Numerical issues in the implementation of high order polynomial multidomain penalty spectral Galerkin methods for hyperbolic conservation laws

Sigal Gottlieb[1] and Jae-Hun Jung [1,*]

[1] *Department of Mathematics, University of Massachusetts at Dartmouth North Dartmouth, MA 02747-2300, U.S.A.*

Abstract. In this paper, we consider high order multi-domain penalty spectral Galerkin methods for the approximation of hyperbolic conservation laws. This formulation has a penalty parameter which can vary in space and time, allowing for flexibility in the penalty formulation. This flexibility is particularly adventageous for problems with an inhomogeneous mesh. We show that the discontinuous Galerkin method is equivalent to the multidomain spectral penalty Galerkin method with a particular penalty term. The penalty parameter has an effect on both the accuracy and stability of the method. We examine the numerical issues which arise in the implementation of high order multi-domain penalty spectral Galerkin methods. The coefficient truncation method is proposed to prevent the rapid error growth due to round-off errors when high order polynomials are used. Finally, we show that an inconsistent evaluation of the integrals in the penalty method may lead to growth of errors. Numerical examples for linear and nonlinear problems are presented.

Key words: High Order Polynomial Galerkin Methods, Penalty Boundary Conditions, Discontinuous Galerkin Methods, Hyperbolic Conservation Laws, Round-off Errors, Truncation Methods.

1 Introduction

Polynomial Galerkin methods are widely used for the numerical solution of hyperbolic conservation laws [1, 2, 9, 10]. These methods seek a polynomial approximation of the solution for which the projected residual of the differential equation to the polynomial space vanishes. Two such classes of methods are the spectral Galerkin methods (sGM) and the discontinuous Galerkin methods (dGM). Traditionally, sGM have used high order polynomials on one element, while dGM use lower order polynomials on many elements. However, multi-domain sGM exist and are known as spectral element methods. In order to increase the accuracy of the approximation, these methods can use more smaller elements

*Corresponding author. *Email addresses:* jjung@umassd.edu (J.-H. Jung), sgottlieb@umassd.edu (S. Gottlieb)

(h-refinement) or raise the degree of the polynomial in each element (p refinement). High order polynomials have numerical issues such as sensitivity to roundoff errors, so it is important to carefully study their effects on accuracy and stability when used in multi-domain penalty spectral Galerkin methods.

The penalty formulation penalizes the boundary or interface conditions at each element by introducing a penalty term which includes a penalty parameter. This penalty parameter confers a great deal of flexibility on the problem, as it can change over space and time. We demonstrate the advantages of the flexibility in the choice of penalty parameter, especially in the case where an inhomogeneous grid system is used. An inhomogeneous grid can be due to a difference in element size or in polynomial order at each element, but this type of grid is subject to non-physical reflecting or dispersive modes which may appear in the solution. By modifying the penalty conditions near the grid discontinuity, we show that the sGM can reduce the non-physical modes while computing the other elements efficiently and accurately. We further consider the effects of the penalty method on the stability and accuracy. In this context, we show that the dG formulation is a special case of the penalty multi-domain sGM.

Next we discuss the effect of round-off errors for high order sGMs. These round-off error effects can arise from the the ill-conditioned mass matrix for high order polynomials, and the numerically inconsistent evaluations of the mass matrix and the load vector. The coefficient truncation method is introduced to reduce round-off errors. This method truncates high order coefficients which do not show rapid decay in the solution of the linear system, and prevents error growth.The second round-off error effect we explore is the error resulting from inconsistent evaluations of the two sides of the equation. While the right-hand-side of the linear system is evaluated by quadrature, the left-hand side can usually be computed exactly. In the case of orthogonal polynomials, this matrix is a diagonal matrix which simplifies the process of solving the system. However, computing one side of the equation exactly and the other by quadrature results in inconsistency errors which rise when the polynomial order is raised. We show these numerical inconsistency errors, and their dependence on the penalty parameter, in numerical computations.

The paper is structured as follows. In Section 2, we formulate the multidomain penalty sGM and demonstrate the effect of the penalty terms on the stability and accuracy. The penalty sGM with non-homogeneous grid and the flexibility of the penalty methods are discussed. The equivalence of the dGM to the sGM is shown as well. In Section 3, the effect of round-off errors on the high order sGM is discussed. The coefficient truncation method is introduced and numerical consistency is studied for the reduction of round-off errors. In Section 4, we summarize our results and discuss future research directions.

2 Penalty Spectral Galerkin method

In this paper, we consider the one-dimensional hyperbolic conservation law

$$u_t + f(u)_x = 0, \quad x \in \Omega = [-1,1], \quad u : [-1,1] \times \mathbf{R}^+ \to \mathbf{R}, t > 0, \tag{2.1}$$

with the initial condition $u(x,0)=g(x)$, and boundary conditions $B^{\pm}u(x,t)=h^{\pm}(t), \quad x\in \partial\Omega$, where $B^{\pm}$ are the boundary operators at the domain boundary $\partial\Omega$. For simplicity, we will consider mainly the case $f'(u)\geq 0$, for which the boundary condition is $u(-1,t)=h^{+}(t)$. In practice, this case can be easily generalized using the flux splitting and treating each flux with the appropriate boundary condition.

In 1988, Gottlieb and Funaro introduced a penalty boundary condition for the spectral collocation approximation of Eq. (2.1) [4] and there have has been much research on the penalty collocation methods [3,5–8]. The main motivation of penalty boundary conditions for collocation methods is that the differential equation is satisfied exactly at a given set of collocation points, while at the boundaries we add a term which penalizes for the approximation's distance from the prescribed boundary conditions to make the given equations satisfied asymptotically at any points. The numerical approximation is the polynomial $U(x,t)=\sum_{k=0}^{m} b_k(t)P_k(x)$, where $P_k(x)$ are the basis polynomials of degree k in x and $b_k(t)$ are the unknown expansion coefficients, which will be determined. The spectral collocation penalty method leads to the requirement that

$$U_t+f(U)_x=\tau Q_m^{+}(x)\left(f(U(-1,t))-f\left(h^{+}(t)\right)\right),$$

at each of the collocation points $x=x_j$. The penalty parameter τ can depend on x and t. $Q_m^{+}(x)$ is a polynomial of degree m which vanishes at all the collocation points x_j, except at the boundary point $x=-1$, so that the penalty term is only applied to the boundary point.

To formulate the penalty Galerkin spectral method, we assume a solution of the form $U(x,t)=\sum_{k=0}^{m} b_k(t)P_k(x)$ and find the expansion coefficients by requiring that the projection of the residual onto the solution subspace vanish. To satisfy the boundary conditions, we can impose a penalty term on the strong formulation (as in the collocation case), and then require the projection of the residual of the penalized equation to vanish. Alternatively, we can impose the penalty term after the projection. If we choose $P_k(x)$ to be a set of orthogonal polynomials, such that $\int_{\Omega} P_k(x)P_j(x)dx=\gamma_k\delta_{jk}$, the penalty Galerkin formulation yields the system

$$\gamma_j b_j'(t)+\int_{\Omega} f(U)_x P_j(x)dx=\tau^j\left(f(U(-1,t))-f(h^{+}(t))\right)\int_{\Omega} Q_m^{+}(x)P_j(x)dx$$

for $\forall j=0,\cdots,m$, where the superscript $'$ denotes the derivative with respect to time and γ_j are the normalization factors.

