Stabilized Finite Element Methods

Finn Joel Bjervig*and John Danielsson†

1TD050 - Advanced Numerical Methods,
Department of Information Technology,
Uppsala university, Sweden

August 31, 2022

*Electronic address: jobj8920@student.uu.se
†Electronic address: joda2016@student.uu.se

The PDE of interest is a linear advection equation

$$\partial_t u + \nabla f(u) = 0 \qquad (x,t) \in \Omega \times (0,T] \qquad f \in \mathcal{C}^1(\mathcal{R}, \mathcal{R}^2) \tag{1a}$$

$$u(x,0) = 0 \tag{1b}$$

$$u(t) = 0 \qquad u \in \partial\Omega \tag{1c}$$

To derive Galerkin Finite Element Method we need to define a relevant vector space. Weakform is not needed because the PDE has no higher order derivatives. We continue with the discrete formulation.

Define the vector space:

$$V_0 = \{v : v \in \mathcal{H}(\Omega)^1 \qquad v = 0 \text{ on } \partial\Omega\} \tag{2}$$

Weak form reads:

Find $u \in V_0$ such that

$$\big(\partial_t u + \nabla f(u), v\big) = 0 \qquad \forall v \in V_0 \tag{3}$$

we continue to define another vector space for the discrete approximation u_h

$$V_{h,0} = \{v(x,t) : v(x,t) \in C^0(\tau_h), \forall t \in (0,T], v|_k \in P_1(k),\ \forall k \in \tau_h \\ v = 0 \text{ on } \partial\Omega\} \tag{4}$$

K denotes the triangular elements of which the mesh consists of.

$$\tau_h = \{K\} \text{ , K elements, } \ h_k = diam(K) \text{ - meshsize} \tag{5}$$

The set of all nodes $\{N\}$ on τ_h is the sum of all internal nodes and boundary nodes as $\{N\} = \{N_h\} + \{N_b\}$. The GFEM formulation reads:

Find $u_h \in V_h$ such that

$$\big(\partial_t u_h + \nabla f(u_h), v\big) = 0 \qquad \forall v \in V_{h,0} \tag{6}$$

Since $u_h \in V_h$, $\exists\{\xi\}_{N_j \in N_h}$, such that

$$u_h(x,t) = \sum_{N_j \in N_h} \xi_j(t)\varphi_j(x) \tag{7}$$

where $\varphi_j(x)$ are the appropriate hat functions.

Before continuing, we express the gradient acting on the flux in the non-conservative form as

$$\nabla f(u) = f'(u) \cdot \nabla u$$

With that said, the above expression for u_h is substituted into 6 and the test function v is chosen as another hat function, and we yield

$$\sum_{N_j \in N_h} \partial_t \xi_j(\varphi_j, \varphi_i) + \sum_{N_j \in N_h} (f'(\xi_j) \cdot \nabla(\xi_j \varphi_j), \varphi_i) = 0 \tag{8}$$

From which we can define the matrices presented in problem 1.1

Applying Crank Nicholson time discretization yields

for $i = 1, 2, \ldots \Pi$, find $\xi_i \in V_{h,0}$ such that

$$\frac{1}{k_n}(\xi_i - \xi_{i-1}, v) + \frac{1}{2}\left(\nabla(f(\xi_{i-1}) + f(\xi_i)), v\right) = 0 \qquad \forall v \in V_h \tag{9}$$

k_n is a sufficiently small timestep. For the project $k_n = CFL\frac{h_{max}}{\left\|f'(\xi_i)\right\|_{L^\infty(\Omega)}}$. The CFL is set to 0.5.

Problem 1.1

Using

$$M_{i,j} = (\varphi_i, \varphi_j) \tag{10a}$$
$$C_{i,j} = (\mathbf{f}'(u) \cdot \nabla\varphi_i, \varphi_j) \tag{10b}$$
$$A_{i,j} = (\nabla\varphi_i, \nabla\varphi_j) \tag{10c}$$

