# ICES REPORT 11-33

## November 2011

# Robust DPG Method for Convection-Dominated Diffusion Problems

**by**

Leszek Demkowicz and Norbert Heuer

**The Institute for Computational Engineering and Sciences**
The University of Texas at Austin
Austin, Texas 78712

*Reference: Leszek Demkowicz and Norbert Heuer, Robust DPG Method for Convection-Dominated Diffusion Problems, ICES REPORT 11-33, The Institute for Computational Engineering and Sciences, The University of Texas at Austin, November 2011.*

in context of wave propagation problems was undertaken in [7, 13] and has recently resulted in a "pollution free" DPG method for the Helmholtz equation [6]. The so-called quasi-optimal test norm (analyzed in this paper) for the confusion problem was recently investigated numerically by Niemi, Bollier and Calo [11,12]. The concept of optimal testing has also been recently pursued by Dahmen, Schwab, Cohen, Welper and Huang in [1, 2].

Given a test norm, the Petrov-Galerkin method with optimal test functions [5] delivers the best approximation error in the corresponding energy norm,

$$\|u - u_{hp}\|_E = \inf_{w_{hp} \in U_{hp}} \|u - w_{hp}\|_E \tag{1.1}$$

where the energy (residual) norm is implied by the very problem we solve, i.e. the bilinear form $b(u, v)$ and the choice of the test norm $\|v\|_V$,

$$\|u\|_E = \sup_{v \in V} \frac{|b(u, v)|}{\|v\|_V}. \tag{1.2}$$

The optimal test function $\hat{e}_i$ corresponding to a trial function $e_i \in U_{hp}$, is determined by solving an auxiliary variational problem dictated by the choice of the test norm,

$$\hat{e}_i \in V : \quad (\hat{e}_i, \delta v)_V = b(e_i, \delta v) \quad \forall \delta v \in V. \tag{1.3}$$

Variational problem (1.3) can rarely be solved exactly and it is solved approximately using the standard, Bubnov-Galerkin method and an "enriched" subspace of $V$. If the corresponding error is negligible, the method delivers its promise.

Obviously, different choices of the test norm lead to different energy norms, and different versions of the method. So, the natural question to ask is,

*What test norm to use ?*

The purpose of this paper is an attempt to answer the question for an important singular perturbation model problem – convection-dominated diffusion, hereafter called the "confusion problem". In fact, the proposed general strategy applies to any perturbation problem involving a parameter, singular or not.

Given a domain $\Omega$ covered with a finite element mesh, we introduce the DPG bilinear form corresponding to the confusion problem [4, 8],

$$b((u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_n), (v, \boldsymbol{\tau})) = \left(\boldsymbol{\sigma}, \epsilon^{-1}\boldsymbol{\tau} + \nabla v\right)_{\Omega_h} + (u, \nabla\cdot \boldsymbol{\tau} - \boldsymbol{\beta} \cdot \nabla v)_{\Omega_h} - \langle [\boldsymbol{\tau} \cdot \boldsymbol{n}], \hat{u} \rangle - \langle \hat{\sigma}_n, [v] \rangle$$

where $\hat{\sigma}_n := \big((\boldsymbol{\sigma} - \boldsymbol{\beta} u) \cdot \boldsymbol{n}\big)|_{\Gamma_h}$ with $\Gamma_h$ being the mesh skeleton and $\boldsymbol{n}$ being a unit normal vector on $\Gamma_h$ in a certain direction (exterior to $\Omega$ on $\Gamma$). Furthermore, $[\cdot]$ denotes the jump across $\Gamma_h$ (reducing to the trace on $\Gamma$).

The functional setting is now well understood. Field variables come from $L^2(\Omega)$ (possibly weighted), the trace $\hat{u}$ comes from the trace of the space $H^1_0(\Omega)$ to $\Gamma_h$, and the flux (or rather another trace) $\hat{\sigma}_n$ lives in the trace of the space $\boldsymbol{H}(\mathrm{div}, \Omega)$ to the mesh skeleton $\Gamma_h$. Test functions come from "broken" Sobolev spaces, $\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, \Omega_h)$, $v \in H^1(\Omega_h)$. $L^2$-inner product notation $(\cdot, \cdot)_{\Omega_h}$ emphasizes that operators of gradient and divergence are to be understood element-wise. Finally, $\langle \cdot, \cdot \rangle$ denotes the duality pairing between the trace spaces $H^{1/2}$ and $H^{-1/2}$ defined over $\Gamma_h$. The trace spaces are equipped with minimum energy extension norms inherited from $H^1(\Omega)$ and $\boldsymbol{H}(\mathrm{div}, \Omega)$.

While we set the spaces, we do not fix at this moment any particular norms.

**Robustness.** Intuitively speaking, we say that a discretization method applied to a singular perturbation problem is *robust*, if the approximate solution does not change much as we vary the perturbation parameter. A precise meaning of robustness is more difficult as we have to decide on which components of the solution we want to look at, and what norms to use. In context of the confusion problem, we typically look at the "density" $u$, and expect its approximate counterpart to vary a "little", as we decrease $\epsilon$. The "little" is typically quantified in terms of the $L^2$-norm of $u$ that is insensitive to steepening of the boundary layer as $\epsilon \to 0$. It is already much less clear whether the same should be expected for the "stress" $\boldsymbol{\sigma} = \epsilon \nabla u$.

Mathematically, we consider the discretization method to be robust, if the stability constant relating actual approximation error and the best approximation error is *independent* of the perturbation parameter. Ideally, the two errors should be measured in the same norm. Most of the time, the perturbation parameter does not "disappear" but is hidden in the definition of the norm.

We outline now our general strategy for designing the test norm for the confusion problem.

**Step 1: Decide what you want.** Notice that the DPG method is automatically robust in the energy[2] norm. Indeed, it delivers the *best approximation error* in the energy norm, so the stability constant is simply equal to one. We are a bit more demanding though, as we want the robustness in *the norm of our choice*. As mentioned above, in context of the confusion problem, we want at least the $L^2$-robustness for $u$. This implies that we need to have

$$\|u\| \lesssim \|(u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_n)\|_E. \tag{1.4}$$

Here, $a \lesssim b$ indicates existence of a constant $C$, of order one (independent of $\epsilon$), such that $a \leq Cb$. When applied to the FE error, the inequality above allows for bounding the $L^2$ error in the field variable $u$, by the best approximation error of *all involved* unknowns and, ultimately, possibly leading to a robust version of the method. Certainly, if the constant $C$ blows up with $\epsilon \to 0$, our chances for a robust method are lost. So we can view assumption (1.4) at least as a necessary condition for designing the test norm.

Given a function $u \in L^2(\Omega)$, we now select special test functions $v, \boldsymbol{\tau}$ that solve the following variational problem,

$$(v, \boldsymbol{\tau}) \in V = H_0^1(\Omega) \times \boldsymbol{H}(\mathrm{div}, \Omega) : \qquad \begin{aligned} \epsilon^{-1}\boldsymbol{\tau} + \nabla v &= 0, \\ \nabla \cdot \boldsymbol{\tau} - \boldsymbol{\beta} \cdot \nabla v &= g. \end{aligned} \tag{1.5}$$

[2] For any choice of the test norm.

Selecting $g = u$, we have (for any test norm)

$$\begin{aligned}\|u\|^2 = b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau})) &= \frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|(v,\boldsymbol{\tau})\|_V}\|(v,\boldsymbol{\tau})\|_V \\ &\leq \sup_{(v,\tau)} \frac{|b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))|}{\|(v,\boldsymbol{\tau})\|_V}\ \|(v,\boldsymbol{\tau})\|_V = \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_E\ \|(v,\boldsymbol{\tau})\|_V. \end{aligned} \tag{1.6}$$

An obvious sufficient condition to request is that the solution to the adjoint problem (1.5) is bounded by $\|g\|$ uniformly in $\epsilon$,

$$\|(v,\boldsymbol{\tau})\|_V \lesssim \|g\|. \tag{1.7}$$

Dividing then (1.6) sidewise by $\|u\|$, we get the inequality (1.4).

The first step in estimating the DPG error has thus been reduced to the stability estimate for a classical, strong version of the adjoint equation. Notice that the assumption on global conformity of the test functions "killed" the trace terms.

**Step 2: Study the stability of the adjoint problem.** We will establish the following stability estimates,

$$\|v\|, \|\boldsymbol{\beta}\cdot\nabla v\|_\phi, \sqrt{\epsilon}\|\nabla v\|, \frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}\|_\phi, \frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\tau}\|, \|\nabla\cdot\boldsymbol{\tau}\|_{\phi+\epsilon} \lesssim \|g\|$$

where $\phi = O(1)$ is a weight function to be introduced later, and

$$\|\cdot\|_\phi := \|\sqrt{\phi}\ \cdot\|.$$

These are now our "Lego blocks" with which we can build different test norms. One obvious choice is the *quasi-optimal test norm* (cf. [13]),

$$\|(v,\boldsymbol{\tau})\|^2_{V,qopt} = \|v\|^2 + \|\underbrace{\frac{1}{\epsilon}\boldsymbol{\tau} + \nabla v}_{=\boldsymbol{f}}\|^2 + \|\underbrace{\nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v}_{=g}\|^2.$$

Another choice, following the original version of the method used in [8], is a *weighted test norm*,

$$\|(v,\boldsymbol{\tau})\|^2_{V,1} = \epsilon\|v\|^2 + \epsilon\|\nabla v\|^2 + \|\boldsymbol{\beta}\cdot\nabla v\|^2_{\phi+\epsilon} + \|\boldsymbol{\tau}\|^2_{\phi+\epsilon} + \|\nabla\cdot\boldsymbol{\tau}\|^2_{\phi+\epsilon}.$$

It turns out that, for both test norms, we also have a robust estimate for the stress:

$$\|\boldsymbol{\sigma}\| \lesssim \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_E.$$

One major difference between the two test norms is their effect on optimal shape functions. Element-wise inversion of the Riesz operator corresponding to the quasi-optimal test norm results in optimal test functions with boundary layers of width equal to the element Peclet number. As we will see, effective resolution of such test functions requires the use of special techniques.

The weighted test norm involves the parameter $\epsilon$ as well. Notice, however, that the small diffusion in the cross-wind derivative of $v$ is balanced with a small reaction term for $v$ (even though we can afford the "full" $L^2$ norm of $v$). Similarly, the weight in the norms of $\boldsymbol{\tau}$ and $\nabla\cdot\boldsymbol{\tau}$ is the same. This produces optimal test functions with no boundary layers, whose resolution is straightforward. Also, components $v$ and $\boldsymbol{\tau}$ are decoupled which makes the inversion of the Riesz operator more efficient.

**Step 3: Estimate the best approximation error in the energy norm.** Once the test norm has been fixed, we can proceed with the estimate of the best approximation error in the energy norm. The choice of the test norm implies now specific norms for $\boldsymbol{\sigma}$ and $u$.

For the quasi-optimal test norm, we will show that one can select norms for traces and fluxes in such a way that we have a two sided estimate:

$$(\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|\hat{u}\|_?^2 + \|\hat{\sigma}_n\|_?^2)^{\frac{1}{2}} \lesssim \|(u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_n)\|_E \lesssim (\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|\hat{u}\|_?^2 + \|\hat{\sigma}_n\|_?^2)^{\frac{1}{2}}$$

A similar result has recently been established for another singular perturbation problem – the Helmholtz equation [7].

Thus, for the quasi-optimal norm, the best approximation error for both field variables is measured in the $L^2$-norm. For the weighted norm, however, the best approximation error estimate involves terms

$$\frac{1}{\epsilon}\|\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)},\ \|u\|_{1/(\phi+\epsilon)}$$

Clearly, we will need many more d.o.f. to control these terms. This is the price we pay for eliminating the boundary layers from the optimal test functions.

Definition of the appropriate norms for the traces and fluxes and the corresponding best approximation error estimates will be explained later.

The rest of this paper is organized as follows. In the next section we formulate the model problem, define specific spaces and norms, recall the DPG method, define several test norms and present the main results (Theorem 1), namely error estimates for the DPG method and any of the test norms. In Section 2.1 we provide stability estimates for the adjoint problem. They are essential for the proofs of the main results which are given in Section 2.2. Numerical experiments with two of the analyzed test norms and are reported in Section 3. We end with some conclusions.

## 2 Model problem and robustness of the DPG method

First, let us recall the precise setting of the convection-dominated diffusion problem, its DPG discretization, and the involved spaces and norms. We introduce several test norms and formulate the main results, namely error estimates for the resulting (abstract) DPG approximations. In Section 2.1 we analyze the stability of the adjoint problem and in Section 2.2 we prove norm equivalences that imply our main results, Theorem 1 below.

**Model problem.** The convection-dominated diffusion ("confusion") problem is

$$\begin{aligned} -\epsilon\Delta u + \nabla\cdot(\boldsymbol{\beta} u) &= f \qquad \text{in} \quad \Omega, \\ u &= 0 \qquad \text{on} \quad \Gamma. \end{aligned} \tag{2.8}$$

Here, $\Omega \subset \mathbb{R}^3$ is a bounded simply connected domain with piecewise smooth boundary $\Gamma := \partial\Omega$. The two-dimensional case can be dealt with analogously with corresponding error estimates. In fact, in Section 3 we report on numerical experiments for a 2D problem.

Introducing $\boldsymbol{\sigma} := \epsilon\nabla u$, (2.8) turns into the first order system

$$\begin{aligned} \epsilon^{-1}\boldsymbol{\sigma} - \nabla u &= 0 \qquad \text{in} \quad \Omega, \\ -\nabla\cdot\boldsymbol{\sigma} + \nabla\cdot(\boldsymbol{\beta} u) &= f \qquad \text{in} \quad \Omega, \\ u &= 0 \qquad \text{on} \quad \Gamma. \end{aligned} \tag{2.9}$$

Let $\Omega_h$ be a non-overlapping partitioning of $\Omega$ into open elements with Lipschitz boundary, $\bar{\Omega} = \cup\{\bar{K};\ K \in \Omega_h\}$, and let us recall the following notation. The mesh skeleton is $\Gamma_h = \cup\{\partial K;\ K \in \Omega_h\}$, $\boldsymbol{n}$ denotes a unit normal vector on $\Gamma_h$, $\hat{\sigma}_n = \big((\boldsymbol{\sigma} - \boldsymbol{\beta} u)\cdot\boldsymbol{n}\big)|_{\Gamma_h}$, and $\hat{u}$ is the trace of $u$ on $\Gamma_h$.