We have flexibility in the choice of the penalty parameter τ^j and Q_m^{+}, provided only that the polynomial $Q_m^{+}(-1)=1$. One way to accomplish this is to set $Q_m^{+}(x)$ to be polynomials of degree 0, that is, $Q_m^{+}(x)=1=P_0(x)$. In this case, the penalty terms appear

only in the equation for $b_0'(t)$, that is,

$$\begin{aligned}\gamma_0 b_0'(t) &= -\int_\Omega f(U)_x dx + 2\tau^0\big(f(U(-1,t))-f(h^+(t))\big)\\ &= f(U(-1,t))-f(U(1,t))+2\tau^0\big(f(U(-1,t))-f(h^+(t))\big),\\ \gamma_j b_j'(t) &= -\int_\Omega f(U)_x P_j(x)dx, \quad j=1,\cdots,m.\end{aligned}$$

We refer to this formulation as the strong formulation. However, this choice of $Q_m^+(x)=1$ does not lead to good stability properties for a linear hyperbolic wave equation, as we will discuss below. A different choice of $Q_m^+(x)$ may resolve this problem. For example, if we let

$$Q_m^+ = \sum_{k=0}^{m} \frac{1}{\gamma_k} P_k(x),$$

then the integral in the penalty term can be evaluated exactly

$$\int_{-1}^{1} Q_m^+(x) P_j(x)dx = \int_{-1}^{1} \sum_{k=0}^{m} \frac{1}{\gamma_l} P_k(x) P_j(x)dx = P_j(-1).$$

For the orthogonal polynomials such as Chebyshev or Legendre polynomials, $P_j(-1) = (-1)^j$. In this case the formulation becomes similar to the collocation case above:

$$\gamma_j b_j'(t) = -\int_\Omega f(U)_x P_j(x)dx + \tau\big(f(U(-1,t))-f(h^+(t))\big)(-1)^j,$$

for $\forall j = 0,\cdots,m$. We refer to this formulation as the weak formulation. For the linear hyperbolic wave equation this penalty term yields a better stability profile.

To see the difference between these two formulations, consider the simple advection equation $u_t + u_x = 0$ with both the strong and weak penalty formulations. Figure 1 shows that the strong formulation (represented by $\times$) has positive real eigenvalues which will lead to instability, while the eigenvalues for the weak formulation (represented by $\circ$) are all in the left half plane. This is a clear indication that the choice of the penalty polynomial, $Q_m^+(x)$ is critical for stability. From now on, we will consider only the weak formulation polynomial.

The penalty parameter τ also plays an important role in the stability of the method. Figure 2 shows the eigenvalues with the dGM and the penalty sGM for two different values of τ. The top figures show the eigenvalues in the complex plane for $\tau=-1$(left) and $\tau=-2$ (right) for polynomial order $m=8$. In each figure, the symbols $\circ$ and $\times$ represent the eigenvalues with the dGM and the sGM respectively. From the left top figure it is clear that the dGM and sGM are identical for $\tau=-1$, but when τ is increased, as in the top right figure, the real part of one eigenvalue increases. This one eigenvalue alone is responsible for the growth in the spectral radius ρ. This eigenvalue will later be implicated in penalty sGM's increased sensitivity to roundoff errors for large τ.

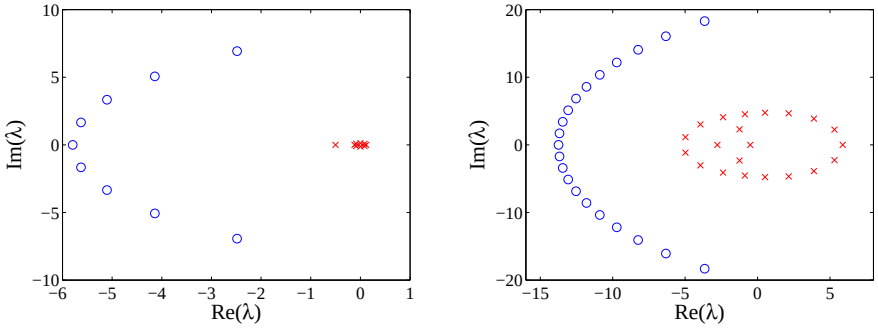

Figure 1: Eigenvalues in complex plane with the strong penalty formulation with $m=8$(left) and $m=20$ (right) for $\tau=-1$. The symbols $\circ$ and $\times$ represent the eigenvalues with the weak and strong penalty methods respectively.

The left bottom figure in Figure 2 shows the spectral radius ρ increasing as a function of the polynomial order m for $\tau=-1$. The right bottom figure shows the spectral radius ρ as a function of the penalty parameter τ, for various polynomial orders, on a logarithmic scale. The figure shows that ρ increases fast around $\tau=-1$.

The figures show that the stability of the formulation depends on the penalty parameters τ_j, as well as $Q_m^+(x)$. Later we will see that the choice of penalty parameter will affect the accuracy as well as stability of the method.

2.1 Multidomain spectral penalty Galerkin methods

To increase the order of accuracy in the penalty sGMs, we can allow the polynomial order to rise or we can divide the domain into many smaller subdomains, and apply the sGM in each domain. To formulate the multi-domain method, we divide the domain $\Omega=[-1,1]$ into N subdomains, or elements, I_l, $l=1,...,N$. For simplicity, we assume here that each element has the same polynomial order, m, but this is not necessary, or in fact advisable, in general. The solution in each domain is given by

$$U^l(x,t)=\sum_{k=0}^{m} b_k^l(t)P_k(\xi(x)), \quad x\in I_l=\left[x_{l-1/2},x_{l+1/2}\right],$$

where I_l is the lth element with the domain interval Δx_l and $x_{l-1/2}=x_l-\frac{\Delta x_l}{2}$, $x_{l+1/2}=x_l+\frac{\Delta x_l}{2}$ where x_l the cell center. Finally, $\xi(x)$ is the linear map $\xi: x\longmapsto[-1,1]$. For each element, we have for each $j=0,\cdots,N$

$$\sum_{k=0}^{m}\frac{db_k^l(t)}{dt}\int_{I_l}P_j(\xi(x))P_k(\xi(x))dx+\int_{I_l}P_j(\xi(x))f(U^l)_x dx=\mathcal{P}_{I_l}^j, \tag{2.2}$$