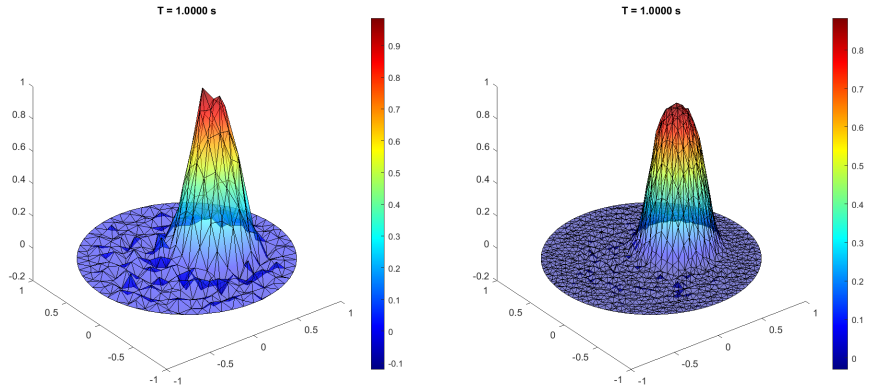
we can turn equation 9 to the following linear system of equations

$$(\frac{M}{k_n} + \frac{C}{2})\xi_{n+1} = (\frac{M}{k_n} - \frac{C}{2})\xi_n \tag{11}$$

With initial conditions being

$$u_0(\mathbf{x}) = \frac{1}{2}(1 - \tanh(\frac{(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2}{r_o^2} - 1)) \tag{12}$$

The solutions at t = 0 are using max mesh sizes of 1/8 and 1/16 are shown in figure 1. $r_0 = 0.25$, $(x_1^0, x_2^0) = (0.3, 0)$



(a) With max mesh size = 1/8

(b) With max mesh size = 1/16

Figure 1: Solutions at T = 1 at different mesh sizes with the continuous initial condition

Problem 1.2

We now compute the L^2 norm of the error for the error using mesh sizes of $h_max = [\frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{32}]$. The error $u - U_h$ can be computed knowing $u(\mathbf{x}, t = 1) = u_0(\mathbf{x})$.

The convergence rate α can be found by using

$$\text{L}^2\text{-norm} \approx O(h^\alpha) \Rightarrow \log(\text{L}^2\text{-norm}) \approx \alpha \log(h) \tag{13}$$

and we can use the Matlab function for polynomial curve fitting *polyfit()* to find a linear interpolation of α, which is shown in figure 2

The convergence rate α can then be seen to be around 1.75.

Problem 1.3

Now we repeat problems 1.1 and 1.2 using the discontinuous initial conditions

$$y = \begin{cases} 1 & \text{if } (x_1 - x_1^0)^2 + (x_2 - x_2^0)^2 \leq r_o^2 \\ 0 & \text{otherwise} \end{cases} \tag{14}$$

using $(x_1, x_2) = (0.3, 0)$, $r_0 = 0.25$. The solutions for meshes 1/8 and 1/16 can be seen in figure 3

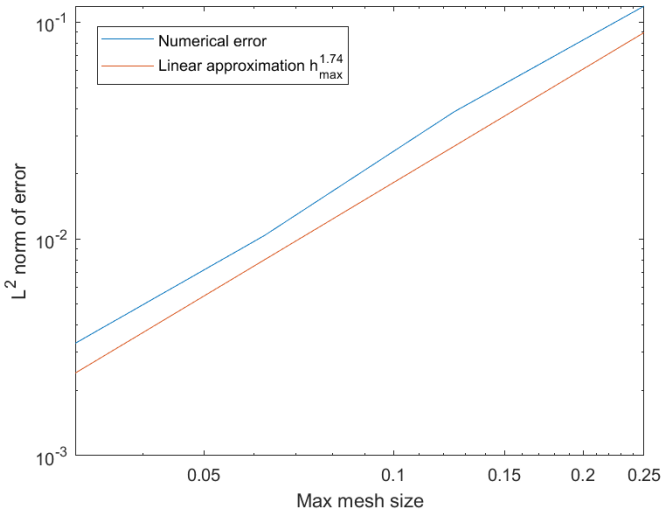
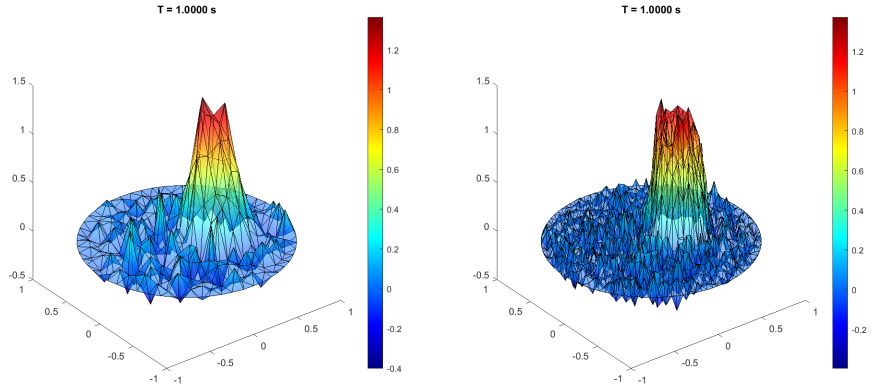