Now, testing (2.9) with appropriate functions $\boldsymbol{\tau}$ and $v$, and integrating by parts on elements, leads to the ultra-weak formulation: *Find $u \in L^2(\Omega)$, $\boldsymbol{\sigma} \in \boldsymbol{L}^2(\Omega)$, $\hat{u} \in H^{1/2}_{00}(\Gamma_h)$, and $\hat{\sigma}_n \in H^{-1/2}(\Gamma_h)$ such that*

$$\begin{aligned} \big(\epsilon^{-1}\boldsymbol{\sigma}, \boldsymbol{\tau}\big)_{\Omega_h} + (u, \nabla\cdot\boldsymbol{\tau})_{\Omega_h} - \langle[\boldsymbol{\tau}\cdot\boldsymbol{n}], \hat{u}\rangle &= 0 \qquad && \forall\boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div}, \Omega_h), \\ (\boldsymbol{\sigma}, \nabla v)_{\Omega_h} - (u, \boldsymbol{\beta}\cdot\nabla v)_{\Omega_h} - \langle\hat{\sigma}_n, [v]\rangle &= (f, v)_\Omega \qquad && \forall v \in H^1(\Omega_h). \end{aligned} \tag{2.10}$$

The left-hand side of system (2.10) is the DPG bilinear form:

$$b((u, \boldsymbol{\sigma}, \hat{u}, \hat{\sigma}_n), (v, \boldsymbol{\tau})) := \big(\boldsymbol{\sigma}, \epsilon^{-1}\boldsymbol{\tau} + \nabla v\big)_{\Omega_h} + (u, \nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v)_{\Omega_h} - \langle[\boldsymbol{\tau}\cdot\boldsymbol{n}], \hat{u}\rangle - \langle\hat{\sigma}_n, [v]\rangle\,. \tag{2.11}$$

Having injectivity of the dual problem under standard assumptions on $\boldsymbol{\beta}$, existence and uniqueness of the solution of (2.10) is guaranteed by the inf-sup condition. This inf-sup condition is immediate for the "optimal test norm" (defined by duality via $b(\cdot,\cdot)$). Under certain conditions on $\boldsymbol{\beta}$, we show equivalence of this "norm" to other test norms so that the optimal test norm is indeed a norm.

Let us recall definitions of some spaces. We use standard (scalar) $L^2(\Omega)$, (vector) $\boldsymbol{L}^2(\Omega)$ and $H^1(\Omega)$ spaces with $L^2, \boldsymbol{L}^2(\Omega)$- and $H^1(\Omega)$-norms denoted by $\|\cdot\|$ and $\|\cdot\|_{1,\Omega}$, respectively. We have the broken spaces

$$\begin{aligned} H^1(\Omega_h) &:= \{v \in L^2(\Omega);\ v|_K \in H^1(K)\ \forall K \in \Omega_h\}, \\ H^1_0(\Omega_h) &:= \{v \in H^1(\Omega_h);\ v|_\Gamma = 0\}, \\ \boldsymbol{H}(\mathrm{div}, \Omega_h) &:= \{\boldsymbol{\tau} \in \boldsymbol{L}^2(\Omega);\ \boldsymbol{\tau}|_K \in \boldsymbol{H}(\mathrm{div}, K)\ \forall K \in \Omega_h\} \end{aligned}$$

with respective product norms, and

$$H^{1/2}_{00}(\Gamma_h) := H^1_0(\Omega)|_{\Gamma_h}, \tag{2.12}$$

$$H^{-1/2}(\Gamma_h) := \{\eta;\ \exists \boldsymbol{\tau} \in \boldsymbol{H}(\mathrm{div},\Omega):\ (\boldsymbol{\tau}\cdot\boldsymbol{n})|_{\Gamma_h} = \eta\} \tag{2.13}$$

with "generic" (canonical) trace norms

$$\|v\|_{1/2,\Gamma_h} := \inf_{w\in H^1_0(\Omega),\ w|_{\Gamma_h}=v} \|w\|_{1,\Omega}, \tag{2.14}$$

$$\|\eta\|_{-1/2,\Gamma_h} := \inf_{\boldsymbol{\tau}\in \boldsymbol{H}(\mathrm{div},\Omega),\ (\boldsymbol{\tau}\cdot\boldsymbol{n})|_{\Gamma_h}=\eta} \|\boldsymbol{\tau}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}. \tag{2.15}$$

Note that, by definition, any $v \in H^{1/2}_{00}(\Gamma_h)$ vanishes on $\Gamma$. Below, we need the restriction onto $\Gamma^0_h := \Gamma_h \cap \Omega$ of the latter norm, denoted by $\|\cdot\|_{-1/2,\Gamma^0_h}$.

To analyze the jumps of $(\boldsymbol{\tau}, v) \in \boldsymbol{H}(\mathrm{div},\Omega_h)\times H^1(\Omega_h)$ across $\Gamma_h$ we define the norms of the dual spaces of $H^{1/2}_{00}(\Gamma_h)$ and $H^{-1/2}(\Gamma_h)$:

$$\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma^0_h} := \sup_{w\in H^{1/2}_{00}(\Gamma_h)} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}], w\rangle}{\|w\|_{1/2,\Gamma_h}} = \sup_{w\in H^1_0(\Omega)} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}], w\rangle}{\|w\|_{1,\Omega}}, \tag{2.16}$$

$$\|\,[v]\,\|_{\Gamma_h} := \sup_{\eta\in H^{-1/2}(\Gamma_h)} \frac{\langle \eta, [v]\rangle}{\|\eta\|_{-1/2,\Gamma_h}} = \sup_{\boldsymbol{\eta}\in \boldsymbol{H}(\mathrm{div},\Omega)} \frac{\langle \boldsymbol{\eta}\cdot\boldsymbol{n}, [v]\rangle}{\|\boldsymbol{\eta}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}}. \tag{2.17}$$

**DPG method.** As discussed in the introduction, the DPG method for the confusion problem consists in

1. selecting a norm with inner product $(\cdot,\cdot)_V$ in $V = \boldsymbol{H}(\mathrm{div},\Omega_h)\times H^1(\Omega_h)$,
2. selecting a piecewise polynomial space induced by the mesh $\Omega_h$ and chosen polynomial degrees, $U_{hp} \subset L^2(\Omega)\times \boldsymbol{L}^2(\Omega)\times H^{1/2}_{00}(\Gamma_h)\times H^{-1/2}(\Gamma_h)$, with basis $\{e_i,\ i = 1,\ldots,\mathrm{dof}\}$,
3. determining test functions $\hat{e}_i \in V$ $(i = 1,\ldots,\mathrm{dof})$ such that

$$(\hat{e}_i, \delta v)_V = b(e_i, \delta v) \quad \forall \delta v \in V,$$

4. and calculating the DPG approximation $(u_{hp}, \boldsymbol{\sigma}_{hp}, \hat{u}_{hp}, \hat{\sigma}_{n,hp}) \in U_{hp}$ defined by

$$b((u_{hp}, \boldsymbol{\sigma}_{hp}, \hat{u}_{hp}, \hat{\sigma}_{n,hp}), \hat{e}_i) = (f, \hat{e}_i)_\Omega, \quad i = 1,\ldots,\mathrm{dof}.$$

**Remark 1** *In order to obtain a practical method, it is essential to work with a localizable norm in $V$. In this way, the calculation of test functions are local problems, defined on individual elements. Note that, by considering standard boundary norms for $\hat{u}$ and $\hat{\sigma}_n$, norms for $[\boldsymbol{\tau}\cdot\boldsymbol{n}]$ and $[v]$ on interfaces defined by duality with respect to the bilinear form $b(\cdot,\cdot)$ (2.11) are not local a priorily. Therefore, critical for the DPG method is to measure the errors of $\hat{u}$ and $\hat{\sigma}_n$ in appropriate norms so that the corresponding $V$-norm is equivalent to a localizable one.*

**Main results.** Recall that the DPG method delivers the best approximation in the energy norm defined by the chosen norm in $V$, cf. (1.1), (1.2). Therefore, proving norm relations in $U := L^2(\Omega) \times \boldsymbol{L}^2(\Omega) \times H^{1/2}_{00}(\Gamma_h) \times H^{-1/2}(\Gamma_h)$

$$\|\cdot\|_{U_1} \lesssim \|\cdot\|_E \lesssim \|\cdot\|_{U_2} \tag{2.18}$$

which are robust in $\epsilon$, one immediately obtains the robust estimates

$$\begin{aligned}\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n) - (u_{hp},\boldsymbol{\sigma}_{hp},\hat{u}_{hp},\hat{\sigma}_{n,hp})\|_{U_1} &\lesssim \inf_{w_{hp}\in U_{hp}} \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n) - w_{hp}\|_E &(2.19)\\ &\lesssim \inf_{w_{hp}\in U_{hp}} \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n) - w_{hp}\|_{U_2}. &(2.20)\end{aligned}$$

To establish these estimates and to introduce the test norms we need some more notation and have to make some assumptions. We define

$$\begin{aligned}\Gamma_- &:= \{\boldsymbol{x} \in \partial\Omega;\ \boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n} < 0\} \quad \text{(inflow)},\\ \Gamma_+ &:= \{\boldsymbol{x} \in \partial\Omega;\ \boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n} > 0\} \quad \text{(outflow)},\\ \Gamma_0 &:= \{\boldsymbol{x} \in \partial\Omega;\ \boldsymbol{\beta}(\boldsymbol{x})\cdot\boldsymbol{n} = 0\}.\end{aligned}$$

Throughout, $\phi \in C^2(\bar{\Omega})$ will be a fixed weight function that satisfies

$$0 \le \phi \le 1 \quad \text{in} \quad \Omega, \qquad \phi = 0 \quad \text{on} \quad \Gamma_-.$$

We will need $L^2(\Omega)$-norms of components of vector fields in directions perpendicular to $\boldsymbol{\beta}$. To this end we formally extend, in any point of $\Omega$, $\boldsymbol{\beta}$ to an orthogonal basis $(\boldsymbol{\beta}, \boldsymbol{\beta}_1^\perp, \boldsymbol{\beta}_2^\perp)$ of $\mathbb{R}^3$ where all three vectors are of equal length. Then we use the generic notation

$$\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\eta}\|^2 := \|\boldsymbol{\beta}_1^\perp\cdot\boldsymbol{\eta}\|^2 + \|\boldsymbol{\beta}_2^\perp\cdot\boldsymbol{\eta}\|^2, \quad \boldsymbol{\eta} \in \boldsymbol{L}^2(\Omega).$$

For our analysis we need to make some assumptions. Throughout we will assume that $\epsilon > 0$ is small enough. There are several conditions $\boldsymbol{\beta}$ has to satisfy. Throughout we require that $\boldsymbol{\beta} \in \boldsymbol{C}^2(\bar{\Omega})$ and $\boldsymbol{\beta} = O(1)$, $\nabla\cdot\boldsymbol{\beta} = O(1)$. Furthermore, depending on the objective, we need one or more of the following properties.

(A1) $\quad \nabla\times\boldsymbol{\beta} = 0, \quad 0 < C \le |\boldsymbol{\beta}|^2 + \frac{1}{2}\nabla\cdot\boldsymbol{\beta}, \quad C = O(1)$

(A2) $\quad \boldsymbol{\Xi} := \nabla(\phi\boldsymbol{\beta}) + \nabla(\phi\boldsymbol{\beta})^T - \nabla\cdot(\phi\boldsymbol{\beta})\boldsymbol{I} = O(1)$

(A3) $\quad \nabla\cdot\boldsymbol{\beta} = 0$

We will analyze the DPG method for the following test norms:

$$\begin{aligned}\|(v,\boldsymbol{\tau})\|^2_{V,qopt} &:= \|v\|^2 + \|\epsilon^{-1}\boldsymbol{\tau} + \nabla v\|^2 + \|\nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v\|^2,\\ \|(v,\boldsymbol{\tau})\|^2_{V,0} &:= \epsilon\|v\|^2 + \epsilon\|\nabla v\|^2 + \|\boldsymbol{\beta}\cdot\nabla v\|^2_{\phi+\epsilon} + \frac{1}{\epsilon^2}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}\|^2_{\phi+\epsilon} + \frac{1}{\epsilon}\|\boldsymbol{\tau}\|^2 + \|\nabla\cdot\boldsymbol{\tau}\|^2_{\phi+\epsilon},\\ \|(v,\boldsymbol{\tau})\|^2_{V,1} &:= \epsilon\|v\|^2 + \epsilon\|\nabla v\|^2 + \|\boldsymbol{\beta}\cdot\nabla v\|^2_{\phi+\epsilon} + \|\boldsymbol{\tau}\|^2_{\phi+\epsilon} + \|\nabla\cdot\boldsymbol{\tau}\|^2_{\phi+\epsilon},\end{aligned}$$

with corresponding energy norms $\|\cdot\|_{E,qopt}$, $\|\cdot\|_{E,0}$ and $\|\cdot\|_{E,1}$, respectively. For the analysis with test norm $\|\cdot\|_{V,qopt}$ we need a specific norm measuring the jumps $[\boldsymbol{\tau}\cdot\boldsymbol{n}]$ and $[v]$ of $(v,\boldsymbol{\tau})\in V$,

$$\|(v,\boldsymbol{\tau})\|_{\Gamma_h} := \|\tilde{v}\|. \tag{2.21}$$

Here, $(v,\boldsymbol{\tau}) = (v_0,\boldsymbol{\tau}_0) + (\tilde{v},\tilde{\boldsymbol{\tau}})$ with $(v_0,\boldsymbol{\tau}_0)\in H_0^1(\Omega)\times\boldsymbol{H}(\mathrm{div},\Omega)$ being the solution of

$$\begin{aligned}\epsilon^{-1}\boldsymbol{\tau}_0 + \nabla v_0 &= \boldsymbol{f} := \epsilon^{-1}\boldsymbol{\tau} + \nabla v,\\ \nabla\cdot\boldsymbol{\tau}_0 - \boldsymbol{\beta}\cdot\nabla v_0 &= g := \nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v.\end{aligned}$$

Also, let us define the dual norm

$$\|(w,\eta)\|_{\Gamma_h'} := \sup_{(v,\boldsymbol{\tau})\in V\setminus\{0\}} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle + \langle\eta,[v]\rangle}{\|(v,\boldsymbol{\tau})\|_{\Gamma_h}} \qquad \forall (w,\eta)\in H_{00}^{1/2}(\Gamma_h)\times H^{-1/2}(\Gamma_h). \tag{2.22}$$

Our main results are as follows.