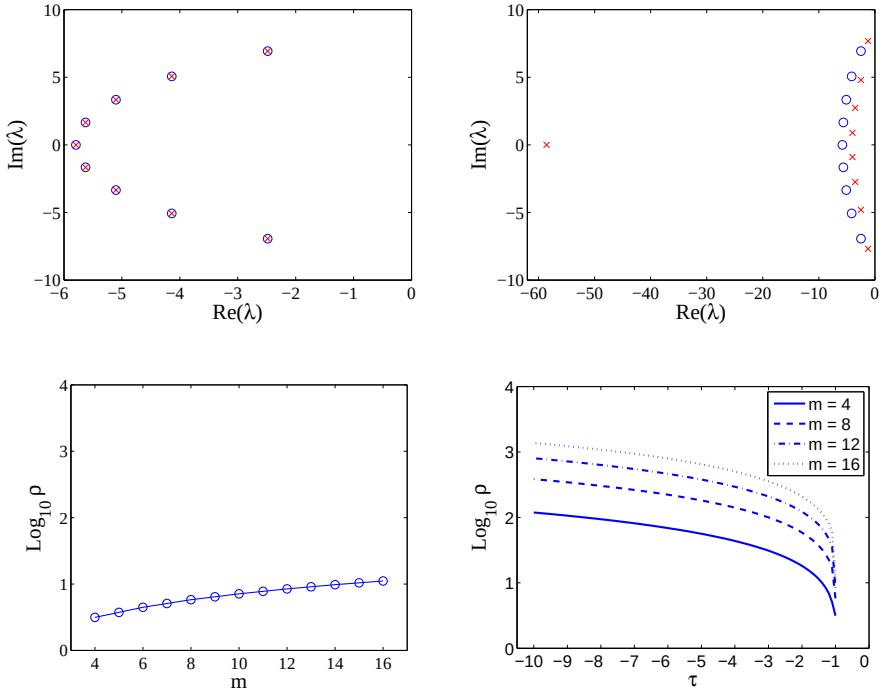

Figure 2: Top: Eigenvalues in complex plane with $\tau=-1$(left) and $\tau=-2$(right) for $m=8$. The symbols $\circ$ and $\times$ represent the eigenvalues with the dGM and the sGM respectively. Bottom: The spectral radius ρ versus τ with the dGM(left) and the sGM(right) in logarithmic scale.

where $\mathcal{P}^j_{I_l}$ is the penalty term for the jth coefficient. Assuming, as above, that $f'(U)\geq 0$, the penalty term can take the form

$$\begin{aligned}\mathcal{P}^j_{I_0} &= \tau^j\left(f(U^0(-1,t))-f(h^+(t))\right)P_j(-1)\\ \mathcal{P}^j_{I_l} &= \tau^j\left(f(U^l(x_{l-\frac{1}{2}},t))-f(U^{l-1}(x_{l-\frac{1}{2}},t))\right)P_j(-1)\quad l=1,\ldots,N\end{aligned}$$

where the τ^js may be different in each subdomain.

This formulation easily generalizes to the case where $f'(U)$ is allowed to be negative. In that case, we split the flux $f=f^+ + f^-$ into its positive ($\frac{df^+}{dU}\geq 0$) and negative ($\frac{df^-}{dU}\leq 0$) parts. For scalar hyperbolic equations, $f=f^+$ when $\frac{\partial f}{\partial U}>0$ and $f=f^-$ when $\frac{\partial f}{\partial U}<0$. For the system of Eq. (2.1), $f^\pm$ are defined as

$$f^\pm = \int A^\pm dU,$$

where A is the Jacobian matrix, i.e., $A=\frac{\partial f}{\partial U}$. The Jacobian A is then given by

$$A^\pm = T\Lambda^\pm T^{-1},$$

where T is the similarity transformation of A and Λ^+ and Λ^- are the matrices composed of the positive and negative eigenvalues respectively with $\Lambda=\Lambda^++\Lambda^-$ so that $f=f^++f^-$.

In this case we can extend the penalty terms to include terms not seen in a characteristic decomposition, so that the multidomain spectral Galerkin penalty method becomes for each interior domain,

$$\begin{aligned}\mathcal{P}^j_{I_l} &= \tau_1^j\Big(f^+(U^l(x_{l-1/2})),t))-f^+(U^{l-1}(x_{l-1/2})),t))\Big)\frac{\Delta x_k}{2}\int_\Omega Q_m^+(\xi)P_j(\xi)d\xi+\\ &\quad \tau_2^j\Big(f^-(U^l(x_{l-1/2})),t))-f^-(U^{l-1}(x_{l-1/2})),t))\Big)\frac{\Delta x_k}{2}\int_\Omega Q_m^+(\xi)P_j(\xi)d\xi+\\ &\quad \tau_3^j\Big(f^+(U^l(x_{l+1/2})),t))-f^+(U^{l+1}(x_{l+1/2})),t))\Big)\frac{\Delta x_k}{2}\int_\Omega Q_m^-(\xi)P_j(\xi)d\xi+\\ &\quad \tau_4^j\Big(f^-(U^l(x_{l+1/2})),t))-f^-(U^{l+1}(x_{l+1/2})),t))\Big)\frac{\Delta x_k}{2}\int_\Omega Q_m^-(\xi)P_j(\xi)d\xi, \quad (2.3)\end{aligned}$$

where τ^j are the penalty parameters, and the polynomials Q can be chosen as before. If $\tau_2^j=0=\tau_3^j$, the above penalty formulation is basically the characteristic decomposition. The case where $\tau_1^j=\tau_2^j=\tau_3^j=\tau_4^j$ is equivalent to no flux-splitting. Adjusting the coefficients allows for flexibility in applying this method. The major advantage of the flexibility of this formulation is in the case of the non-homogeneous grid, as we see in the following example.

Example 2.1: Consider the following equation

$$q_t+f_x=0, \quad (2.4)$$

where $q=(u,v)^T$, and $f=(v,u)$. The initial conditions are $u(x,0)=\sin(\omega\pi x)$ and $v(x,0)=0$ with $\omega=5$. With these initial conditions, the exact solutions are given by $u(x,t)=\frac{1}{2}(-\sin(\omega\pi(x-t))-\sin(\omega\pi(x+t)))$ and $v(x,t)=\frac{1}{2}(-\sin(\omega\pi(x-t))+\sin(\omega\pi(x+t)))$. This equation was previously investigated by Hu and Atkins for the dGM in [11]. The positive and negative fluxes are given by $f^+=(u+v,u+v)^T$ and $f^-=(u-v,u-v)^T$.

We can easily show that the sGM is equivalent to the dGM with the characteristic decomposition if the following penalty parameters are taken, that is, for $q_1=u$,

$$\tau_1=-\frac{1}{2}=\tau_4, \quad \text{and} \quad \tau_2=0=\tau_3,$$

and for $q_2=v$,

$$\tau_1=-\frac{1}{2}=-\tau_4, \quad \text{and} \quad \tau_2=0=\tau_3.$$

On the other hand, the case $\tau_1=\tau_2=-\frac{1}{2}$ and $\tau_3=\tau_4=-\frac{1}{2}$ are taken for $q_1=u$ and $\tau_1=\tau_2=-\frac{1}{2}$ and $\tau_3=\tau_4=\frac{1}{2}$ for $q_2=v$, there is no flux-splitting.