Figure 2: Loglog plot of error vs mesh size



(a) With max mesh size = 1/8

(b) With max mesh size = 1/16

Figure 3: Solutions at T = 1 at different mesh sizes with the discontinuous initial condition

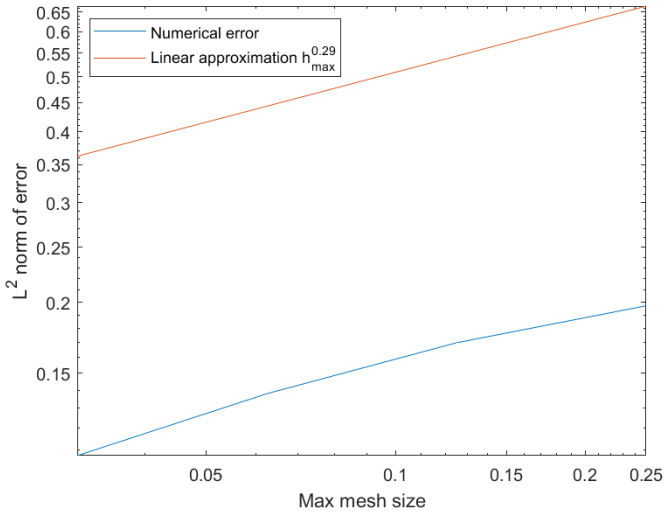


Figure 4: Loglog plot of error vs mesh size

It can be seen that the solution is clearly worse with more turbulent behavior occurring throughout the space where the solution should be zero. The convergence rate at around 0.32 is also much lower than for the continuous problem.

1 Part 2. Stabilized FEM

Problem 2.1

Another way to stabilize the GFEM method is to add an stabilization term. Least squares method combined with the GFEM method results in a stabilized FEM method and is called Galerkin Least Squares method, abbreviated as GLS.

The GLS method for 1a is formulated as

Find $u_h \in V_{h,0}$ such that:

$$(Au_h, v) + \delta(Au_h, Av) = 0 \qquad \forall v \in V_{h,0} \tag{15}$$

$$A = (\partial_t + \beta\nabla) \tag{16}$$

$$V_h = \{v(x,t) : v(x,t) \in C^0(\tau_h), \forall t \in (0,T], v|_k \in P_1(k), \forall k \in \tau_h, \\ v(t) = 0 \text{ at } \partial\Omega\} \quad (17)$$

Chose $v = u_h$ as testfunction and expand the operator A

$$(\partial_t u_h + \beta\nabla u_h, u_h) + \delta(\partial_t u_h + \beta\nabla u_h, \partial_t u_h + \beta\nabla u_h) = \\ (\partial_t u_h, u_h) + (\beta\nabla u_h, u_h) + \delta\|\partial_t u_h + \beta\nabla u_h\|^2 = 0 \quad (18)$$

The first term on the RHS can be written as

$$(\partial_t u_h, u_h) = \int_\Omega \partial_t u_h u_h dx = \int_\Omega \frac{1}{2}\frac{\partial}{\partial t}(u_h^2)dx = \frac{1}{2}\partial_t\|u_h\|^2$$

concerning the next term, β is divergence free and thus $\nabla \cdot \beta = 0$, making the second term zero. The last term is the residual $R(u_h)$. With this, the GLS formulation becomes

$$\frac{1}{2}\partial_t\|u_h\|^2 + \delta\|\partial_t u_h + \beta\nabla u_h\|^2 = 0 \quad (19)$$

We attempt to integrate in time

$$\int_0^T \partial_t\|u_h\|^2 + \delta\|\partial_t u_h + \beta\nabla u_h\|^2 dt = 0$$
$$\implies \left\|u_h(t=T)\right\|^2 + \delta\int_0^T \left\|\partial_t u_h(\cdot,t) + \beta\nabla u_h(\cdot,t)\right\|^2 dt = \left\|u_{h,0}\right\|^2$$
$$\implies \left\|u_h(T)\right\|^2 + \delta\int_0^T \left\|R(u_h)\right\|^2 dt = \left\|u_h(t=0)^2\right\| \quad (20)$$