**Theorem 1**

(i) (DPG method with quasi-optimal test norm $\|\cdot\|_{V,qopt}$, $\|\cdot\|_E = \|\cdot\|_{E,qopt}$.)
*If $\boldsymbol{\beta}$ satisfies* (A1) *then there hold* (2.19) *and* (2.20) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_1} := \|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_2} := \Big(\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \|(\hat{u},\hat{\sigma}_n)\|_{\Gamma_h'}^2\Big)^{1/2}.$$

*If $\boldsymbol{\beta}$ satisfies* (A1) *and* (A3) *then there holds* (2.19) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_1} := \Big(\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \epsilon^2\|\hat{u}\|_{1/2,\Gamma_h}^2 + \epsilon\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2\Big)^{1/2}.$$

(ii) (DPG method with weighted test norm $\|\cdot\|_{V,0}$, $\|\cdot\|_E = \|\cdot\|_{E,0}$.)
*There holds* (2.20) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_2} := \Big(\|u\|_{1/(\phi+\epsilon)}^2 + \|\boldsymbol{\beta}\cdot\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)}^2 + \frac{1}{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\sigma}\|^2 + \frac{1}{\epsilon}\|\hat{u}\|_{1/2,\Gamma_h}^2 + \frac{1}{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2\Big)^{1/2}.$$

*If $\boldsymbol{\beta}$ satisfies* (A1), (A2) *and* (A3) *then there holds* (2.19) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_1} := \Big(\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \epsilon^2\|\hat{u}\|_{1/2,\Gamma_h}^2 + \epsilon\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2\Big)^{1/2}.$$

(iii) (DPG method with weighted test norm $\|\cdot\|_{V,1}$, $\|\cdot\|_E := \|\cdot\|_{E,1}$.)
*There holds* (2.20) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_2} := \Big(\|u\|_{1/(\phi+\epsilon)}^2 + \frac{1}{\epsilon^2}\|\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)}^2 + \frac{1}{\epsilon}\|\hat{u}\|_{1/2,\Gamma_h}^2 + \frac{1}{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2\Big)^{1/2}.$$

*If $\boldsymbol{\beta}$ satisfies* (A1), (A2) *and* (A3) *then there holds* (2.19) *with*

$$\|u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U_1} := \Big(\|u\|^2 + \|\boldsymbol{\sigma}\|^2 + \epsilon^2\|\hat{u}\|_{1/2,\Gamma_h}^2 + \epsilon\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2\Big)^{1/2}.$$

**Proof.** It is enough to prove the respective equivalence (2.18). These are given by Lemmas 6, 4 and 5, respectively, in Section 2.2. □

**Remark 2** *In Section* 3 *we will present numerical results for a two-dimensional model problem with quasi-optimal test norm and the weighted test norm $\|\cdot\|_{V,1}$, among others. Under the conditions of separating norms for $v$ and $\boldsymbol{\tau}$, and robustness for $\|u\|$, the theoretical result for the weighted norm $\|\cdot\|_{V,0}$ is the best we could achieve, i.e. the corresponding norm $\|\cdot\|_{U_2}$ for the error estimate in* (2.20) *deteriorates least when $\epsilon \to 0$. In particular, the upper bound resulting from $\|\cdot\|_{V,0}$ is stronger than the one corresponding to $\|\cdot\|_{V,1}$, cf. Theorem* 1(ii), (iii)*. From the point of view of practicality, however, the latter case is superior since its test functions in $\boldsymbol{\tau}$ do not suffer from boundary layers whereas the ones from the former case do.*

## 2.1 Stability of the adjoint problem

As discussed before, stability estimates for the adjoint problem are vital to prove robust $L^2$ error estimates for the field variables, cf. (1.7). The stability will be analyzed in this section.

Considering the bilinear form (2.11) one finds that the adjoint problem of (2.9) is

$$\epsilon^{-1}\boldsymbol{\tau} + \nabla v = \boldsymbol{f}, \tag{2.23}$$

$$\nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v = g, \tag{2.24}$$

in $\boldsymbol{L}^2(\Omega)$ and $L^2(\Omega)$, respectively. Furthermore, the function $v$ solves

$$-\epsilon\Delta v - \boldsymbol{\beta}\cdot\nabla v = g - \epsilon\nabla\cdot\boldsymbol{f} \quad \text{on } \Omega_h. \tag{2.25}$$

Here, $\nabla\cdot\boldsymbol{f} \in (H_0^1(\Omega))'$ by defining the divergence operator as the adjoint operator of

$$-\nabla : \ H_0^1(\Omega) \to \boldsymbol{L}^2(\Omega).$$

**Lemma 1** *Let $v \in H_0^1(\Omega)$ satisfy* (2.25)*. If $\boldsymbol{\beta}$ satisfies* (A1) *then there holds for sufficiently small $\epsilon > 0$*

$$\epsilon\,\|\nabla v\|^2 + \|v\|^2 \lesssim \|g\|^2 + \epsilon\|\boldsymbol{f}\|^2.$$

**Proof.** Since $\nabla\times\boldsymbol{\beta} = 0$ and $\Omega$ is simply connected, $\boldsymbol{\beta}$ has a scalar potential $\psi$, $\boldsymbol{\beta} = \nabla\psi$, and we adjust the constant such that $e^{\psi} = O(1)$. We consider the transformed function $w := e^{\psi}v$ and calculate

$$\begin{aligned}
\nabla w &= e^{\psi}(\nabla v + \boldsymbol{\beta}v),\\
e^{\psi}\nabla v &= \nabla w - \boldsymbol{\beta}w,\\
\Delta w &= e^{\psi}\boldsymbol{\beta}\cdot(\nabla v + \boldsymbol{\beta}v) + e^{\psi}(\Delta v + \nabla\cdot\boldsymbol{\beta}\,v + \boldsymbol{\beta}\cdot\nabla v)\\
&= e^{\psi}(\Delta v + \nabla\cdot\boldsymbol{\beta}\,v + |\boldsymbol{\beta}|^2\,v + 2\boldsymbol{\beta}\cdot\nabla v),\\
\boldsymbol{\beta}\cdot\nabla w &= e^{\psi}(\boldsymbol{\beta}\cdot\nabla v + |\boldsymbol{\beta}|^2\,v).
\end{aligned}$$

Using these relations and equation (2.25) it follows that

$$\begin{aligned}\epsilon\Delta w + \boldsymbol{\beta}\cdot\nabla w &= e^{\psi}\Big(\epsilon\Delta v + \epsilon\nabla\cdot\boldsymbol{\beta}\, v + \epsilon\,|\boldsymbol{\beta}|^2\, v + 2\epsilon\boldsymbol{\beta}\cdot\nabla v + \boldsymbol{\beta}\cdot\nabla v + |\boldsymbol{\beta}|^2\, v\Big)\\ &= e^{\psi}(\epsilon\nabla\cdot\boldsymbol{f} - g) + (1-\epsilon)\,|\boldsymbol{\beta}|^2\, w + \epsilon\nabla\cdot\boldsymbol{\beta}\, w + 2\epsilon\boldsymbol{\beta}\cdot\nabla w\end{aligned}$$

that is,

$$-\epsilon\Delta w - (1-2\epsilon)\boldsymbol{\beta}\cdot\nabla w + \Big((1-\epsilon)\,|\boldsymbol{\beta}|^2 + \epsilon\nabla\cdot\boldsymbol{\beta}\Big)w = e^{\psi}(g - \epsilon\nabla\cdot\boldsymbol{f}).$$

We multiply by $w$ and integrate to deduce that

$$\epsilon\int_\Omega |\nabla w|^2 - (1-2\epsilon)\int_\Omega \boldsymbol{\beta}\cdot\nabla w\, w + \int_\Omega\Big((1-\epsilon)\,|\boldsymbol{\beta}|^2 + \epsilon\nabla\cdot\boldsymbol{\beta}\Big)w^2 = \int_\Omega e^{\psi}(g-\epsilon\nabla\cdot\boldsymbol{f})w. \quad (2.26)$$

Since $w\in H^1_0(\Omega)$, integration by parts proves that

$$\int_\Omega \boldsymbol{\beta}\cdot\nabla w\, w = -\frac{1}{2}\int_\Omega \nabla\cdot\boldsymbol{\beta}\, w^2$$

so that (2.26) is equivalent to

$$\begin{aligned}\epsilon\int_\Omega |\nabla w|^2 + \int_\Omega\Big((1-\epsilon)\,|\boldsymbol{\beta}|^2 + \frac{1}{2}\nabla\cdot\boldsymbol{\beta}\Big)w^2 &= \int_\Omega e^{\psi}(g-\epsilon\nabla\cdot\boldsymbol{f})w\\ &= \int_\Omega e^{\psi} g w + \epsilon\int_\Omega e^{\psi}\boldsymbol{f}\cdot(\boldsymbol{\beta}w + \nabla w). \quad (2.27)\end{aligned}$$

Now, for $\epsilon>0$ sufficiently small, the assumptions about $\boldsymbol{\beta}$ and Young's inequality show that for $\delta>0$ there holds

$$\epsilon\,\|\nabla w\|^2 + \|w\|^2 \lesssim \frac{1}{\delta}\|g\|^2 + \delta\|w\|^2 + \frac{\epsilon}{\delta}\|\boldsymbol{f}\|^2 + \epsilon\delta\|w\|^2 + \frac{\epsilon}{\delta}\|\boldsymbol{f}\|^2 + \epsilon\delta\|\nabla w\|^2.$$

Selecting $\delta$ small enough this yields

$$\epsilon\,\|\nabla w\|^2 + \|w\|^2 \lesssim \|g\|^2 + \epsilon\|\boldsymbol{f}\|^2.$$

The assertion is proved by bounding

$$\|v\|^2 = \|e^{-\psi}w\|^2 \lesssim \|w\|^2 \lesssim \|g\|^2 + \epsilon\|\boldsymbol{f}\|^2$$

and

$$\epsilon\|\nabla v\|^2 = \epsilon\|e^{-\psi}(\nabla w - \boldsymbol{\beta}w)\|^2 \lesssim \epsilon\|\nabla w\|^2 + \epsilon\|w\|^2 \lesssim \|g\|^2 + \epsilon\|\boldsymbol{f}\|^2.$$

□

**Lemma 2** *Let $\boldsymbol{\beta}$ satisfy* (A1) *and* (A2)*. Then there holds*

$$\|\boldsymbol{\beta}\cdot\nabla v\|_\phi \lesssim \|g\|$$

*for any solution $v\in H^1_0(\Omega)$ of* (2.25) *with $\nabla\cdot\boldsymbol{f}=0$, and sufficiently small $\epsilon>0$.*

**Proof.** Let $v_\beta := \boldsymbol{\beta}\cdot\nabla v$. We multiply (2.25) by $\phi v_\beta$ and integrate over $\Omega$ to obtain

$$\|v_\beta\|_\phi^2 = -\int_\Omega \phi g v_\beta - \epsilon\int_\Omega \Delta v \phi v_\beta. \tag{2.28}$$

Integration by parts yields

$$-\int_\Omega \Delta v\phi v_\beta = \int_\Omega \nabla v\cdot\nabla(\phi\boldsymbol{\beta})\nabla v + \int_\Omega \nabla v\cdot\nabla\nabla v(\phi\boldsymbol{\beta}) - \int_{\partial\Omega}\phi\big(\boldsymbol{n}\cdot\nabla v\big)v_\beta. \tag{2.29}$$

Now

$$\int_\Omega \nabla v\cdot\nabla\nabla v(\phi\boldsymbol{\beta}) = \frac{1}{2}\int_\Omega \phi\boldsymbol{\beta}\cdot\nabla|\nabla v|^2 = -\frac{1}{2}\int_\Omega \nabla\cdot(\phi\boldsymbol{\beta})\,|\nabla v|^2 + \frac{1}{2}\int_{\partial\Omega}\phi\boldsymbol{\beta}\cdot\boldsymbol{n}\,|\nabla v|^2.$$

Using this relation for (2.29), substituting in (2.28) and using that $\phi = 0$ on $\Gamma_-$ and $\boldsymbol{\beta}\cdot\boldsymbol{n} = 0$ on $\Gamma_0$ yields

$$\|v_\beta\|_\phi^2 + \frac{\epsilon}{2}\int_{\Gamma_+}\phi\boldsymbol{\beta}\cdot\boldsymbol{n}\,|\nabla v|^2 = -\int_\Omega \phi g v_\beta + \frac{\epsilon}{2}\int_\Omega \nabla v\cdot\Big(2\nabla(\phi\boldsymbol{\beta}) - \nabla\cdot(\phi\boldsymbol{\beta})\boldsymbol{I}\Big)\nabla v.$$

Here, $\boldsymbol{I}$ denotes the identity tensor and we also used that, due to the homogeneous boundary condition for $v$, there holds $\nabla v = \big(\boldsymbol{n}\cdot\nabla v\big)\boldsymbol{n}$ on $\partial\Omega$.

Recalling the notation $\boldsymbol{\Xi} := \nabla(\phi\boldsymbol{\beta}) + \nabla(\phi\boldsymbol{\beta})^T - \nabla\cdot(\phi\boldsymbol{\beta})\boldsymbol{I}$, and using the fact that $\boldsymbol{\beta}\cdot\boldsymbol{n} > 0$ on $\Gamma_+$, we eventually obtain

$$\|v_\beta\|_\phi^2 \le -\int_\Omega \phi g v_\beta + \frac{\epsilon}{2}\int_\Omega \nabla v\cdot\boldsymbol{\Xi}\nabla v.$$

By assumption, $\boldsymbol{\Xi} = O(1)$. Therefore, Young's inequality and an application of Lemma 1 prove the assertion. □

**Lemma 3** *Let $\boldsymbol{\beta}$ satisfy* (A1) *and* (A3)*. Then there holds*

$$\|\nabla v\| = \frac{1}{\epsilon}\|\boldsymbol{\tau}\| \lesssim \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}$$

*for any solution $\boldsymbol{\tau}\in\boldsymbol{H}(\mathrm{div},\Omega_h)$, $v\in H^1(\Omega_h)$ of* (2.23), (2.24) *with $\boldsymbol{f} = 0$ and $g = 0$.*

**Proof.** We follow the proof of [4, Lemma 4.3] and use Lemma 1 rather than standard stability results.

There is $\boldsymbol{z}\in\boldsymbol{H}(\mathrm{curl},\Omega)$ and $\psi\in H_0^1(\Omega)$ such that

$$\boldsymbol{\tau} = (\epsilon\nabla\psi - \boldsymbol{\beta}\psi) + \nabla\times\boldsymbol{z} \tag{2.30}$$

is a stable decomposition of $\boldsymbol{\tau}$. In particular, $\psi$ can be chosen as the solution of

$$-\epsilon\Delta\psi + \nabla\cdot(\boldsymbol{\beta}\psi) = -\nabla\cdot\boldsymbol{\tau} \quad \text{in } \Omega.$$

Since $\nabla\cdot\boldsymbol{\beta}=0$, this equation is of the form (2.25) with opposite sign of $\boldsymbol{\beta}$ and $\boldsymbol{f}=\frac{1}{\epsilon}\boldsymbol{\tau}$. Again, since $\nabla\cdot\boldsymbol{\beta}=0$, assumption (A1) is invariant with respect to a sign change of $\boldsymbol{\beta}$ and Lemma 1 is applicable to the problem defining $\psi$. This gives

$$\epsilon\|\nabla\psi\|^2+\|\psi\|^2 \lesssim \frac{1}{\epsilon}\|\boldsymbol{\tau}\|^2, \tag{2.31}$$

and by the triangle inequality this estimate proves that

$$\|\nabla\times\boldsymbol{z}\| \le \|\epsilon\nabla\psi\|+\|\boldsymbol{\beta}\psi\|+\|\boldsymbol{\tau}\| \lesssim \frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\tau}\|. \tag{2.32}$$