In our numerical example, we consider the sGM approximation with a non-homogeneous grid. A total of 50 elements are used. In the interval $x=[-1,0]$, there are 47 elements

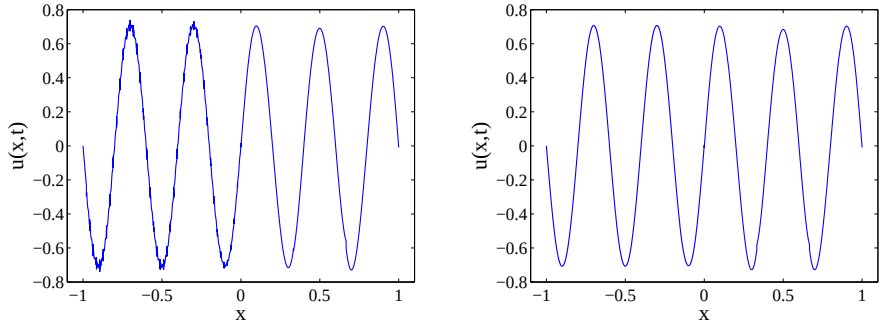

Figure 3: Solution $u(x,t)$ to Eq. (2.4) at $t=0.15$ with $m=5$ with totally 50 elements. Left: Inhomogeneous grid with no splitting penalty method. Right: Inhomogeneous grid with no splitting penalty method except 3 interface elements near $x=0$. The interface of two inhomogeneous media is at $x=0$.

each of size $\Delta x_i=\frac{1}{47},\forall i=1,\cdots,47$. In the interval $x=[0,1]$, there are 3 elements with of size $\Delta x_i=\frac{1}{3},\forall i=48,49,50$. In each element, polynomial order $m=5$ is used.

In each interval $x=[-1,0]$ and $x=[0,1]$, the solutions are smooth and there is no need to use flux splitting at all. In that case, we use the penalty method such that there is no flux-splitting, i.e. $\tau_1=\tau_2=\tau_3=\tau_4=-\frac{1}{2}$ for u and $\tau_1=\tau_2=-\frac{1}{2}=-\tau_3=-\tau_4$ for v. If there is no flux splitting, the method is easily implemented without computing the local flux conditions for the characteristic decomposition. However, the inhomogeneity of the grid raises the issue of possible non-physical reflection waves at the grid discontinuity at $x=0$. The penalty method can be easily implemented to reduce the artificial reflections at the grid discontinuity $x=0$. By adopting the characteristic flux splitting only for the domains near the grid discontinuity, one can reduce the magnitude of the non-physical reflections significantly. Furthermore, one can take an advantage of the no flux splitting method inside each homogeneous domain.

The left figure of Figure 3 shows the non-splitting penalty method at $t=0.15$. As expected, there are small fluctuations in the solution in the region $x\leq 0$. These fluctuations are the non-physical reflecting solutions reflected at $x=0$ and propagating to the left. The right figure shows the result with the penalty method where only three elements near $x=0$ and two boundary elements, use the characteristic penalty method while the other elements use the non-splitting method. For the numerical experiment, for example, I_{46},I_{47} and I_{48} are implemented based on the characteristic penalty method. Here note that the grid discontinuity $x=0$ exists between I_{47} and I_{47}. We also note that we use the characteristic penalty method for I_1 and I_{50} to minimize the boundary effects. As shown in the figures, the penalty method is efficiently flexible to adopt the grid inhomogeneity and obtains an accurate result without any considerable non-physical reflecting modes.

In [3] the same issue, but in the collocation method, was also briefly discussed.

2.2 The dGM as a special case of the the multi-domain penalty sGM

In this section, we will present the dGM and then show that it can be seen as an example of the multi-domain penalty sGM with a particular penalty term $\tau=-1$. To formulate the dGM we begin with the Galerkin form for every element I_l,

$$\sum_{k=0}^{m}\frac{db_k^l(t)}{dt}\int_{I_l}P_j(\xi(x))P_k(\xi(x))dx+\int_{I_l}P_j(\xi(x))f(U^l)_x dx=0,\qquad j=0,\cdots,m.$$

and integrate by parts to obtain,

$$\sum_{k=0}^{m}\frac{db_k^l(t)}{dt}\int_{I_l}P_j(\xi(x))P_k(\xi(x))dx-\int_{I_l}\frac{dP_j(\xi(x))}{dx}f(U^l)dx=-P_j(\xi(x))f(U^l)\,|_{x_{l-\frac{1}{2}}}^{x_{l+\frac{1}{2}}}. \quad (2.5)$$

for $\forall j=0,\cdots,m$.

As usual, we assume that $f'(U)\geq 0$ without loss of generality. Under this assumption, the boundary term in the above equation can be evaluated by replacing the left flux with the incoming flux based on the characteristic direction for $l>0$,

$$\mathcal{B}_{I_l}^j \;:=\; -P_j(\xi(x))f(U^l(x,t))\,|_{\partial I_l}=P_j(-1)f(U^{l-1}(x_{l-\frac{1}{2}},t))-P_j(1)f(U^l(x_{l+\frac{1}{2}},t)).$$

In the case $l=0$ where the left-most interval in considered, we replace the incoming value by the given boundary condition,

$$\mathcal{B}_{I_0}^j \;=\; P_j(-1)f(h^+(t))-P_j(1)f(U^l(x_{l+\frac{1}{2}},t)). \quad (2.6)$$

To see the relationship between the dGM and sGM, we define the auxiliary integrals $\mathcal{F}$ and $\mathcal{G}$

$$\mathcal{F}=\int_{I_l}\frac{dP_j(\xi(x)))}{dx}f(U^l)dx\qquad \mathcal{G}=\int_{I_l}P_j(\xi(x))f(U^l)_x dx,$$

and we note that

$$\mathcal{F}+\mathcal{G}=\int_{I_l}\frac{d}{dx}\Big(P_j(\xi(x))f(U^l(x,t))\Big)dx=P_j(1)f(U^l(x_{l+\frac{1}{2}},t))-P_j(-1)f(U^l(x_{l-\frac{1}{2}},t)). \quad (2.7)$$

With these two integrals and the penalty and boundary terms, the dGM and sGM can be rewritten, for each domain l:

$$\text{(dG)}\quad \mathbf{M}\cdot\mathbf{b}'-\mathcal{F}=\mathcal{B}_{I_l},\qquad\qquad \text{(sG)}\quad \mathbf{M}\cdot\mathbf{b}'+\mathcal{G}=\mathcal{P}_{I_l}, \quad (2.8)$$

where the mass matrix $\mathbf{M}$ and the coefficient vector $\mathbf{b}$ are defined by

$$M_{jk}=\int_{I_l}P_j(\xi(x))P_k(\xi(x))dx,\;\; b=\Big(b_0^l(t),\cdots,b_N^l(t)\Big)^T.$$