The residual can be approximated to

$$R_h(u_h) = \partial_t u_h + \beta\nabla u_h \approx \frac{u_k - u_{k-1}}{\Delta t} + \beta\nabla u_h \quad (21)$$

where the second term does not live in V_h.

$$\beta\nabla u_h \notin V_h$$

For this reason we need to find a $R_h(u_h) \in V_h$ using the L_2 projection such that

$$\begin{aligned}(R_h, v) &= \left(\frac{u_k - u_{k-1}}{\Delta t} + \beta \cdot \nabla u_k, v\right) \\ &= \frac{1}{\Delta t}\left[(u_k, v) - (u_{k-1}, v)\right] + (\beta \cdot \nabla u_k, v) \end{aligned} \tag{22}$$

set the test function v to be a hat function $v = \phi_i$ and construct the solution u_h as a linear combination of hat functions

$$u_k = \sum_j \xi_j^k \phi_j \qquad\qquad u_{k-1} = \sum_j \xi_j^{k-1} \phi_j$$

substitute these definitions into 22 to yield

$$\frac{1}{\Delta t}\left[\sum_j (\phi_j^k, \phi_i) - \sum_j (\phi_j^{k-1}, \phi_i)\right] + (\beta \sum_j \nabla \phi_j^k, \phi_i) \tag{23}$$

Where the stiffness and convection matrices are realized

$$M\eta = \frac{1}{\Delta t}(M\xi^k - M\xi^{k-1}) + A\xi^k \tag{24}$$

and solve for η.

Back to formulating the GLS method for implementation, we have:

$$(\partial_t u, v) + (\beta \nabla u, v) + \delta(\partial_t u, \beta \nabla v) + (\beta \nabla u, \beta \nabla v) = 0 \tag{25}$$

Using Crank Nicholson for time discretization and assembling the matrices we get

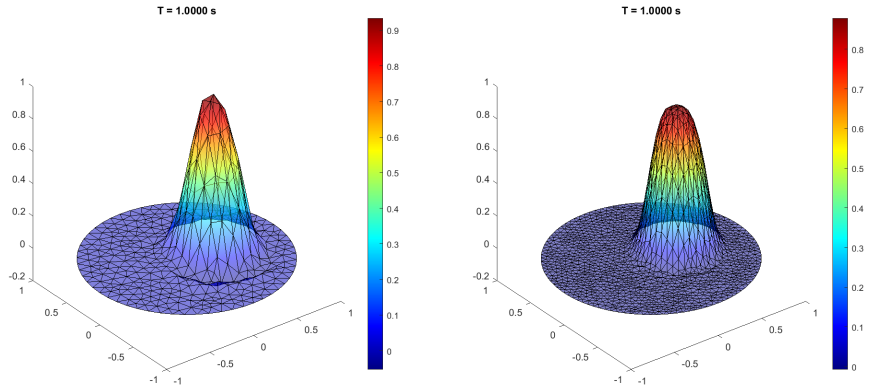
$$\frac{M}{k_n}(U_n - U_{n-1}) + \frac{\delta C^T}{k_n}(U_n - U_{n-1}) + \frac{C}{2}(U_n + U_{n-1}) + \frac{A_\beta}{2}(U_n + U_{n-1}) = 0 \tag{26}$$

Factor into U_n and U_{n-1}

$$U_n\left(\frac{M}{k_n} + \frac{\delta C^T}{k_n} + \frac{C}{2} + \frac{A_\beta}{2}\right) = U_{n-1}\left(\frac{M}{k_n} + \frac{\delta C^T}{k_n} - \frac{C}{2} - \frac{A_\beta}{2}\right) \tag{27}$$

solve for U_n for each time step incremented by k_n with a optimized matrix solver such as pseudo-inverse in Matlab.