Now, using the decomposition of $\boldsymbol{\tau}$, equation (2.23), integrating by parts on $\Omega_h$ and using (2.24), we obtain

$$\begin{aligned}\|\boldsymbol{\tau}\|^2 &= (\boldsymbol{\tau},\epsilon\nabla\psi-\boldsymbol{\beta}\psi+\nabla\times\boldsymbol{z})_{\Omega_h} = (\boldsymbol{\tau},\epsilon\nabla\psi)_{\Omega_h}+(\epsilon\nabla v,\boldsymbol{\beta}\psi)_{\Omega_h}-(\epsilon\nabla v,\nabla\times\boldsymbol{z})_{\Omega_h}\\ &= -\epsilon\,(\nabla\cdot\boldsymbol{\tau},\psi)_{\Omega_h}+\epsilon\,\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\psi\rangle+\epsilon\,(\nabla\cdot(\boldsymbol{\beta}v),\psi)_{\Omega_h}-\epsilon\,\langle\boldsymbol{n}\cdot\nabla\times\boldsymbol{z},[v]\rangle\\ &= \epsilon\,\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\psi\rangle-\epsilon\,\langle\boldsymbol{n}\cdot\nabla\times\boldsymbol{z},[v]\rangle \lesssim \epsilon\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0}\|\psi\|_{1,\Omega}+\epsilon\|\nabla\times\boldsymbol{z}\|\|\,[v]\,\|_{\Gamma_h}.\end{aligned} \tag{2.33}$$

In the last step we made use of the definition of the dual boundary norms (2.16), (2.17). An application of (2.31) and (2.32) yields

$$\|\boldsymbol{\tau}\| \lesssim \|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0}+\sqrt{\epsilon}\|\,[v]\,\|_{\Gamma_h}.$$

Since $\nabla v=\frac{1}{\epsilon}\boldsymbol{\tau}$ by (2.23) this finishes the proof. □

## 2.2 Energy norm equivalences.

For the proof of Theorem 1 we need several norm equivalences in the solution space

$$U := L^2(\Omega)\times\boldsymbol{L}^2(\Omega)\times H^{1/2}_{00}(\Gamma_h)\times H^{-1/2}(\Gamma_h).$$

They are proved in this section. Recall that, for a given norm in the test space

$$V = H^1(\Omega_h)\times\boldsymbol{H}(\mathrm{div},\Omega_h),$$

the energy norm $\|\cdot\|_E$ is defined by duality with respect to the bilinear form $b(\cdot,\cdot)$, cf. (1.2). Throughout we will use that the $V$-norm is recovered by duality from $\|\cdot\|_E$ (see [13, (2.11)]):

$$\|(v,\boldsymbol{\tau})\|_V = \sup_{(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\in U\setminus\{0\}} \frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_E} \quad \forall(v,\boldsymbol{\tau})\in V. \tag{2.34}$$

Here and in the following lemmas, and misusing notation, $(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)$ denotes any element of the solution space $U$, and not necessarily the solution of (2.10). Also, without further indication, $(v,\boldsymbol{\tau})\in V$ will always be related to $\boldsymbol{f}\in\boldsymbol{L}^2(\Omega)$ and $g\in L^2(\Omega)$ through the adjoint problem (2.23), (2.24).

The equivalences needed for the three parts of Theorem 1 are proved following identical lines. We start proving the relations needed for Theorem 1(ii) (Lemma 4) and re-use estimates for the proofs of (iii) and (i), in this order (Lemmas 5, 6).

**Lemma 4** *Let* $(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\in U$ *be arbitrary. There holds*

$$\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,0}\lesssim \|u\|_{1/(\phi+\epsilon)}+\|\boldsymbol{\beta}\cdot\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)}+\frac{1}{\sqrt{\epsilon}}\Big(\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\sigma}\|+\|\hat{u}\|_{1/2,\Gamma_h}+\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}\Big). \quad (2.35)$$

*Furthermore, if* $\boldsymbol{\beta}$ *satisfies* (A1), (A2) *and* (A3) *then there holds*

$$\|u\|+\|\boldsymbol{\sigma}\|+\epsilon\|\hat{u}\|_{1/2,\Gamma_h}+\sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}\lesssim\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,0}. \quad (2.36)$$

**Proof.** We first prove (2.35). Let us denote

$$\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U,0}^2:=\|u\|_{1/(\phi+\epsilon)}^2+\|\boldsymbol{\beta}\cdot\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)}^2+\frac{1}{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\sigma}\|^2+\frac{1}{\epsilon}\|\hat{u}\|_{1/2,\Gamma_h}^2+\frac{1}{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}^2.$$

Using (2.34) and the definition of the bilinear form (2.11), we see that the relation

$$\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,0}\lesssim\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U,0}$$

is equivalent to

$$\begin{aligned}\|(v,\boldsymbol{\tau})\|_{V,0}&\gtrsim\sup_{(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\neq 0}\frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{U,0}} && (2.37)\\ &\simeq\|g\|_{\phi+\epsilon}+\|\boldsymbol{\beta}\cdot\boldsymbol{f}\|_{\phi+\epsilon}+\sqrt{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{f}\|+\sqrt{\epsilon}\sup_{(\hat{u},\hat{\sigma}_n)\neq 0}\frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\hat{u}\rangle+\langle\hat{\sigma}_n,[v]\rangle}{\|\hat{u}\|_{1/2,\Gamma_h}+\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}}.\end{aligned}$$

The bound

$$\|g\|_{\phi+\epsilon}+\|\boldsymbol{\beta}\cdot\boldsymbol{f}\|_{\phi+\epsilon}+\sqrt{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{f}\|\lesssim\|(v,\boldsymbol{\tau})\|_{V,0}$$

is immediate by the triangle inequality and the definition of the norms. To estimate the supremum we integrate by parts to represent the boundary dualities and to bound:

$$\begin{aligned}&\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle+\langle(\boldsymbol{\eta}-\boldsymbol{\beta}w)\cdot\boldsymbol{n},[v]\rangle\\ &=(\boldsymbol{\eta},\nabla v)_{\Omega_h}+(\nabla\cdot\boldsymbol{\eta},v)_{\Omega_h}-(w,\boldsymbol{\beta}\cdot\nabla v)_{\Omega_h}-(\nabla\cdot(\boldsymbol{\beta}w),v)_{\Omega_h}+(w,\nabla\cdot\boldsymbol{\tau})_{\Omega_h}+(\nabla w,\boldsymbol{\tau})_{\Omega_h}\\ &\lesssim\Big(\sqrt{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\nabla v\|+\|\boldsymbol{\beta}\cdot\nabla v\|_{\phi+\epsilon}+\sqrt{\epsilon}\|v\|+\|\nabla\cdot\boldsymbol{\tau}\|_{\phi+\epsilon}+\frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}\|_{\phi+\epsilon}+\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\tau}\|\Big)\\ &\quad\Big(\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\eta}\|+\|\boldsymbol{\beta}\cdot\boldsymbol{\eta}\|_{1/(\phi+\epsilon)}+\frac{1}{\sqrt{\epsilon}}\|\nabla\cdot\boldsymbol{\eta}\|\\ &\qquad+\|w\|_{1/(\phi+\epsilon)}+\frac{1}{\sqrt{\epsilon}}\|w\|+\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\beta}\cdot\nabla w\|+\epsilon\|\boldsymbol{\beta}\cdot\nabla w\|_{1/(\phi+\epsilon)}+\sqrt{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\nabla w\|\Big)\\ &\lesssim\|(v,\boldsymbol{\tau})\|_{V,0}\frac{1}{\sqrt{\epsilon}}\left(\|\boldsymbol{\eta}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}+\|w\|_{1,\Omega}\right) \qquad (2.38)\end{aligned}$$

for any $w\in H_0^1(\Omega)$, $\boldsymbol{\eta}\in\boldsymbol{H}(\mathrm{div},\Omega)$. Here, we used that $|1/(\phi+\epsilon)|\le 1/\epsilon$ and $\nabla\cdot\boldsymbol{\beta}=O(1)$.

Now, for given $\rho \in H^{1/2}_{00}(\Gamma_h)$ and $\varphi \in H^{-1/2}(\Gamma_h)$ there are $w \in H^1_0(\Omega)$ and $\tilde{\boldsymbol{\eta}} \in \boldsymbol{H}(\mathrm{div}, \Omega)$ such that $w|_{\Gamma_h^0} = \rho$ and $(\tilde{\boldsymbol{\eta}} \cdot \boldsymbol{n})|_{\Gamma_h} = \varphi$. Defining $\boldsymbol{\eta} := \tilde{\boldsymbol{\eta}} + \boldsymbol{\beta} \cdot w$ and using the definition of the trace norms (2.14), (2.15), we obtain

$$\sup_{(\rho,\varphi)\neq 0} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\rho\rangle + \langle\varphi,[v]\rangle}{\|\rho\|_{1/2,\Gamma_h} + \|\varphi\|_{-1/2,\Gamma_h}} = \sup_{(w,\boldsymbol{\eta})\neq 0} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle + \langle(\boldsymbol{\eta}-\boldsymbol{\beta}w)\cdot\boldsymbol{n},[v]\rangle}{\|w\|_{1,\Omega} + \|\tilde{\boldsymbol{\eta}}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}}. \tag{2.39}$$

Estimate (2.38) and the triangle inequality yield

$$\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle + \langle(\boldsymbol{\eta}-\boldsymbol{\beta}w)\cdot\boldsymbol{n},[v]\rangle \lesssim \|(v,\boldsymbol{\tau})\|_{V,0}\frac{1}{\sqrt{\epsilon}}\Big(\|\tilde{\boldsymbol{\eta}}\|_{\boldsymbol{H}(\mathrm{div},\Omega)} + \|w\|_{1,\Omega}\Big)$$

so that

$$\sup_{(w,\boldsymbol{\eta})\neq 0} \frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle + \langle(\boldsymbol{\eta}-\boldsymbol{\beta}w)\cdot\boldsymbol{n},[v]\rangle}{\|w\|_{1,\Omega} + \|\tilde{\boldsymbol{\eta}}\|_{\boldsymbol{H}(\mathrm{div},\Omega)}} \lesssim \frac{1}{\sqrt{\epsilon}}\|(v,\boldsymbol{\tau})\|_{V,0}.$$

Referring to (2.39) this finishes the proof of (2.37) and, therefore, of (2.35).

To prove (2.36) we note that, as before,

$$\|u\| + \|\boldsymbol{\sigma}\| + \epsilon\|\hat{u}\|_{1/2,\Gamma_h} + \sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h} \lesssim \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,0}$$

is equivalent to

$$\begin{aligned}
\|(v,\boldsymbol{\tau})\|_{V,0} &\lesssim \sup_{(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\neq 0} \frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|u\| + \|\boldsymbol{\sigma}\| + \epsilon\|\hat{u}\|_{1/2,\Gamma_h} + \sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}} \\
&\simeq \|g\| + \|\boldsymbol{f}\| + \frac{1}{\epsilon}\sup_{\hat{u}\neq 0}\frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\hat{u}\rangle}{\|\hat{u}\|_{1/2,\Gamma_h}} + \frac{1}{\sqrt{\epsilon}}\sup_{\hat{\sigma}_n\neq 0}\frac{\langle\hat{\sigma}_n,[v]\rangle}{\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}} \\
&= \|g\| + \|\boldsymbol{f}\| + \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}. 
\end{aligned} \tag{2.40}$$

Now, for given $(v,\boldsymbol{\tau}) \in H^1_0(\Omega_h) \times \boldsymbol{H}(\mathrm{div},\Omega_h)$ we decompose $(v,\boldsymbol{\tau}) = (v_1,\boldsymbol{\tau}_1) + (v_2,\boldsymbol{\tau}_2) + (v_0,\boldsymbol{\tau}_0)$ with $v_1, v_2 \in H^1_0(\Omega)$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2 \in \boldsymbol{H}(\mathrm{div},\Omega)$ and $(v_0,\boldsymbol{\tau}_0) = (v,\boldsymbol{\tau}) - (v_1,\boldsymbol{\tau}_1) - (v_2,\boldsymbol{\tau}_2)$. Here, $(v_1,\boldsymbol{\tau}_1)$ and $(v_2,\boldsymbol{\tau}_2)$ solve

$$\begin{aligned}
\epsilon^{-1}\boldsymbol{\tau}_1 + \nabla v_1 &= 0 && \text{in } \Omega,\\
\nabla\cdot\boldsymbol{\tau}_1 - \boldsymbol{\beta}\cdot\nabla v_1 &= g = \nabla\cdot\boldsymbol{\tau} - \boldsymbol{\beta}\cdot\nabla v && \text{in } \Omega
\end{aligned}$$

and

$$\begin{aligned}
\epsilon^{-1}\boldsymbol{\tau}_2 + \nabla v_2 &= \boldsymbol{f} = \epsilon^{-1}\boldsymbol{\tau} + \nabla v && \text{in } \Omega,\\
\nabla\cdot\boldsymbol{\tau}_2 - \boldsymbol{\beta}\cdot\nabla v_2 &= 0 && \text{in } \Omega,
\end{aligned}$$

respectively. Moreover, by construction,

$$\begin{aligned}
\epsilon^{-1}\boldsymbol{\tau}_0 + \nabla v_0 &= 0 \quad \text{in } \Omega_h,\\
\nabla\cdot\boldsymbol{\tau}_0 - \boldsymbol{\beta}\cdot\nabla v_0 &= 0 \quad \text{in } \Omega_h
\end{aligned}$$

and

$$[\boldsymbol{n}\cdot\boldsymbol{\tau}] = [\boldsymbol{n}\cdot\boldsymbol{\tau}_0] \quad \text{on } \Gamma_h^0, \qquad [v] = [v_0] \quad \text{on } \Gamma_h. \tag{2.41}$$

We finish the proof by bounding each of the three pairs.

