$ which follows form (4.60) with $g=\xi$ and the definition of boundary value problem, given the assumed smoothness of f. Thus, we deduce that

$$
C_I|||\xi|||_{\gamma,\delta}\le|\tilde B_\delta(\eta,\xi)|\le|A_\delta(\eta,\xi)|+|\tilde D_\delta(\eta,\xi)|. \tag{4.74}
$$

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

$$
[[\eta]]=0 \qquad \text{on } \mathcal{E}_v, \tag{4.75}
$$

and also

$$
\begin{aligned}
R(\eta)&=0 &&\text{on } \Omega,\\
r_e(\eta)&=0 &&\text{on } \Omega, \qquad \forall e\in\mathcal{E}_v.
\end{aligned}
\tag{4.76}
$$

Hence,

$$
|\tilde D_\delta(\eta,\xi)|\le I+II+III, \tag{4.77}
$$

where

$$
\begin{aligned}
I=\sigma|(\nabla_v\eta,\nabla_v\xi)_{Q_L}|, \qquad & II=\sigma|(\nabla_v\eta,R(\xi))_{Q_L}|,\\
III=\sum_{K\in\mathcal{C}_h}\sigma\delta_K|(\Delta_v\eta,\xi_x+v\cdot\nabla_\perp\xi)_K|. &
\end{aligned}
\tag{4.78}
$$

The term I can be estimated similarly as in the proof of Lemma 4.2. For the term II, using the definition of orthogonal projector to $L^2(Q_L)$, we obtain

$$
\begin{aligned}
\sigma(\nabla_v\eta,R(\xi))_{Q_L}&=\sigma(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L}+\sigma(\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L}\\
&=\sigma(\nabla_v\eta-\mathrm{P}_v\nabla_v\eta,R(\xi))_{Q_L}+\sigma(\nabla_v\eta,R(\xi))_{Q_L}=T_1+T_2.
\end{aligned}
\tag{4.79}
$$

For the term T_1, by the definition of the operator R and the shape regularity of $\mathcal{C}_h$ to relate h_e to h_K we obtain

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

where, in the first inequality, we used the notation

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

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

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

Hence,

$$\begin{aligned} T_1 \leq & C\sigma\|\sqrt{\gamma}[[\xi]]\|_{\mathcal{E}_v} \\ & \left(\sum_{I_m\times\tau_{x_\perp}} \sum_{\tau_v\in T_h^v} p^{-2}h_K\|\nabla_v f - \mathrm{P}_v\nabla_v f\|^2_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)}\right)^{\frac{1}{2}}. \end{aligned} \tag{4.80}$$

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

$$T_2 \leq \sqrt{\sigma}\|(\nabla_v\eta)\|_{Q_L}(\sigma\sum_{e\in\mathcal{E}_v}\|r_e(\xi)\|^2_{Q_L})^{\frac{1}{2}}. \tag{4.81}$$

It remains to estimate the term III. Following the proof of Theorem 3.2, i.e. applying inverse inequality and assumption (3.31), we get

$$\sigma\delta_K|(\Delta_v\eta, \xi_x + v\cdot\nabla_\perp\xi)_K| \leq \sqrt{\sigma\delta_K}\|\nabla_v\eta\|_K\|\xi_x + v\cdot\nabla_\perp\xi\|_K. \tag{4.82}$$

Substituting T_1 and T_2 into (4.79) and then inserting (4.78) into (4.77), by Cauchy-Schwarz inequality, we obtain

$$\begin{aligned} |\tilde{D}_\delta(\eta,\xi)| \leq C_1|||\xi|||^2_{\gamma,\delta} + C\sigma\Bigg(& \|(\nabla_v\eta)\|^2_{Q_L} \\ & + \sum_{I_m\times\tau_{x_\perp}} \sum_{\tau_v\in T_h^v} p^{-2}h_K\|\nabla_v f - \mathrm{P}_v\nabla_v f\|^2_{L^2(I_m,\tau_{x_\perp},\partial\tau_v)}\Bigg), \end{aligned} \tag{4.83}$$

where $C_1 \leq \frac{1}{3}C_I$. For the term $|A_\delta(\eta,\xi)|$, using Lemma 4.2 we have

$$|A_\delta(\eta,\xi)| \leq C_2|||\xi|||^2_{\gamma,\delta} + C\sum_{K\in\mathcal{C}_h}\left(\delta_K^{-1}\|\eta\|^2_K + \delta_K\|\eta_x + v.\nabla_\perp\eta\|^2_K + \|\eta\|^2_{\partial K}\right), \tag{4.84}$$

where $C_2 \leq \frac{1}{3}C_I$. Substituting the estimates (4.83) and (4.84) into (4.74), using the standard kick back argument and applying the approximation error estimate (4.69) and (4.66) and trace inequality (3.36) we deduce that

$$|||\xi|||^2_{\gamma,\delta} \leq C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\delta_K h_K^{-1} + \sigma h_K^{-1} + \frac{h_K}{p^2\delta_K}\right)\|f\|^2_{s,K} + \frac{h_K^{2\mu-1}}{p^{2\mu-1}}\|f\|^2_{\mu,K}. \tag{4.85}$$

Similarly, due to (4.75) and (4.76) for the interpolation error we get

$$|||\eta|||^2_{\gamma,\delta} \leq C\sum_{K\in\mathcal{C}_h}\left(h_K^{-1}\|\eta\|^2_K + \sigma\|\nabla_v\eta\|^2_K + \delta_K\|\eta_x + v.\nabla_\perp\eta\|^2_K\right). \tag{4.86}$$