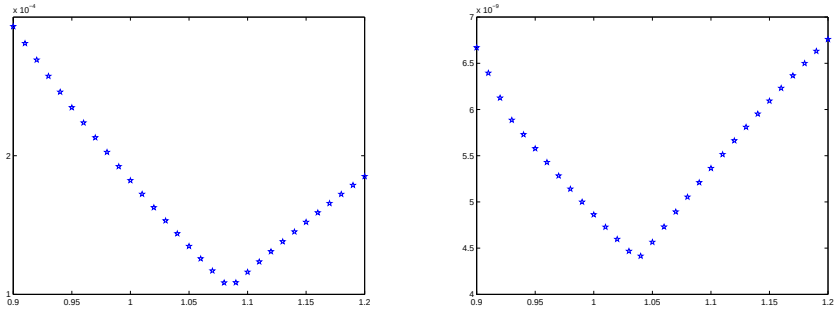

Figure 4: Example 2.2: On the horizonal axis is the value of $-\tau$ while on the vertical axis the L_2 error of the numerical solution at time $t=0.1$, with $N=10$. Left: The errors as a function of $-\tau$ for $m=3$. Right: The errors as a function of $-\tau$ for $m=6$.

To show that the dGM and sGM formulations are equivalent, we can rearrange the dG formulation Eq. (2.8) for $l>0$

$$\mathbf{M}\cdot\mathbf{b}'=\mathcal{F}+\mathcal{B}_{I_l}=P_j(1)f(U_N^l(x_{l+\frac{1}{2}},t))-P_j(-1)f(U_N^l(x_{l-\frac{1}{2}},t))-\mathcal{G}+\mathcal{B}_{I_l},$$

$$\begin{aligned}\mathbf{M}\cdot\mathbf{b}'+\mathcal{G} &= P_j(1)f(U^l(x_{l+\frac{1}{2}},t))-P_j(-1)f(U^l(x_{l-\frac{1}{2}},t))+\mathcal{B}_{I_l}\\ &= P_j(1)f(U^l(x_{l+\frac{1}{2}},t))-P_j(-1)f(U^l(x_{l-\frac{1}{2}},t))\\ &\quad +P_j(-1)f(U^{l-1}(x_{l-\frac{1}{2}},t))-P_j(1)f(U^l(x_{l+\frac{1}{2}},t))\\ &= P_j(-1)\Big(f(U^{l-1}(x_{l-\frac{1}{2}},t))-f(U^l(x_{l-\frac{1}{2}},t))\Big)\end{aligned}$$

which is equal to the penalty term $\mathcal{P}_{I_l}$ with $\tau=-1$. For the case $l=0$, the left boundary term $U^{l-1}(x_{l-\frac{1}{2}},t)$ is replaced by the boundary condition $h^+(t)$. Thus, the dGM formulation is just a special case of the penalty multidomain sGM. We saw in Figures 1 and 2 that different values of the penalty parameter yield different stability properties. Additionally, changing the value of τ will also change the size of the errors. To see this, consider the following example.

Example 2.2: Consider the sGM for $u_t+u_x=0$ with the initial condition $u(x,0)=-\sin(\pi x)$ and periodic boundary conditions. Using the multidomain sGM with $N=10$ and the weak penalty formulation, we examine the the effect of the penalty parameter τ on the errors. Figure 4 shows the L_2 errors for different τ and $m=3$(left) and $m=6$(right) at $t=0.1$. The figures show that the optimal value of τ is not $\tau=-1$, i.e. the dGM is not the optimal method within the class of sGMs. Furthermore, we observe that the optimal value of τ also depends on the polynomial order. Note that L_2 errors are $\sim 10^{-4}$ and $\sim 10^{-9}$ for $m=3$ and $m=6$, respectively.

2.3 The flexibility of the penalty parameter

One of the main advantages of the penalty sGM is that the penalty parameters can vary depending on the problem. To see how the penalty parameter affects the peformance of the evaluation of hyperbolic conservation laws, consider the following numerical example.

Example 2.3: Consider Burgers' equation

$$u_t+(u^2/2)_x=0, \quad x\in[-1,1], \quad t>0, \tag{2.9}$$

with the initial condition $u(x,0)=x$. Then the exact solution $u^e(x,t)$ is given by $u^e(x,t)=\frac{x}{1+t}$. With the given initial condition, there is no incoming boundary condition for the boundaries both at $x=-1$ and $x=1$. Also, if $x>0$, the characteristic *incoming* boundary conditions are applied at the left boundary of each element, ∂I_l^-. If $x<0$, the characteristic *incoming* boundary conditions are applied at the right boundary of each element, ∂I_l^+. Since the solution is a polynomial of degree only 1, the approximation in each element can be sufficiently resolved with $m=1$, i.e. $U=\sum_{k=0}^{1} b_k(t)P_k(x)$. However, the expansion coefficients b_k, for $\forall k>1$ are not necessarily zero in the Galerkin approximation. The weak sGM imposes the interface conditions for every mode for which the overall approximation is affected by round-off errors when the high order approximation is sought. The penalty sG formulation has more flexibility than the dG formulation to deal with such issue by exploiting the penalty parameters. Since the solution is only a polynomial of degree 1, we employ the following penalty parameters

$$\tau^j=\begin{cases} \tau & \text{if } j=0,1 \\ 0 & \text{otherwise} \end{cases}. \tag{2.10}$$

This procedure can be automated by examining the regularity of the coefficients, and then designing penalty parameters that take this into account. Figure 5 shows the sG formulation with $\tau^j=-1,\forall j=0,\cdots,m$ which is equivalent to the dGM and the modified penalty formulation with the condition of Eq. (2.10). The figures show the L_2 and L_∞ errors at $t=0.15$. We use the Legendre polynomials with the total number of elements being 3, with $m=20$, $\tau=-5$ and $CFL=0.0001$. As shown in the figure, the modified penalty method represented by $\circ$ yields the best results. We also compute the errors between the exact expansion coefficients and the approximated coefficients. The exact expansion coefficients for each domain are given in Table 1 at any time t.

Figure 6 shows the errors in the expansion coefficients for both the dGM approximations and the modified penalty approximations at $t=t_f=0.15$. The figures show that the errors of each expansion coefficient are larger than 10^{-14} for the dGM approximation for $b_k, k>1$ while they are close or below 10^{-14} for the modified penalty sGM.

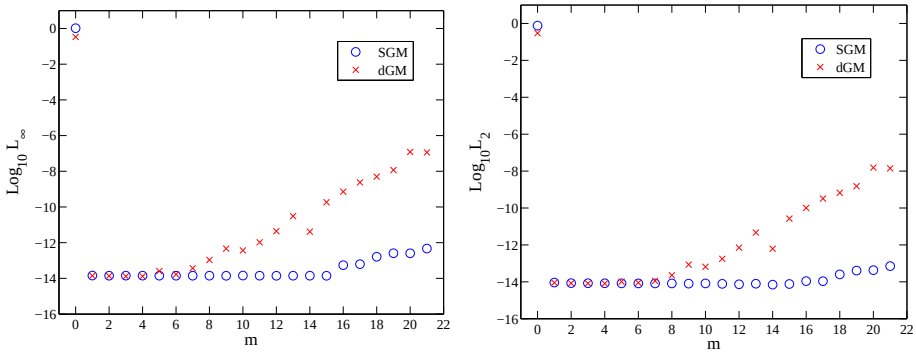

Figure 5: Example 2.3: The L_∞ (left) and L_2 (right) errors versus m for Burgers' equation. $N=3$, $m=20$, $\tau=-5$, $CFL=0.0001$, and $t_f=0.15$. The symbols $\circ$ and $\times$ represent the dGM and the modified sGM.