(i) **Bound of $\|(v_1,\boldsymbol{\tau}_1)\|_{V,0}$.** Throughout we use the relations defining $(v_1,\boldsymbol{\tau}_1)$. Lemma 1 proves

$$\sqrt{\epsilon}\|\nabla v_1\| + \sqrt{\epsilon}\|v_1\| \lesssim \|g\|.$$

Again by Lemma 1, and Lemma 2:

$$\|\boldsymbol{\beta}\cdot\nabla v_1\|_{\phi+\epsilon} \lesssim \max\{\sqrt{\epsilon}\|\boldsymbol{\beta}\cdot\nabla v_1\|, \|\boldsymbol{\beta}\cdot\nabla v_1\|_\phi\} \lesssim \|g\|.$$

By Lemmas 1, 2:

$$\|\nabla\cdot\boldsymbol{\tau}_1\|_{\phi+\epsilon} \lesssim \max\{\sqrt{\epsilon}\|\nabla\cdot\boldsymbol{\tau}_1\|, \|\nabla\cdot\boldsymbol{\tau}_1\|_\phi\} = \max\{\sqrt{\epsilon}\|g+\boldsymbol{\beta}\cdot\nabla v_1\|, \|g+\boldsymbol{\beta}\cdot\nabla v_1\|_\phi\} \lesssim \|g\|$$

and

$$\frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}_1\|_{\phi+\epsilon} \lesssim \max\{\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}_1\|, \frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}_1\|_\phi\} = \max\{\frac{1}{\sqrt{\epsilon}}\|\epsilon\boldsymbol{\beta}\cdot\nabla v_1\|, \frac{1}{\epsilon}\|\epsilon\boldsymbol{\beta}\cdot\nabla v_1\|_\phi\} \lesssim \|g\|.$$

And finally by Lemma 1:

$$\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\tau}_1\| = \sqrt{\epsilon}\|\nabla v_1\| \lesssim \|g\|.$$

Therefore,

$$\|(v_1,\boldsymbol{\tau}_1)\|_{V,0} \lesssim \|g\|.$$

(ii) **Bound of $\|(v_2,\boldsymbol{\tau}_2)\|_{V,0}$.** As before, we use the relations defining $(v_2,\boldsymbol{\tau}_2)$, and Lemma 1 to bound the individual terms of $\|(v_2,\boldsymbol{\tau}_2)\|_{V,0}$. Specifically,

$$\begin{aligned}
\sqrt{\epsilon}\|\nabla v_2\| + \sqrt{\epsilon}\|v_2\| &\lesssim \|\boldsymbol{f}\|,\\
\|\boldsymbol{\beta}\cdot\nabla v_2\|_{\phi+\epsilon} &\lesssim \|\nabla v_2\| \lesssim \|\boldsymbol{f}\|,\\
\|\nabla\cdot\boldsymbol{\tau}_2\|_{\phi+\epsilon} &= \|\boldsymbol{\beta}\cdot\nabla v_2\|_{\phi+\epsilon} \lesssim \|\boldsymbol{f}\|,\\
\frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}_2\|_{\phi+\epsilon} &= \frac{1}{\epsilon}\|\epsilon\boldsymbol{\beta}\cdot(\boldsymbol{f}-\nabla v_2)\|_{\phi+\epsilon} \lesssim \|\boldsymbol{f}\| + \|\nabla v_2\| \lesssim \|\boldsymbol{f}\|,
\end{aligned}$$

and

$$\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\tau}_2\| = \sqrt{\epsilon}\|\boldsymbol{f}-\nabla v_2\| \lesssim \|\boldsymbol{f}\|.$$

This proves that

$$\|(v,\boldsymbol{\tau})\|_{V,0} \lesssim \|\boldsymbol{f}\|.$$

(iii) **Bound of $\|(v_0,\boldsymbol{\tau}_0)\|_{V,0}$.** By Lemma 3 and (2.41) there holds

$$\|\nabla v_0\| = \frac{1}{\epsilon}\|\boldsymbol{\tau}_0\| \lesssim \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}.$$

Moreover, by [4, Lemma 4.2],

$$\|v_0\| \lesssim \|\nabla v_0\| + \|\,[v]\,\|_{\Gamma_h}. \tag{2.42}$$

Therefore,

$$\begin{aligned}\|(v_0,\boldsymbol{\tau}_0)\|_{V,0} &= \sqrt{\epsilon}\|v_0\| + \sqrt{\epsilon}\|\nabla v_0\| + \|\boldsymbol{\beta}\cdot\nabla v_0\|_{\phi+\epsilon} + \frac{1}{\epsilon}\|\boldsymbol{\beta}\cdot\boldsymbol{\tau}_0\|_{\phi+\epsilon} + \frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\tau}_0\| + \|\nabla\cdot\boldsymbol{\tau}_0\|_{\phi+\epsilon}\\ &\lesssim \sqrt{\epsilon}\|v_0\| + \|\nabla v_0\| + \frac{1}{\epsilon}\|\boldsymbol{\tau}_0\| + \|\nabla\cdot\boldsymbol{\tau}_0\| \lesssim \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}.\end{aligned}$$

□

**Lemma 5** *Let* $(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\in U$ *be arbitrary. There holds*

$$\|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,1} \lesssim \|u\|_{1/(\phi+\epsilon)} + \frac{1}{\epsilon}\|\boldsymbol{\sigma}\|_{1/(\phi+\epsilon)} + \frac{1}{\sqrt{\epsilon}}\Big(\|\hat{u}\|_{1/2,\Gamma_h} + \|\hat{\sigma}_n\|_{-1/2,\Gamma_h}\Big). \tag{2.43}$$

*Furthermore, if* $\boldsymbol{\beta}$ *satisfies* (A1), (A2) *and* (A3) *then there holds*

$$\|u\| + \|\boldsymbol{\sigma}\| + \epsilon\|\hat{u}\|_{1/2,\Gamma_h} + \sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h} \lesssim \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,1}. \tag{2.44}$$

**Proof.** The proof is analogous to the one of Lemma 4. Using (2.34) and the definition (2.11) of the bilinear form, one sees that estimate (2.43) is equivalent to

$$\|(v,\boldsymbol{\tau})\|_{V,1} \gtrsim \|g\|_{\phi+\epsilon} + \epsilon\|\boldsymbol{f}\|_{\phi+\epsilon} + \sqrt{\epsilon}\sup_{(\hat{u},\hat{\sigma}_n)\neq 0}\frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\hat{u}\rangle + \langle\hat{\sigma}_n,[v]\rangle}{\|\hat{u}\|_{1/2,\Gamma_h} + \|\hat{\sigma}_n\|_{-1/2,\Gamma_h}}. \tag{2.45}$$

The bound

$$\|g\|_{\phi+\epsilon} + \epsilon\|\boldsymbol{f}\|_{\phi+\epsilon} \lesssim \|(v,\boldsymbol{\tau})\|_{V,1}$$

is immediate by the triangle inequality and the definition of the norms. To estimate the supremum we integrate by parts to represent the boundary dualities and to bound:

$$\begin{aligned}&\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],w\rangle + \langle(\boldsymbol{\eta}-\boldsymbol{\beta}w)\cdot\boldsymbol{n},[v]\rangle\\ &= (\boldsymbol{\eta},\nabla v)_{\Omega_h} + (\nabla\cdot\boldsymbol{\eta},v)_{\Omega_h} - (w,\boldsymbol{\beta}\cdot\nabla v)_{\Omega_h} - (\nabla\cdot(\boldsymbol{\beta}w),v)_{\Omega_h} + (w,\nabla\cdot\boldsymbol{\tau})_{\Omega_h} + (\nabla w,\boldsymbol{\tau})_{\Omega_h}\\ &\lesssim \Big(\sqrt{\epsilon}\|\boldsymbol{\beta}^\perp\cdot\nabla v\| + \|\boldsymbol{\beta}\cdot\nabla v\|_{\phi+\epsilon} + \sqrt{\epsilon}\|v\| + \|\nabla\cdot\boldsymbol{\tau}\|_{\phi+\epsilon} + \|\boldsymbol{\tau}\|_{\phi+\epsilon}\Big)\\ &\quad\left(\frac{1}{\sqrt{\epsilon}}\|\boldsymbol{\beta}^\perp\cdot\boldsymbol{\eta}\| + \|\boldsymbol{\beta}\cdot\boldsymbol{\eta}\|_{1/(\phi+\epsilon)} + \frac{1}{\sqrt{\epsilon}}\|\nabla\cdot\boldsymbol{\eta}\| + \|w\|_{1/(\phi+\epsilon)} + \frac{1}{\sqrt{\epsilon}}\|w\| + \frac{1}{\sqrt{\epsilon}}\|\nabla w\| + \|\nabla w\|_{1/(\phi+\epsilon)}\right)\\ &\lesssim \|(v,\boldsymbol{\tau})\|_{V,1}\frac{1}{\sqrt{\epsilon}}\left(\|\boldsymbol{\eta}\|_{\boldsymbol{H}(\mathrm{div},\Omega)} + \|w\|_{1,\Omega}\right)\end{aligned} \tag{2.46}$$

for any $w\in H_0^1(\Omega)$, $\boldsymbol{\eta}\in\boldsymbol{H}(\mathrm{div},\Omega)$. The last factor is the same as in (2.38). Therefore, exactly as in the proof of Lemma 4, we find that

$$\sup_{(\hat{u},\hat{\sigma}_n)\neq 0}\frac{\langle[\boldsymbol{\tau}\cdot\boldsymbol{n}],\hat{u}\rangle + \langle\hat{\sigma}_n,[v]\rangle}{\|\hat{u}\|_{1/2,\Gamma_h} + \|\hat{\sigma}_n\|_{-1/2,\Gamma_h}} \lesssim \frac{1}{\sqrt{\epsilon}}\|(v,\boldsymbol{\tau})\|_{V,1}.$$

This finishes the proof of (2.43). To prove (2.44) we show the equivalent estimate (cf. (2.40))

$$\|(v,\boldsymbol{\tau})\|_{V,1} \lesssim \|g\| + \|\boldsymbol{f}\| + \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}.$$

From the proof of Lemma 4 we already have the estimate

$$\sqrt{\epsilon}\|v\| + \sqrt{\epsilon}\|\nabla v\| + \|\boldsymbol{\beta}\cdot\nabla v\|_{\phi+\epsilon} + \|\nabla\cdot\boldsymbol{\tau}\|_{\phi+\epsilon} \lesssim \|g\| + \|\boldsymbol{f}\| + \frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}$$

so that it remains to bound $\|\boldsymbol{\tau}\|_{\phi+\epsilon}$. For a given $(v,\boldsymbol{\tau}) \in H_0^1(\Omega_h)\times\boldsymbol{H}(\mathrm{div},\Omega_h)$ we make use of the decomposition from the proof of Lemma 4: $(v,\boldsymbol{\tau}) = (v_1,\boldsymbol{\tau}_1) + (v_2,\boldsymbol{\tau}_2) + (v_0,\boldsymbol{\tau}_0)$ with $v_1, v_2 \in H_0^1(\Omega)$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2 \in \boldsymbol{H}(\mathrm{div},\Omega)$ and $(v_0,\boldsymbol{\tau}_0) = (v,\boldsymbol{\tau}) - (v_1,\boldsymbol{\tau}_1) - (v_2,\boldsymbol{\tau}_2)$.

From the definitions of $\boldsymbol{\tau}_1$ and $\boldsymbol{\tau}_2$, and Lemma 1 we obtain

$$\|\boldsymbol{\tau}_1\|_{\phi+\epsilon} = \epsilon\|\nabla v_1\|_{\phi+\epsilon} \lesssim \|g\|$$

and

$$\|\boldsymbol{\tau}_2\|_{\phi+\epsilon} = \|\epsilon(\boldsymbol{f} - \nabla v_2)\|_{\phi+\epsilon} \lesssim \|\boldsymbol{f}\|.$$

Lemma 3 and (2.41) prove that

$$\|\boldsymbol{\tau}_0\|_{\phi+\epsilon} \lesssim \|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0} + \sqrt{\epsilon}\|\,[v]\,\|_{\Gamma_h}.$$

The triangle inequality yields the wanted bound for $\|\boldsymbol{\tau}\|_{\phi+\epsilon}$ and finishes the proof. □

**Lemma 6** *Let $(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n) \in U$ be arbitrary. If $\boldsymbol{\beta}$ satisfies* (A1) *then there holds*

$$\|u\| + \|\boldsymbol{\sigma}\| + \|(\hat{u},\hat{\sigma}_n)\|_{\Gamma_h'} \simeq \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,qopt}. \tag{2.47}$$

*If $\boldsymbol{\beta}$ satisfies* (A1) *and* (A3) *then there holds*

$$\|u\| + \|\boldsymbol{\sigma}\| + \epsilon\|\hat{u}\|_{1/2,\Gamma_h} + \sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h} \lesssim \|(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\|_{E,qopt}. \tag{2.48}$$

**Proof.** Using duality as in the proof of Lemma 4, and the equivalence

$$\sup_{(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\neq 0} \frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|u\| + \|\boldsymbol{\sigma}\| + \|(\hat{u},\hat{\sigma}_n)\|_{\Gamma_h'}} \simeq \|g\| + \|\boldsymbol{f}\| + \|(v,\boldsymbol{\tau})\|_{\Gamma_h} = \|g\| + \|\boldsymbol{f}\| + \|\tilde{v}\|,$$

cf. (2.21), (2.22), we see that (2.47) is equivalent to

$$\|g\| + \|\boldsymbol{f}\| + \|\tilde{v}\| \simeq \|(v,\boldsymbol{\tau})\|_{V,qopt} = \|g\| + \|\boldsymbol{f}\| + \|v\|.$$

Both estimates are immediate by the triangle inequality and by bounding $\|v_0\| \lesssim \|g\| + \|\boldsymbol{f}\|$ with Lemma 1 and the superposition principle.