Hence, using (4.69) and trace inequality (3.36) we get

$$|||\eta|||^2_{\gamma,\delta} \leq C\sum_{K\in\mathcal{C}_h}\frac{h_K^{2\mu-1}}{p^{2s-2}}\left(\delta_K h_K^{-1} + \frac{1}{p^2} + \frac{1}{p} + \sigma h_K^{-1}\right)\|f\|^2_{s,K}. \tag{4.87}$$

Now inserting the resulting bound on $|||\eta|||_{\gamma,\delta}$ and (4.85) in (4.74) we obtain the desired result. □

Remark 4.1. Suppose in Theorem 4.4 that $2 \le s \le p+1$ and the streamline diffusion parameter is chosen as $\delta_K = \frac{h_K^2}{\sigma C_I^2 p^4}$ for all $K \in \mathcal{C}_h$, then if we assume $\mathcal{O}(\frac{\sigma}{h_K}) \sim 1$ for all $K \in \mathcal{C}_h$, we deduce from Theorem 4.4 that the discretization error, in the norm $|||.|||_{\gamma,\delta}$, converges like $\mathcal{O}(\frac{h^{(\mu-\frac{1}{2})}}{p^{(\mu-1)}})$. We see that the error bound is optimal in both h and p. The parameter δ_K may be selected as

$$\delta_K = C_\delta \frac{h_K}{p}, \qquad \forall K \in \mathcal{C}_h, \tag{4.88}$$

where the constant C_δ is chosen subject to the constraint on δ_K in Theorem 4.4. In this case the parameter $\delta_K h_K^{-1}$ in (4.71) is equal to $\frac{1}{p}$, and the error of the method measured in DG-norm is of order $\mathcal{O}(\frac{h^{(\mu-\frac{1}{2})}}{p^{(\mu-\frac{1}{2})}})$. We note that in this case the error bound (4.71) is again simultaneously optimal in h and p.

Remark 4.2. The choices for δ_K made in Remark 4.1, end to optimal error bounds, simultaneously, in h and p. These choices 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. Using The later ones the suboptimal stabilization terms in the method (4.60) would vanish.

Conclusion: Our analysis extend the result of [3] to a three dimensional degenerate type convection-dominated convection-diffusion problem with a small and variable diffusion coefficient. We have presented an h- and hp-a priori error analysis of both SD- and DG- schemes for Fermi Pencil equation. We have shown that the schemes are optimally convergent with respect to the mesh size h and the degree p of approximating polynomial. This estimates are sharp in the sense that omitting any power of the diffusion coefficient on the left hand side of our stability norms will cause the same amount of reduced convergence rate. In our error analysis the availability of continuous interpolant have played a crucial role.

References

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

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

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

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

[5] M. Asadzadeh, *On the stability of characteristic schemes for the Fermi equation*, Appl. Comput. Math. 1 (2002), no. 2, 158-174.

[6] M. Asadzadeh and E. W. Larsen, *Linear transport equations in flatland with small angular diffusion and their finite element approximations*, Math. Comput. Modelling, 47 (2008), no. 3-4, 495-514.

[7] 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.

[8] M. Asadzadeh and A. Sopasakis, *On fully discrete schemes for the Fermi pencil-beam equation*, Comput. Methods Appl. Mech. Engrg. 191 (2002), no. 41-42, 4641-4659.

[9] 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.

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

[11] C. Börgers and E. W. Larsen, *The transversely integrated scalar density of a narrowly focused particle beam*, SIAM journal on Applied Mathematics, 55 (1995), no. 1, 1-22.

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

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

[14] F. Brezzi, J. Douglas and L. D. Marini, *Two families of mixed finite elements for second order elliptic problems*, Numer. Math., 47 (1985), no. 2, 217-235.
[15] 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.
[16] P. G. Ciarlet, *The Finite Element Method for Elliptic Problems*, North-Holland, Amsterdam, 1978.
[17] P. Houston, C. Schwab and E. Süli, *Stabilized hp-finite element methods for first order hyperbolic problems*, SIAM J. Numer. Anal. 37 (2000), 1618-1643.
[18] P. Houston and E. Süli, *Stabilized hp-finite element approximation of partial differential equations with nonnegative characteristic form*, Computing 66 (2001), 99-119.
[19] 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.
[20] 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.
[21] J. Jaffré, C. Johnson and A. Szepessy, *Convergence of the discontinuous Galerkin finite element method for hyperbolic conservation laws*, Math. Models Methods Appl. Sci. 5 (1995), no. 3, 367-386.
[22] C. Johnson and J. Saranen, *Streamline diffusion methods for the incompressible Euler and Navier-Stokes equations*, Math. Comp. 47 (1986), no. 175, 1-18.
[23] G. C. Pomraning, *The Focker-Planck operator as an asymptotic limit*, M3AS. 2 (1992), no. 1, 21-36.
[24] H. Risken, *The Focker-Planck Equation, Methods of solution and applications*, Springer-Verlag, Berlin, Heildlberg, New York, (1989).
[25] C. Schwab, *p- and hp-finite element methods, Theory and Applications in Solid and Fluid Mechanics*, Oxford (1998).
[26] L. R. Scott and S. Zhang, *Finite element interpolation of non-smooth functions satisfying boundary conditions*, Math. Comp. 54, (1990), 484–493.
[27] 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 address: mohammad@chalmers.se

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