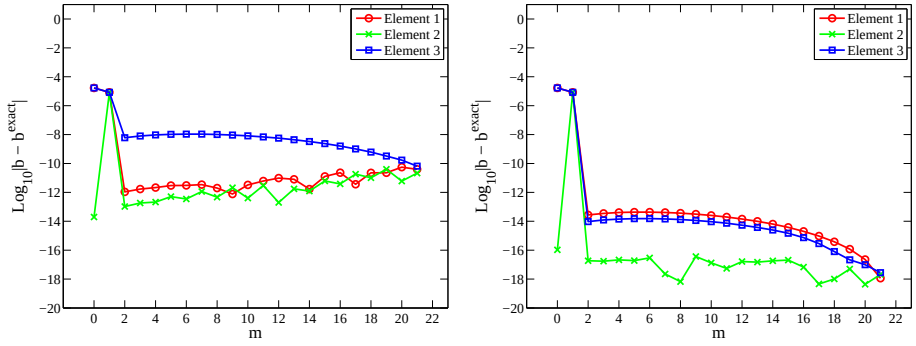

Figure 6: The errors between the exact and approximated coefficients in Example 2.3. Left: dGM approximation. Right: Modified penalty approximation. The element 1 is in $x\in[-1,-\frac{1}{3}]$, the element 2 in $x\in[-\frac{1}{3},\frac{1}{3}]$ and the element 3 in $x\in[\frac{1}{3},1]$.

Table 1: *The exact expansion coefficients for Example 2.3, for each of three domains at time t.*

	$x\in[-1,-\frac{1}{3}]$	$x\in[-\frac{1}{3},\frac{1}{3}]$	$x\in[\frac{1}{3},1]$
$b_0(t)$	$-\frac{2}{3}\frac{1}{1+t}$	0	$\frac{2}{3}\frac{1}{1+t}$
$b_1(t)$	$\frac{1}{3}\frac{1}{1+t}$	$\frac{1}{3}\frac{1}{1+t}$	$\frac{1}{3}\frac{1}{1+t}$
$b_i(t), i>1$	0	0	0

3 Roundoff errors of high order sGM

3.1 The coefficient truncation method

Although a larger polynomial order should produce a more accurate solution, in practice we see that Galerkin approximations with higher order polynomials are more sensitive to

roundoff errors. This sensitivity to roundoff errors can destroy the accuracy of a solution for large polynomial values. In this section, we present numerical examples of this problem, and suggest the coefficient truncation method to resolve it.

Example 3.1: Consider, once again, the linear advection equation $U_t+U_x=0, \quad x\in[-1,1], \quad t>0$. The solution of this equation is approximated using a dG formulation with the monomial basis function $\psi_i(x)=x^i$. Figure 7 shows the decays of the transformed vector $\mathbf{b}$ (left) and the obtained expansion coefficient vector $\mathbf{x}$ (right) for the last element which contains the the right boundary $x=1$. The parameters used for the numerical approximation are total number of domain = 5, $m=37$, $t_f=0.11$, and $CFL=0.005$.

The left figure in Figure 7 shows that the element of $\mathbf{b}$ decays with m until around $m\sim33$ with the magnitude $\sim10^{-6}$. Beyond $m>33$ it is observed not to decrease. This is due to the large condition number, κ, of the stiffness matrix $\mathbf{M}$, i.e. $\kappa\sim5.1445\times10^{17}$. The right figure in Figure 7 shows the decay or growth of the evaluated expansion coefficient vector $\mathbf{x}$. The figure clearly shows that the expansion coefficients grow rapidly with m. Figure 8 demonstrates that this results in large errors. This simple example shows that for high order polynomials, the calculation of the coefficients is very sensitive to round-off error, and this affects the accuracy of the solution.

To resolve this problem, we modify the penalty method with the truncation method introduced in [13] for use in the inverse polynomial reconstruction method (IPRM) [12,14] to reduce deterioration of the error when the polynomial error was large.

To apply the coefficient truncation method, we look at the coefficients resulting from the intermediate step in the Gaussian elimination. Suppose that the set of expansion coefficients is to be found from the linear equation $\mathbf{Mb'}=\mathbf{h}$, then the system is solved by Gaussian elimination to yield an upper triangular matrix system $\mathbf{U}\cdot\mathbf{b'}=\mathbf{c}$. The coefficients of this system should be rapidly decaying, so to reduce the problem of round-off errors we impose this requirement by setting

$$c_i=\begin{cases} c_i & \text{if } c_i>\epsilon_t \\ 0 & \text{otherwise} \end{cases} . \tag{3.1}$$

Here ϵ_t is the tolerance level which is to be determined depending on the decay rate of $\mathbf{c}$. The motivation behind this method is that, as shown in Figure 1, the spectral radius ρ increases with m for the Galerkin method and this makes the method sensitive to round-off errors. The truncation method tries to reduce the magnitude of ρ by truncating the RHS of the linear system of the Galerkin method by reducing the rank of the matrix. For example, if the truncation order is n, then the rank of the truncated matrix becomes $m+1-n$. With the rank reduced, ρ is also reduced.

Example 3.2: To demonstrate the success of the truncation method in reducing the effects of round-off errors, we return to Example 3.1 above. We apply the truncation method with various tolerence levels, i.e. $\epsilon_t=10^{-14}(+)$, $\epsilon=10^{-12}(\circ)$, $\epsilon=10^{-10}(\square)$, and $\epsilon=10^{-6}(\nabla)$. The left figure in Figure 7 shows that the the elements of $\mathbf{b}$ decay faster when the truncation error is applied for different values of ϵ. The right figure in Figure 7 shows the evaluated expansion coefficient vector $\mathbf{x}$. The figure clearly shows that the expansion

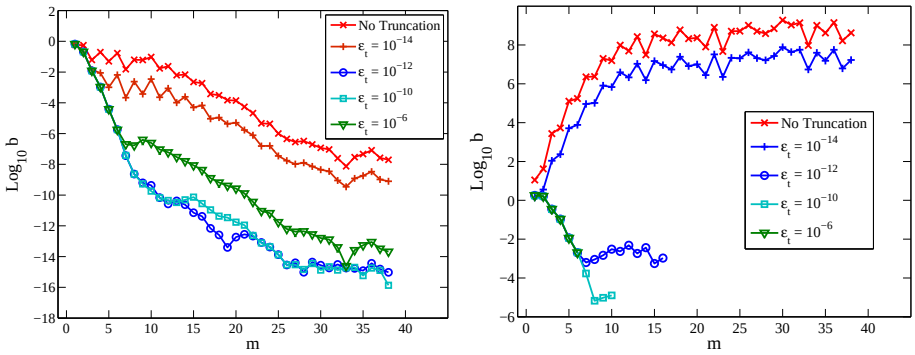

Figure 7: Examples 3.1 and 3.2: Left: The decay of the elements of the load vector $\mathbf{c}$ with and without the truncation method. Right: The decay of the obtained expansion coefficient vector $\mathbf{x}$ with and without the truncation method. The dGM is used for the advection equation with monomial basis functions. Total number of domain is 5, $m=37$, $t_f=0.11$, and $CFL=0.005$. Various tolerance levels ϵ_t are used, i.e. $\epsilon_t=10^{-14},10^{-12},10^{-10},10^{-6}$.