To show (2.48) we prove the equivalent estimate

$$\begin{aligned}\|(v,\boldsymbol{\tau})\|_{V,qopt} &\lesssim \sup_{(u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n)\neq 0} \frac{b((u,\boldsymbol{\sigma},\hat{u},\hat{\sigma}_n),(v,\boldsymbol{\tau}))}{\|u\|+\|\boldsymbol{\sigma}\|+\epsilon\|\hat{u}\|_{1/2,\Gamma_h}+\sqrt{\epsilon}\|\hat{\sigma}_n\|_{-1/2,\Gamma_h}} \\ &\simeq \|g\|+\|\boldsymbol{f}\|+\frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0}+\frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h},\end{aligned}$$

that is, it is enough to bound

$$\|v\| \lesssim \|g\|+\|\boldsymbol{f}\|+\frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0}+\frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}. \tag{2.49}$$

Exactly as in the proof of Lemma 4 we decompose $(v,\boldsymbol{\tau})\in H_0^1(\Omega_h)\times\boldsymbol{H}(\mathrm{div},\Omega_h)$ like $(v,\boldsymbol{\tau})=(v_1,\boldsymbol{\tau}_1)+(v_2,\boldsymbol{\tau}_2)+(v_0,\boldsymbol{\tau}_0)$ with $v_1,v_2\in H_0^1(\Omega)$, $\boldsymbol{\tau}_1,\boldsymbol{\tau}_2\in\boldsymbol{H}(\mathrm{div},\Omega)$ and $(v_0,\boldsymbol{\tau}_0)=(v,\boldsymbol{\tau})-(v_1,\boldsymbol{\tau}_1)-(v_2,\boldsymbol{\tau}_2)$. Then,

$$\|v\|\le\|v_0\|+\|v_1\|+\|v_2\|$$

and by Lemma 1,

$$\|v_1\|+\|v_2\|\lesssim\|g\|+\|\boldsymbol{f}\|.$$

Finally, combining (2.42) and Lemma 3, we obtain

$$\|v_0\|\lesssim\|\nabla v_0\|+\|\,[v]\,\|_{\Gamma_h}\lesssim\frac{1}{\epsilon}\|\,[\boldsymbol{\tau}\cdot\boldsymbol{n}]\,\|_{\Gamma_h^0}+\frac{1}{\sqrt{\epsilon}}\|\,[v]\,\|_{\Gamma_h}.$$

This proves (2.49). □

# 3 Numerical experiments

In this section, we report on numerical experiments with the DPG method for a 2D model problem introduced by Eriksson and Johnson [10],

$$\begin{aligned}-\epsilon\Delta u+\frac{\partial u}{\partial x}&=f \quad \text{in } \Omega=(0,1)^2,\\ u&=u_0 \quad \text{on } \partial\Omega,\end{aligned}$$

with coordinates $(x,y)$ so that $\boldsymbol{\beta}=(1,0)^T$ in $\Omega$. We shall study numerically the case with $f=0$ and boundary condition

$$u_0(x,y)=\begin{cases}u_0(y), & x=0,\ y\in(0,1),\\ 0, & \text{otherwise.}\end{cases}$$

We use separation of variables to compute the exact solution,

$$u(x,y)=\sum_{n=1}^{\infty}C_n\frac{e^{r_{1,n}(x-1)}-e^{r_{2,n}(x-1)}}{e^{-r_{1,n}}-e^{-r_{2,n}}}\sin(n\pi y) \tag{3.50}$$

where

$$u_0(y) = \sum_{n=1}^{\infty} C_n \sin(n\pi y), \qquad r_{1(2),n} = \frac{-1 \pm \sqrt{1 + 4\epsilon^2 n^2 \pi^2}}{-2\epsilon}.$$

We will use a manufactured solution corresponding to the leading term $n = 1$ of the series,

$$u(x, y) = \frac{e^{r_1(x-1)} - e^{r_2(x-1)}}{e^{-r_1} - e^{-r_2}} \sin(\pi y)$$

with

$$r_{1,2} = \frac{-1 \pm \sqrt{1 + 4\epsilon^2 \pi^2}}{-2\epsilon}.$$

The solution is displayed in Fig. 1.

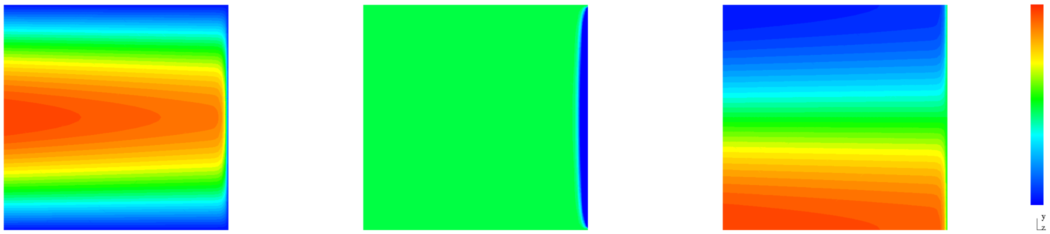

Figure 1: 2D model problem, $\epsilon = 10^{-2}$. Velocity $u$ and "stresses" $\sigma_x, \sigma_y$ (using scale for $\sigma_y$).

**Meshes.** We shall use quadrilateral (rectangular) meshes and focus on $h$-adaptivity. Our standard choice is going to be bilinear elements for the field variables $\sigma_1, \sigma_2, u$. The logic of the code is based on the exact sequence, so the corresponding $H^1$-conforming element is $Q^{(2,2)} = \mathcal{P}^2 \otimes \mathcal{P}^2$ (tensor products of quadratic polynomials) which implies the default polynomial degree for traces to be $p = 2$. In turn, the corresponding Raviart-Thomas element is $(\mathcal{P}^2 \otimes \mathcal{P}^1) \times (\mathcal{P}^1 \otimes \mathcal{P}^2)$, and this implies the linear approximation of fluxes, $p = 1$. All the experiments are done in the context of adaptive meshes. For $h$-adaptivity, we use the standard greedy algorithm with coefficient 0.5, i.e. all elements which contribute with an error that is above 50 percent of the maximum element contribution, are selected for refinement. For all other details on adaptivity we refer to [8].

## 3.1 Experiments with the quasi-optimal test norm

Recall the definition of the quasi-optimal test norm:

$$\|(\boldsymbol{\tau}, v)\|^2_{V,qopt} = \|\epsilon^{-1}\boldsymbol{\tau} + \nabla v\|^2 + \|\nabla\cdot \boldsymbol{\tau} - \frac{\partial v}{\partial x}\|^2 + \|v\|^2.$$

We did not implement an adaptive solver for determining optimal test functions. In view of our 1D experiments [9], the examples that we present are thus rather limited. We start with a diffusion dominated regime, $\epsilon = 0.1$. The optimal test functions were determined using standard enriched spaces with polynomial degree increment $\Delta p = 5, 6$. Moreover, we use only $h$-adaptivity. Results for the first six meshes are reported in Fig. 2.

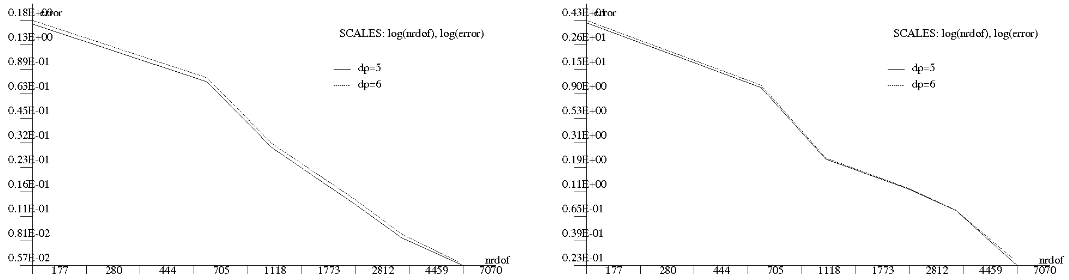

Figure 2: 2D model problem, quasi-optimal norm, $\epsilon = 10^{-1}, \Delta p = 5, 6$. Left: convergence of the error in energy norm. Right: convergence in relative $L^2$-norm for the field variables (in percent of their $L^2$-norm).

The small difference between the two energy- and $L^2$-norm curves indicates that the optimal test functions have practically been resolved. The ratio of the $L^2$ to the energy error, displayed in Fig 3, is no longer one as in the 1D case (cf. [9]), but varies between 0.02 and 0.12. The method does deliver a picture perfect solution. For illustration, Fig. 4 presents the error in $\sigma_1$ on the mesh after six $h$-refinements. The method delivers 3 digits in the solution.

**Experiments with Shishkin meshes for $\epsilon = 10^{-2}$.** An attempt to resolve the optimal test functions for $\epsilon = 10^{-2}$ using the simple enrichment with $\Delta p = 6$ has failed. The optimal test functions are clearly underresolved which manifests itself for instance in a loss of monotonic decrease of the error with refinements. Following [11], we have implemented Shishkin sub-meshes to resolve optimal test functions. The Shishkin-type sub-element tensor product mesh corresponds to a 1D master element sub-mesh consisting of three subelements with nodes placed at

$$x_0 = 0, \quad x_1 = \frac{\epsilon}{h}\hat{p}, \quad x_2 = 1 - \frac{\epsilon}{h}\hat{p}, \quad x_3 = 1$$

where $h$ is the element size and $\hat{p}$ is the degree of polynomials employed in the sub-mesh. Upon transforming to the actual element of length $h$, the sub-mesh nodes are placed at

$$0, \quad \epsilon\hat{p}, \quad h - \epsilon\hat{p}, \quad h.$$

It is assumed, of course, that $\frac{\epsilon}{h}\hat{p} < \frac{1}{2}$, so that the experiments with just Shishkin meshes can be performed only in the preasymptotic regime.

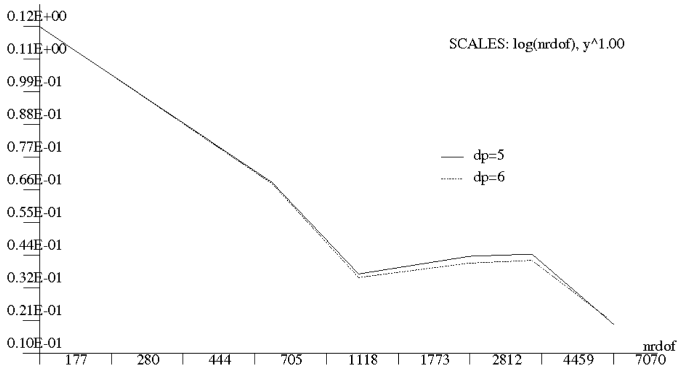

Figure 3: 2D model problem, quasi-optimal norm, $\epsilon = 10^{-1}, \Delta p = 5, 6$. Ratio of $L^2$ and energy norms.

Resolution of optimal functions involves discretization of the $H^1$-component $v$, and the $H(\text{div})$ component $\boldsymbol{\tau}$. We have used a Nedéléc-like space corresponding to the Shishkin submesh to resolve $\boldsymbol{\tau}$. In other words, if $W^{\hat{p}}$ denotes the $H^1$-conforming space of order $\hat{p}$ in 1D, for the corresponding $H(\text{div})$-conforming space in 2D we choose

$$(W^{\hat{p}} \otimes W^{\hat{p}-1}) \times (W^{\hat{p}-1} \otimes W^{\hat{p}}).$$

We have tried to run the adaptive code with $\hat{p} = 1, 2, 3, 4$. The results for different $\hat{p}$ differ significantly from each other, and the behavior of the energy error in none of the cases was monotone.

In view of the results, we conclude that the use of Shiskin meshes does not guarantee the resolution of optimal mesh functions. The problem may be related to the use of irregular meshes.

### 3.2 Experiments with weighted test norms

We have experimented with three weighted test norms. The first one is

$$\|(v, \boldsymbol{\tau})\|_{V,1}^2 := \epsilon\|v\|^2 + \|\boldsymbol{\beta} \cdot \nabla v\|_{\phi+\epsilon}^2 + \epsilon\|\nabla v\|^2 + \|\boldsymbol{\tau}\|_{\phi+\epsilon}^2 + \|\nabla \cdot \boldsymbol{\tau}\|_{\phi+\epsilon}^2$$

and corresponds to the one covered by Theorem 1(iii). We select the weight function $\phi(x, y) = x$. Recall the reasoning used to construct the norm. The stability estimates for the adjoint equation tell us that we have a robust control of the streamline component of $\nabla v$, i.e. simply $\partial v/\partial x$ in the case of our model problem. The stability comes from the convection, not diffusion. In fact, the weight reflects a "compromise" between the convection and diffusion. Without the diffusion, we would have had the control of the streamline component of the gradient in the standard $L^2$

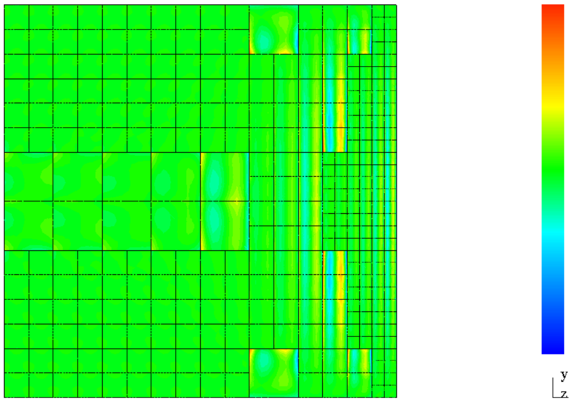

Figure 4: 2D model problem, quasi-optimal norm, $\epsilon = 10^{-1}, \Delta p = 5$. Pointwise error in $\sigma_1$, and optimal mesh corresponding to 0.04 percent relative $L^2$-error. Range: ±0.002.

(unweighted) norm. Stability of the crosswind component of $\nabla v$ comes from the diffusion and it depends upon $\epsilon$. This is reflected by the third term of the test norm. The $L^2$ term is also prescaled with $\epsilon$. This is *not* because of the stability results, we *do* have a robust control of $\|v\|$. By scaling the $L^2$ term though, we avoid boundary layers in the optimal test functions. Indeed, without the $\epsilon$ scaling of the $L^2$-norm of $v$, some of the optimal test functions develop boundary layers along north and south boundaries of elements, and their resolution is as problematic as for the quasi-optimal test norm. With the prescaling, there is no conflict between the lower and higher order terms in the definition of the norm, and the boundary layers are avoided.

A similar reasoning is used to define the norm for the $\boldsymbol{\tau}$ component. From the control of $\boldsymbol{\beta} \cdot \nabla v$ in norms with weights $\phi$ and $\epsilon$, it follows the use of weight $\phi + \epsilon$ in the norm for $\nabla \cdot \boldsymbol{\tau}$. The same weight is then used to define the lower order term to avoid possible boundary layers.

In essence, the weighted norm reflects everything we have learned on the stability, and an attempt to avoid test functions with boundary layers.

The second weighted norm is the one used in [8]. At that point, we had little understanding of robustness in multi-dimensions and the choice represented an extrapolation of our 1D results. The norm is defined as follows.

$$\|(\boldsymbol{\tau}, v)\|^2_{V,2} = \|\nabla \cdot \boldsymbol{\tau}\|^2_{\phi+\epsilon} + \|\boldsymbol{\tau}\|^2_{\phi+\epsilon} + \|\nabla v\|^2_{\phi+\epsilon} + \|v\|^2_{\phi+\epsilon}.$$

The third weighted norm is tailored for this specific example,

$$\|(v, \boldsymbol{\tau})\|^2_{V,3} := \min\{\frac{\epsilon}{h_2^2}, 1\}\|v\|^2 + \|\frac{\partial v}{\partial x}\|^2_{\phi+\epsilon} + \epsilon\|\frac{\partial v}{\partial y}\|^2 + \min\{\frac{1}{\epsilon}, \frac{1}{h_1 h_2}\}\|\boldsymbol{\tau}\|^2_{\phi+\epsilon} + \|\nabla \cdot \boldsymbol{\tau}\|^2_{\phi+\epsilon}.$$

With the use of rectangular meshes aligned with coordinates, $h_1, h_2$ denote the element lengths in $x$ and $y$ directions. The definition reflects the fact that the conflict between diffusion and reaction terms (producing boundary layers in optimal test functions) diminishes with element size, and rescales the reaction terms accordingly to the scaling of the crosswind diffusion term for $v$, and divergence for $\boldsymbol{\tau}$. The norm changes with the mesh (it depends upon the element size) but it satisfies the criteria implied by the stability analysis for the adjoint equation. By rescaling the $L^2$ terms with mesh dependent constants, we not only fight off the round-off error but improve the best approximation error estimates. The norm can be defined in a similar way for a general advection vector $\boldsymbol{\beta}$ and arbitrary meshes.

All results for the weighted norms have been obtained using the (anisotropic) $hp$-adaptive strategy described in detail in [8]. In short, we use the greedy strategy combined with a switch from $h$-refinement to $p$-refinement when the element size (in any direction) reaches $\epsilon$. The minimum element size is thus limited by $\epsilon$. To control the round-off error, we also impose a limit on the order of approximation, $p \leq p_{max} = 4$.

Fig. 5 presents the convergence history for the first weighted norm and three values of the diffusion coefficient $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. As predicted by the theory, the energy error decreases monotonically in the whole range of refinements. For $\epsilon = 10^{-4}$, the convergence stalls which we attribute to the round-off effects for the smallest elements (cf. [8]).

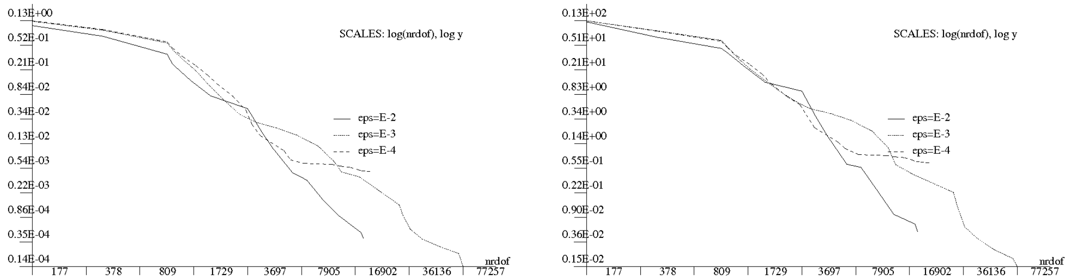

Figure 5: 2D model problem, weighted test norm $\|\cdot\|_{V,1}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Left: convergence of the error in energy norm. Right: convergence in relative $L^2$-norm for the field variables (in percent of their $L^2$-norm).

Fig. 6 shows the ratio of the $L^2$-error for the field variables $\sigma_1, \sigma_2, u$ and the energy error. The ratio stays uniformly bounded not only from above, as predicted by the theory, but also from below. Contrary to 1D experiments in [9], it does not converge asymptotically to any number.

Finally, Figures 7 and 8 present results for the "non-kosher" weighted norm $\|\cdot\|_{V,2}$. Consistently with the experience reported in [8], the overall performance of the method is as good as for the weighted norm $\|\cdot\|_{V,1}$ justified by the theory. However, the ratio of the $L^2$ and energy norms does not stay uniformly bounded from above as it begins to grow in the end of the convergence process.

Finally, Figures 9 and 10 present results for the rescaled "kosher" weighted norm $\|\cdot\|_{V,3}$.

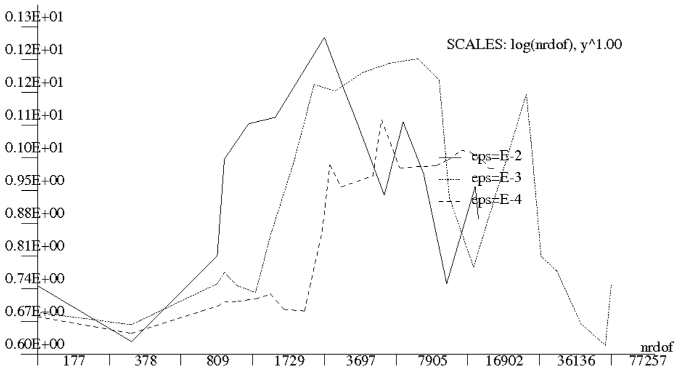

Figure 6: 2D model problem, weighted test norm $\| \cdot \|_{V,1}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Ratio of $L^2$ and energy norms.

Clearly, the results are the best out of the three norms. We were able to complete the computations for all three values of $\epsilon$. Notice the sudden drop in both energy and $L^2$-errors for $\epsilon = 10^{-3}, 10^{-4}$ due to $p$-refinements in the boundary layer. Consistently with the theory, the ratio of $L^2$ and energy errors stays bounded uniformly in $\epsilon$.

### 3.3 An Extra Example.

In order to pin down better the difference between the three weighted norms, we have considered again the model problem of Eriksson and Johnson but with an inflow boundary condition corresponding to the first twenty terms of the Fourier expansion (3.50) of a discontinuous piecewise linear function. The exact (manufactured) solution for $\epsilon = 10^{-2}$ is displayed in Fig. 11.

#### Case: $\epsilon = 10^{-2}$

We start with a moderate diffusion constant $\epsilon = 10^{-2}$. Comparison of results for the three weighted norms is presented in Figures 12 and 13. As predicted by theory, the ratio of the $L^2$ and energy norms shoots higher for the non-kosher norm $\| \cdot \|_{V,2}$. The overall performances of the three methods, though, are quite comparable.

#### Case: $\epsilon = 10^{-4}$

We now turn to a more demanding and interesting case for $\epsilon = 10^{-4}$. Recall that this was the smallest value of the diffusion constant for which we have managed to solve the previous example. Solution for smaller values of $\epsilon$ requires the use of mesh-dependent norms [8].

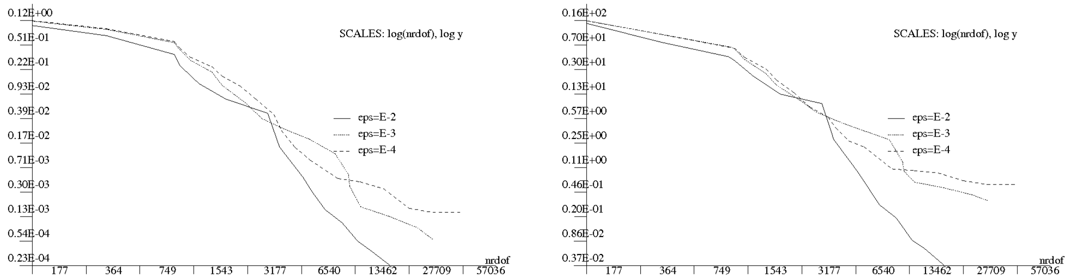

Figure 7: 2D model problem, weighted test norm $\| \cdot \|_{V,2}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Left: convergence of the error in energy norm. Right: convergence in relative $L^2$-norm for the field variables (in percent of their $L^2$-norm).

**Kosher norm $\| \cdot \|_{V,1}$.** Figures 14 and 15 present results obtained with the test norm $\| \cdot \|_{V,1}$ and enriched spaces with $\Delta p = 2, 3$. The computations with $\Delta p = 2$ were stopped shortly after the energy error began to increase, a behavior inconsistent with theory and indicating that the optimal test functions had not been resolved. The two curves begin to diverge from each other only after 16 iterations. This could be explained with an observation that most of the refinements in the first steps take place near the inflow boundary where the weight $\phi = x$ is small and, consequently, the anisotropy present in the test norm of $v$ is less pronounced. The ratio of $L^2$ and energy error shoots a bit higher for $\Delta p = 2$ but the difference between the two plots is not dramatic.

**Non-kosher test norm $\| \cdot \|_{V,2}$.** The results for the "non-kosher" test norm $\| \cdot \|_{V,2}$ are shown in Figures 16 and 17. The norm is less sensitive to $\Delta p$, the two convergence curves overlap with each other. This is consistent with the results reported in [8]. The two curves begin to slightly diverge from each other by the end of the process. The convergence for $\Delta p = 3$ eventually breaks down which indicates round-off error problems with the smallest elements. In [8], we were able to finish the convergence process for $\epsilon = 10^{-4}$. There are essentially only two differences between the codes. With linear elements for field variables, we use now continuous quadratics to approximate traces, whereas in [8], we used discontinuous linears. The second difference is in the solution of local problems for the optimal test functions. In [8] we used $H^1$-conforming shape functions for $\boldsymbol{\tau}$. In the present work, we use the $H(\text{div})$-conforming shape functions.

The ratio of $L^2$ and energy norms, presented in Fig. 17, shoots up for both values of $\Delta p$, consistently with the theory.

**Rescaled kosher test norm $\| \cdot \|_{V,3}$.** The results for the rescaled kosher norm $\| \cdot \|_{V,3}$ are presented in Figures 18 and 19. Similarly as for the second norm, we were able to push the

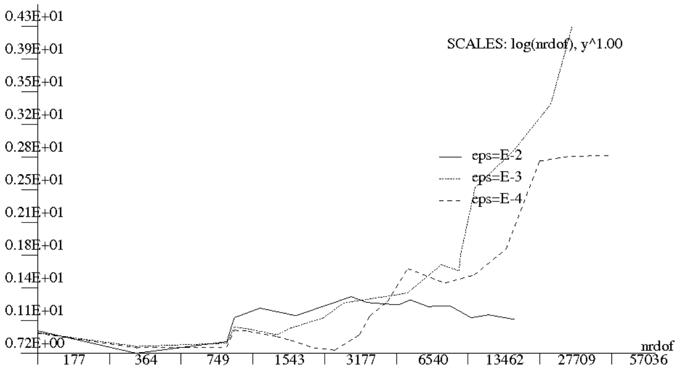

Figure 8: 2D model problem, weighted test norm $\| \cdot \|_{V,2}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Ratio of $L^2$ and energy norms.

computations to an error level that guarantees the resolution of the boundary layer. Notice that, for the smaller constant $\epsilon$, this requires reaching a smaller value of the $L^2$ error. The difference in convergence histories indicates some sensitivity to the resolution of optimal test functions. The acceleration in convergence that occurs when the $p$-refinements occur, happens later with $\Delta p = 3$. We cannot claim thus that the optimal test functions have been fully resolved with $\Delta p = 2$.

Finally, Fig. 20 presents the $L^2$-convergence history for all three weighted test norms and $\Delta p = 2$. Clearly, the performance of the rescaled norm $\| \cdot \|_{V,3}$ supported by the theory is the best. First of all, we were able to finish the computations all the way to a full resolution of the boundary layer. The rescaling in the small element regime not only successfully fights off the round-off error but it also delivers lower approximation error. This is in agreement with the best approximation error estimates.

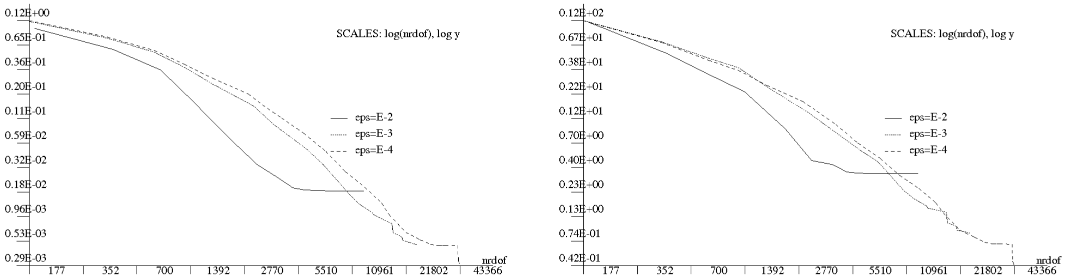

Figure 9: 2D model problem, weighted test norm $\|\cdot\|_{V,3}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Left: convergence in energy norm. Right: convergence in relative $L^2$-norm for the field variables (in percent of their $L^2$-norm).

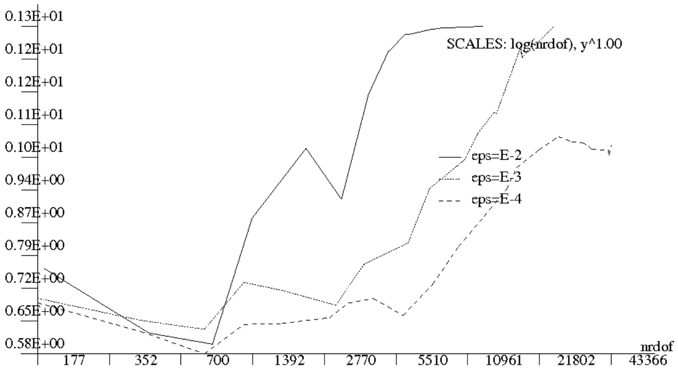

Figure 10: 2D model problem, weighted test norm $\|\cdot\|_{V,3}$, $\epsilon = 10^{-2}, 10^{-3}, 10^{-4}$. Ratio of $L^2$ and energy norms.

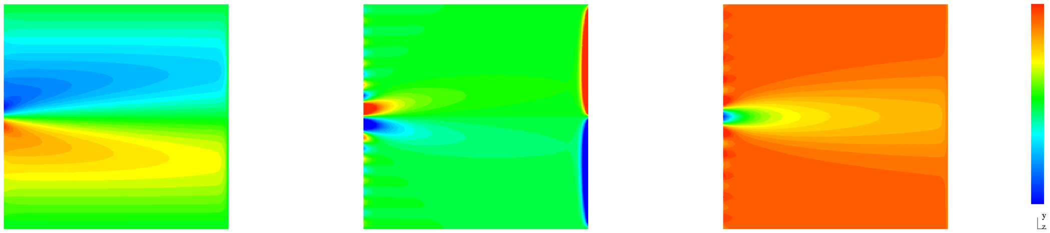

Figure 11: 2D model problem II (with “discontinuous” inflow data), $\epsilon = 10^{-2}$. Velocity $u$ and “stresses” $\sigma_x, \sigma_y$ (using scale for $\sigma_y$).

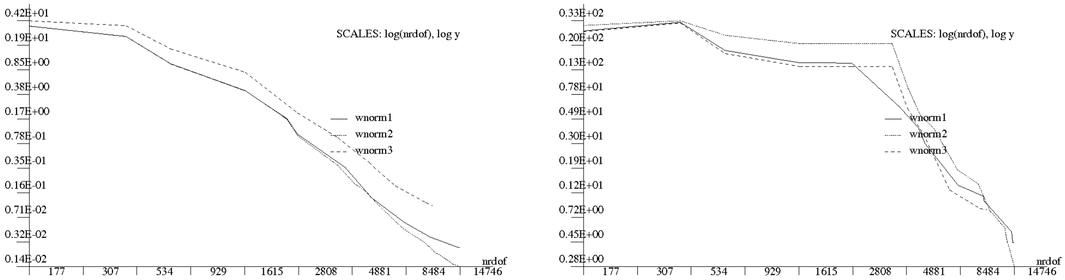

Figure 12: 2D model problem II, $\epsilon = 10^{-2}$. Comparison of three weighted norms. Left: convergence in energy norm. Right: convergence in relative $L^2$-error for the field variables (in percent of their $L^2$-norm).