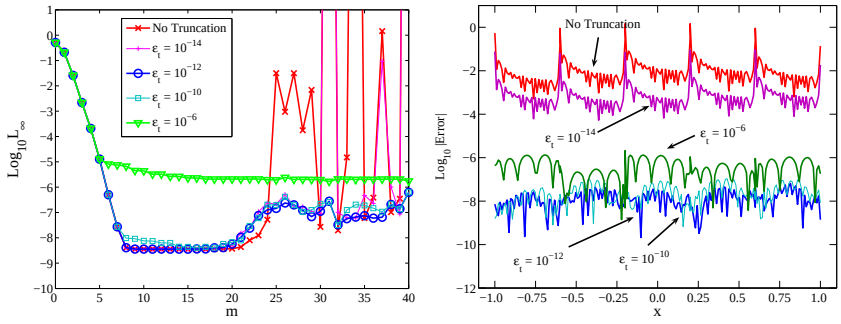

Figure 8: Examples 3.1 and 3.2: Left: The L_∞ errors versus m. Right: The pointwise errors when $m=37$ is used with and without the truncation method. The dGM is used for the advection equation with monomial basis functions. Total number of domain is 5, $t_f=0.11$, and $CFL=0.005$. Various tolerance levels ϵ_t are used, i.e. $\epsilon_t=10^{-14},10^{-12},10^{-10},10^{-6}$.

coefficients grow with m exponentially without the truncation method while they decay if the truncation method is applied. The figures also shows that the expansion coefficients become truncated to almost machine zeros or very close to zeros if the truncation method is applied. In the figure, these truncated values are not displayed as the plot uses the logarithmic scale. The larger ϵ_t is used, the faster the truncation of $\mathbf{x}$ occurs.

The L_∞ errors with m (left) and the pointwise errors with $m=37$ (right) are given in Figure 8 with the same symbols as in Figure 7. The left figure in Figure 8 shows the L_∞ decay or growth with m with and without the truncation method. The figure shows that the L_∞ errors decay up to $m\sim7$ and they start to grow beyond $m\sim20$ if the truncation method is not applied or the truncation method is applied with $\epsilon_t=10^{-14}$. The truncation method with $\epsilon=10^{-10}$ or 10^{-12} yields the best results up to $m\sim40$. If $\epsilon_t=10^{-6}$ is used,

the L_∞ remains around 10^{-6} for large m. It is observed that for $m>40$, the L_∞ errors remains around 10^{-6} when $\epsilon_t=10^{-6}$ while the other cases result in the growth of L_∞ errors. The right figure in Figure 8 shows the pointwise errors with $m=37$. The figure shows that the best result is obtained when $\epsilon_t=10^{-12}$ is used.

Figure 8 suggests that the tolerance level used with the truncation method can be chosen by observing the decay rate of $\mathbf{c}$ for a given m. Also, one can use the different tolerance level ϵ_t for different element since each element has different decay rate of $\mathbf{c}$.

3.2 Numerical consistency

Another factor which increases the sensitivity to round-off errors is the inconsistent evaluation of the integrals in the formulation of the system of ODEs for the approximation coefficients. On the Galerkin formulationwe need to compute two sets of integrals, $\mathcal{I}_1=\int_\Omega P_k(x)P_j(x)dx$ and $\mathcal{I}_2=\int_\Omega -f(U)_xP_j(x)dx$. While $\mathcal{I}_1$ can be evaluated exactly, because the form of the polynomials is known explicitly in general, $\mathcal{I}_2$ can not generally be evaluated, and so quadrature must be used. We define the integral operators $\mathcal{E}$ and $\mathcal{Q}$, where $\mathcal{E}$ is the exact integration operator $\mathcal{E}(g)=\int_\Omega g(x)$, and $\mathcal{Q}$ some quadrature rule such as Gauss quadrature rule used to evaluate the given integral $\mathcal{Q}(g)=\sum_{i=0}^{M} g_i\omega_i \sim \int_\Omega g$, where ω_i are the weights. The Galerkin formulation can now be performed in one of four ways:

$$\begin{aligned} &\mathcal{E}(P_lP_j)=\mathcal{E}(f(U)_xP_j)+\mathcal{P}_l, \qquad &&\mathcal{E}(P_lP_j)=\mathcal{Q}(f(U)_xP_j)+\mathcal{P}_l, \\ &\mathcal{Q}(P_lP_j)=\mathcal{E}(f(U)_xP_j)+\mathcal{P}_l, \qquad &&\mathcal{Q}(P_lP_j)=\mathcal{Q}(f(U)_xP_j)+\mathcal{P}_l. \end{aligned} \tag{3.2}$$

We say that the Galerkin formulation is *numerically consistent* if the integrals are evaluated in the same way on both sides of the equations (both by $\mathcal{E}$ or both by $\mathcal{Q}$), and it is *numerically inconsistent* if the two integrals are evaluated differently on the two sides of the equation.

The phenomenon we are considering is essentially a numerical one. For high polynomial order, the numerically consistent formulation yields more accurate results, and lower sensitivity to roundoff errors, than the inconsistent formulation. Consequently the above 4 different formulations, Eq. (3.2), show different errors.

Example 3.3: We first consider the following steady-state problem

$$u_t+u_x=\sin(\pi x), \quad x\in[-1,1], t>0,$$

with initial condition $u(x,0)=u^s(x)+\epsilon_M\delta(x)$, where ϵ_M is machine accuracy, $\epsilon_M=10^{-16}$ and $\delta(x)$ is the random function, $0\le\delta(x)\le 1$ with the normal distribution. The boundary condition is $u(-1,t)=-\frac{1}{\pi}\cos(-\pi), \quad \forall t>0$. The steady-state solution, $u^s=\lim_{t\to\infty}u(x,t)=u^s(x)$ for $t\to\infty$ is then given by $u^s(x,t)=-\frac{1}{\pi}\cos(\pi(x-t))$. For the numerical experiment, the final time is $t_f=1$, the total number of domain $N=30$, the penalty parameter $\tau=-5$, and the CFL number $CFL=0.001$ for $dt=CFL\times dx=6.6667\times 10^{-5}$ are used where each element has the same element size and polynomial order.