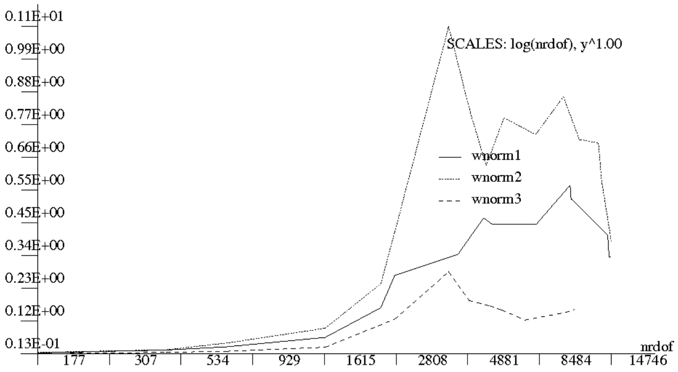

Figure 13: 2D model problem II, $\epsilon = 10^{-2}$. Comparison of three weighted norms. Ratio of $L^2$ and energy norms.

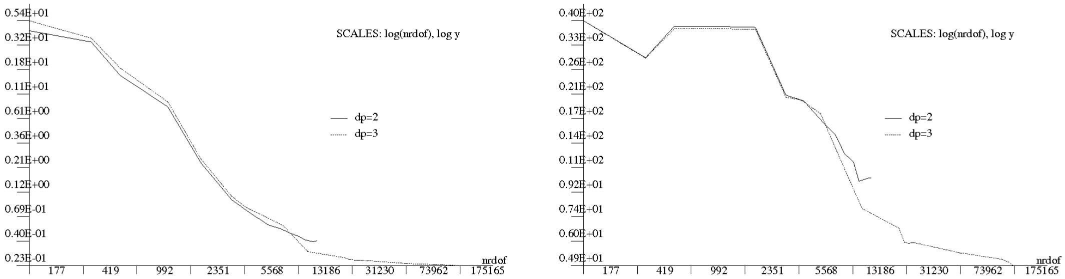

Figure 14: 2D model problem II, $\epsilon = 10^{-4}$, kosher weighted norm $\| \cdot \|_{V,1}$. Comparison of results for $\Delta p = 2, 3$. Left: convergence in energy error. Right: convergence in relative $L^2$-error for the field variables (in percent of their $L^2$-norm).

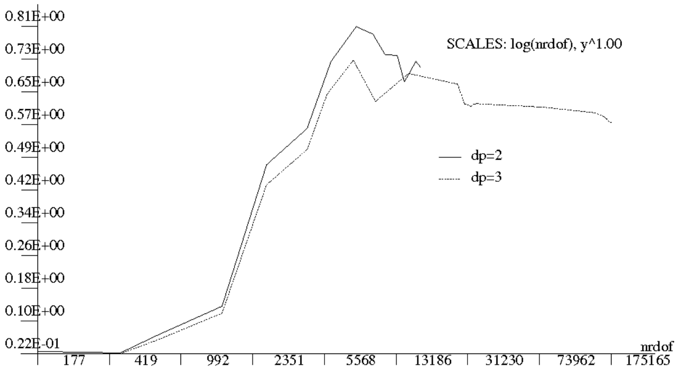

Figure 15: 2D model problem II, $\epsilon = 10^{-4}$, kosher weighted norm $\| \cdot \|_{V,1}$. Comparison of results for $\Delta p = 2, 3$. Ratio of $L^2$ and energy norms.

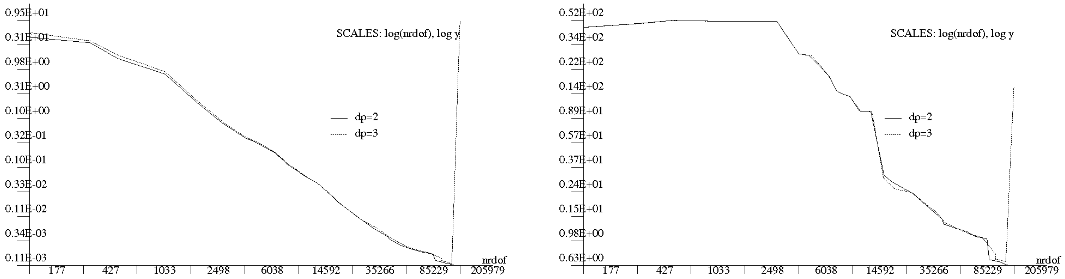

Figure 16: 2D model problem II, $\epsilon = 10^{-4}$, non-kosher weighted norm $\| \cdot \|_{V,2}$. Comparison of results for $\Delta p = 2, 3$. Left: convergence in energy norm. Right: convergence in relative $L^2$-error for the field variables (in percent of their $L^2$-norm).

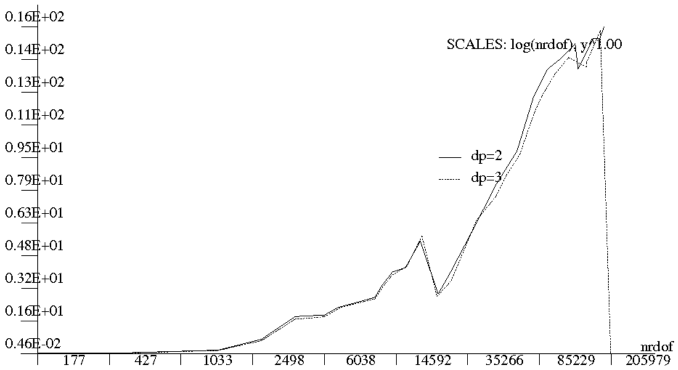

Figure 17: 2D model problem II, $\epsilon = 10^{-4}$, non-kosher weighted norm $\| \cdot \|_{V,2}$. Comparison of results for $\Delta p = 2, 3$. Ratio of $L^2$ and energy norms.

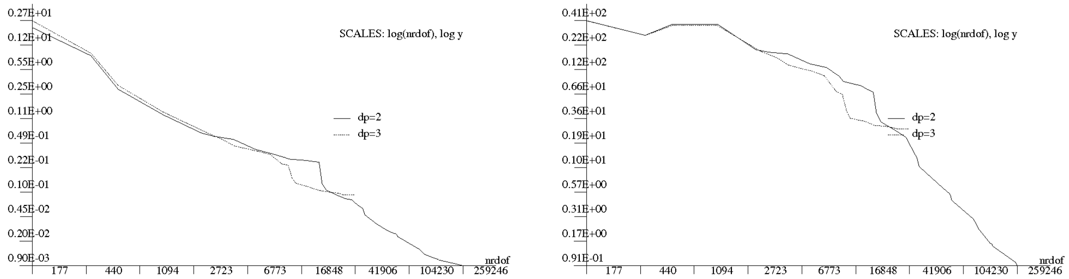

Figure 18: 2D model problem II, $\epsilon = 10^{-4}$, rescaled kosher weighted norm $\| \cdot \|_{V,3}$. Comparison of results for $\Delta p = 2, 3$. Left: convergence in energy norm. Right: convergence in relative $L^2$-error for the field variables (in percent of their $L^2$-norm).

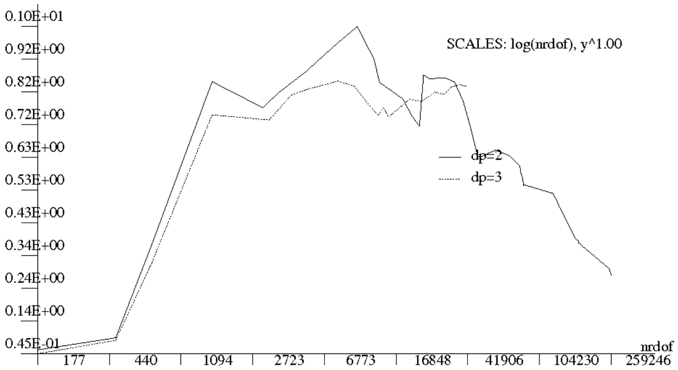

Figure 19: 2D model problem II, $\epsilon = 10^{-4}$, rescaled kosher weighted norm $\| \cdot \|_{V,3}$. Comparison of results for $\Delta p = 2, 3$. Ratio of $L^2$ and energy norms.

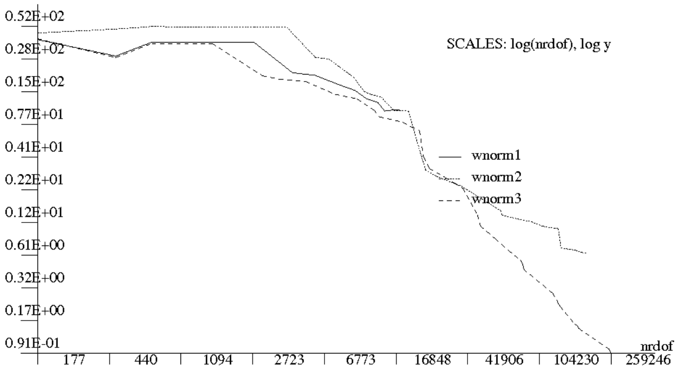

Figure 20: 2D model problem II, $\epsilon = 10^{-4}$. Convergence history in $L^2$-norm for the three weighted norms.

## 4 Conclusions

We have presented a general strategy aiming at the construction of an optimal test norm for the DPG method applied to singularly perturbed problems. The strategy requires a stability analysis of the *classical* adjoint problem accounting for the dependence upon the perturbation parameter, i.e. *robust* error estimates.

We have applied the strategy to the important model problem of convection-dominated diffusion. The obtained stability estimates for the adjoint problem extend earlier results from Eriksson and Johnson [10].

Given the stability results, one can construct many test norms that guarantee bounding the $L^2$-error in the field variables by the DPG energy (residual) error in a robust way, i.e. uniformly in the diffusion coefficient $\epsilon$. The quasi-optimal test norm, implied by the Banach closed range theorem [13] and proposed in context of the confusion problem in [12], admits also a robust upper bound, i.e. the energy error can be bounded uniformly in $\epsilon$ by the best approximation error involving the $L^2$-norm for the field variables.

Unfortunately, the optimal test functions corresponding to the quasi-optimal test norm exhibit boundary layers whose resolution is very demanding.

Based on the stability analysis, we have proposed two new weighted test norms that are consistent with theory and produce test functions that can be easily resolved with enriched spaces and $\Delta p = 2$. Especially attractive is the rescaled norm that presents an optimal compromise between three conditions:

- it satisfies robust stability estimate (1.7) and, therefore, guarantees a robust bound of the $L^2$-error in the field variables with the energy error;
- it enables resolution of the optimal test functions in the small elements limit;
- it implies better best approximation error estimates than the unscaled weighted norm.

Presented numerical results illustrate and confirm our theory.

In conclusion, we believe that the weighted test norm presents an optimal compromise between robustness and practicality of the DPG method for convection-dominated diffusion problems.

**Current work.** The methodology presented in this paper opens up doors for a systematic construction of robust DPG methods for different singularly perturbed problems. There is no miracle, without a systematic analysis of the stability of classical, strong formulations, we act blind folded, and may only try to guess appropriate test norms. Construction of such test norms is clearly very problem dependent. Our current efforts focus on generalizing presented results to linearized compressible Navier-Stokes equations.

The main ambition of the DPG method is to provide a framework for *fully automatic adaptive methods* that allow for *a full resolution* of boundary layers and other irregularities related to the nature of singularly perturbed problems. The unprecedented feature of the method is that it offers stability for a very general class of grids including *hp*-adaptive meshes. This enables anisotropic *hp*-adaptivity, a technique which is crucial for the resolution of boundary layers, (smeared) shocks and singularities. The use of highly irregular *hp* grids presents several challenges. We mention just two:

- The round-off error limitations are critical, and result in limiting both element size and order of approximation, possibly also aspect ratios. For $\epsilon \leq 10^{-5}$, one has to use rescaled test norms proposed in [8]. Stability and convergence analysis for such norms remains an open issue.

- We need a genuine *hp*-adaptive strategy. The strategy proposed in [8] and used in this paper, is limited to a rather narrow class of problems and has been intended mostly to illustrate the stability of the DPG method.

### Acknowledgments

Demkowicz was supported by the Department of Energy [National Nuclear Security Administration] under Award Number [DE-FC52-08NA28615]. Heuer was supported by Fondecyt-Chile under Grant Number 1110324. We thank for the J.T. Oden fellowship supporting a visit of the second author to ICES.

## References

[1] A. Cohen, W. Dahmen, and G. Welper, *Adaptivity and variational stabilization for convection-diffusion equations*, Bericht 323, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 2011.

[2] W. Dahmen, C. Huang, C. Schwab, and G. Welper, *Adaptive petrov Galerkin methods for first order transport equations*, Bericht 321, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, 2011.

[3] L. Demkowicz and J. Gopalakrishnan, *A new paradigm for discretizing difficult problems: Discontinuous Petrov Galerkin method with optimal test functions.* Expressions (publication of International Association for Computational Mechanics), November 2010.

[4] ——, *Analysis of the DPG method for the Poisson problem*, SIAM J. Numer. Anal., (2011).

[5] ——, *A class of discontinuous Petrov-Galerkin methods. Part II: Optimal test functions*, Numer. Methods Partial Differential Eq., (2011), pp. 70–105. See also ICES Report 9-16.

[6] L. Demkowicz, J. Gopalakrishnan, I. Muga, and D. Pardo, *A pollution free DPG method for multidimensional Helmholtz equation*, ICES Report, The University of Texas at Austin, 2011. In preparation.

[7] L. Demkowicz, J. Gopalakrishnan, I. Muga, and J. Zitelli, *Wavenumber explicit analysis for a DPG method for the multidimensional Helmholtz equation*, ICES Report 11-24, The University of Texas at Austin, 2011. Submitted to CMAME.

[8] L. Demkowicz, J. Gopalakrishnan, and A. Niemi, *A class of discontinuous Petrov-Galerkin methods. Part III: Adaptivity*, Appl. Numer. Math., (2011). Accepted.

[9] L. Demkowicz and N. Heuer, *Robust DPG methods for 1d convection-dominated diffusion problems*, ICES Report, The University of Texas at Austin, 2011. In preparation.

[10] K. Eriksson and C. Johnson, *Adaptive streamline diffusion finite elements methods for stationary convection-diffusion problems*, Math. Comp., 60 (1993), pp. 167–188.

[11] A. Niemi, N. Collier, and V. Calo, *Automatically stabilized discontinuous Petrov-Galerkin methods for stationary transport problems: Quasi-optimal test space norm*, (2011). In preparation.

[12] ——, *Discontinuous Petrov-Galerkin method based on the optimal test space norm for one-dimensional transport problems*, Journal of Computational Science, (2011). Submitted.

[13] J. Zitelli, I. Muga, L. Demkowicz, J. Gopalakrishnan, D. Pardo, and V. Calo, *A class of discontinuous Petrov-Galerkin methods. Part IV: Wave propagation problems*, J. Comp. Phys., 230 (2011), pp. 2406–2432. See also ICES Report 10-17.