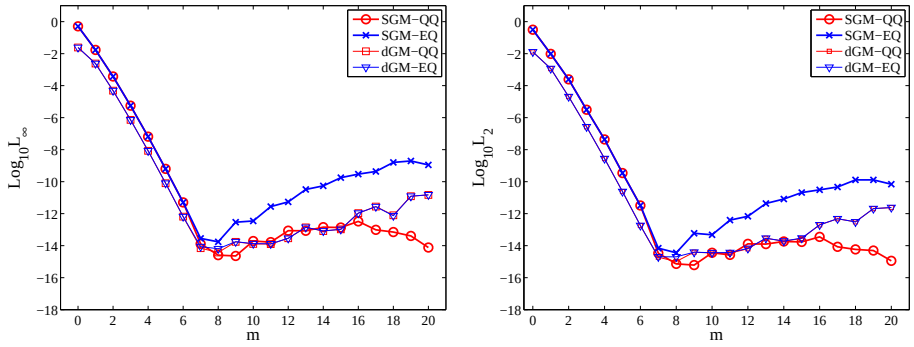

Figure 9: Left: L_∞ errors vs m for the steady-state problem $u_t + u_x = \sin(\pi x)$. The symbols $\circ$ and $\times$ denote the quadrature-quadrature (QQ) and exact-quadrature (EQ) formulations for the penalty sGM and $\square$ and ∇ the QQ and EQ formulations for the dGM. The final time $t_f = 1$ and the total number of domain is 30. The penalty paramter is $\tau = -5$ and the CFL number is $CFL = 0.001$, i.e. $dt = 6.6667 \times 10^{-5}$.

Figure 9 shows the convergence of L_2 and L_∞ errors with m for the the weak penalty formulation and the dG formulation. The LHS or RHS are evaluated either using the quadrature rules labeled by Q or the exact formula denotedlabeled by E for the penalty formulation with $\tau = -5$. As the figures show, both the L_∞ and L_2 errors decay exponentially until around $m \sim 7$. For $m > 7$, these errors grow slightly due to round-off errors. The figures show that the results with the consistent evaluation of the stiffness matrix of the sGM (QQ) show better performance for the weak penalty sGM when round-off error become dominant. We note that although the dGM does not suffer from an inconsistent formulation, the consistent formulation for the penalty sGM yields the best results when m is large.

4 Summary and conclusion

We describe the formulation of the multi-domain penalty sGM and demonstrated that the flexibility of the penalty formulation can be advantageous, because it allows us to tailor the penalty parameter to match the problem. This is especially relevant in the case where the sub-domains have different mesh size or polynomial order, as the flexibility in the penalty parameters can allow us to avoid costly flux splitting except near the grid discontinuity. However, different values of the penalty parameter effect the stability, accuracy, and sensitivity to round-off errors. For example, the dGM is simply a sGM with a particular choice of penalty parameter, which for a linear wave equation has nice stability properties, though it is not optimal in terms of accuracy.

We presented two numerical issues which arise in the solution of hyperbolic conservation laws using the multi-domain penalty sGM with high order polynomials. The first is the sensitivity of high order sGMs to roundoff errors, which can potentially ruin the accu-

racy of the solution. To resolve this issue we introduce the coefficient truncation method which prevents the rapid growth of the errors with m, and has a stabilizing effect on the method. The second numerical issue is that consistent evaluation of the stiffness matrix and the load vector yields a better result when the high order polynomials are used with the penalty formulation. This sensitivity, too, depends on the penalty parameter, and we note that the case of $\tau = -1$, (the dGM case) does not seem affected by this numerical consistency issue.

Future studies will center around methods for choosing the penalty parameter to optimize for accuracy and stability, as well as further development of the coefficient truncation method for multi-dimensional problems.

5 Acknowledgements

The authors gratefully acknowledges the advice of David Gottlieb and Jennifer Ryan. This work has been supported by the NSF under Grant No. DMS-0608844.

References

[1] B. Cockburn, C. Johnson, C.-W. Shu, E. Tadmor, Advanced numerical approximation of nonlinear hyperbolic equations, Lecture Notes in Mathematics, Springer, Berlin, 1999.

[2] B. Cockburn, G. Karniadakis, C.-W. Shu, The development of discontinuous Galerkin methods, in: *Discontinuous Galerkin methods* (Newport, RI, 1999), Lecture Notes in Computer Science and Engineering, vol. 11, Springer, Berlin, 2000 pp. 3-50.

[3] W.-S. Don, D. Gottlieb, J.-H. Jung, A multidomain spectral method for supersonic reactive flows, J. Comput. Phys. 192 (2003) 325-354.

[4] D. Funaro, D.Gottlieb, A New Method of Imposing Boundary Conditions for Hyperbolic Equations, Math. Comput. 51 (1988) 599-613.

[5] D. Funaro, D. Gottlieb, Convergence Results for Pseudospectral Approximations of Hyperbolic Systems by a Penalty-type Boundary Treatment, Math. Comput. 57 (1991) 585-596.

[6] J. S. Hesthaven , D. Gottlieb, A Stable Penalty Method for the Compressible Navier–Stokes Equations: I. Open Boundary Conditions, SIAM Journal on Scientific Computing 17 (1996) 579-612.

[7] J. S. Hesthaven, A Stable Penalty Method for the Compressible Navier–Stokes Equations: II. One-Dimensional Domain Decomposition Schemes, SIAM Journal on Scientific Computing, 18 (1997) 658 - 685.

[8] J. S. Hesthaven, S. Gottlieb, D. Gottlieb, Spectral methods for time-dependent problems (Cambridge, 2007), Cambridge UP.

[9] J. S. Hesthaven and T. Warburton, 2002, High-Order Nodal Methods on Unstructured Grids. I. Time-Domain Solution of Maxwell's Equations, J. Comput. Phys. 181(1), 186-221.

[10] F. X. Giraldo, J. S. Hesthaven, and T. Warburton, 2002, Nodal High-Order Discontinuous Galerkin Method for the Spherical Shallow Water Equations, J. Comput. Phys. 181(2), 499-525.

[11] F. Q. Hu , H. L. Atkins, Eigensolution analysis of the discontinuous Galerkin method with nonuniform grids: I. one space dimension, J. Comput. Phy. 182 (2002) 516-545.

[12] J.-H. Jung, B. D. Shizgal, Generalization of the inverse polynomial reconstruction method in the resolution of the Gibbs phenomena, J. Comput. Appl. Math. 172 (2004) 131-151.
[13] J.-H. Jung, B. D. Shizgal, On the numerical convergence with the inverse polynomial reconstruction method for the resolution of the Gibbs phenomenon, J. Comput. Phys. 224 (2007) 477-488.
[14] B. D. Shizgal, J.-H Jung, Towards the resolution of the Gibbs phenomena, J. Comput. Appl. Math. 161 (2003) 41-65.