Problem 2.2



(a) GLS method with max mesh size = 1/8 (b) GLS method with max mesh size = 1/16

Figure 5: Solutions for Galerkin-Least-Squares at T = 1 at different mesh sizes with continuous initial condition

The GLS method applied to 1a with the continuous initial condition 12 converges with increasing mesh resolution as seen in 5, and converges to the right solution with a rate of $\alpha_{GLS} = 1.91$, seen in 6, which is within a reasonable margin to the theoretical rate $\alpha_{GLS} = 2$. Contrary to the GFEM, GLS computes a solution with significantly less fluctuations, thanks to the stabilization term, especially in the "wake" of the pulse.

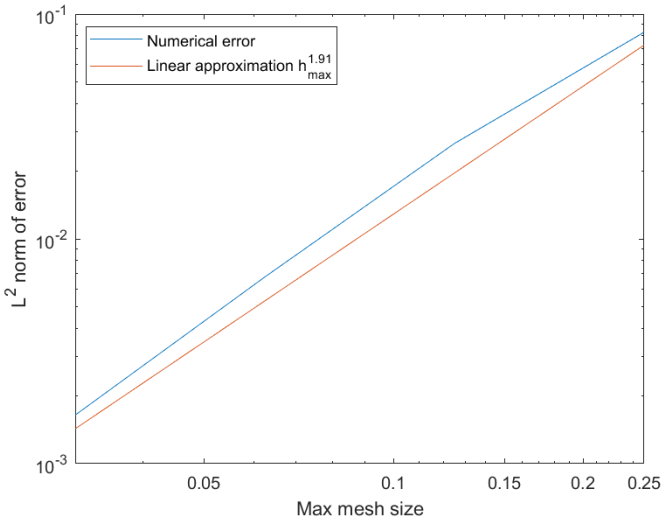
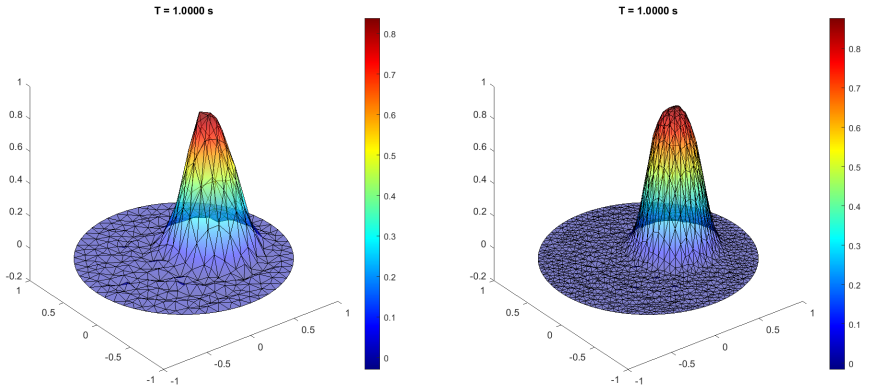


Figure 6: Loglog plot of error at $T = 1$ (one period) vs mesh size for GLS with continous initial condition

Residual Viscosity also converges with increased mesh resolution and practically has a convergence rate of at $\alpha_{RV} = 1.96\ 8$, practically the same as GLS. It also agrees with the theoretical value of $\alpha_{RV} = 2$.



(a) RV method with max mesh size = 1/8 (b) RV method with max mesh size = 1/16

Figure 7: Solutions for Artificial Residual Viscosity method at T = 1 at different mesh sizes with continuous initial condition

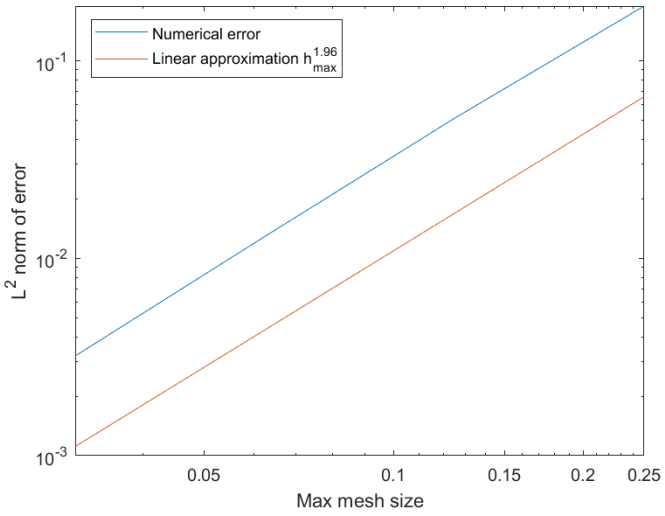
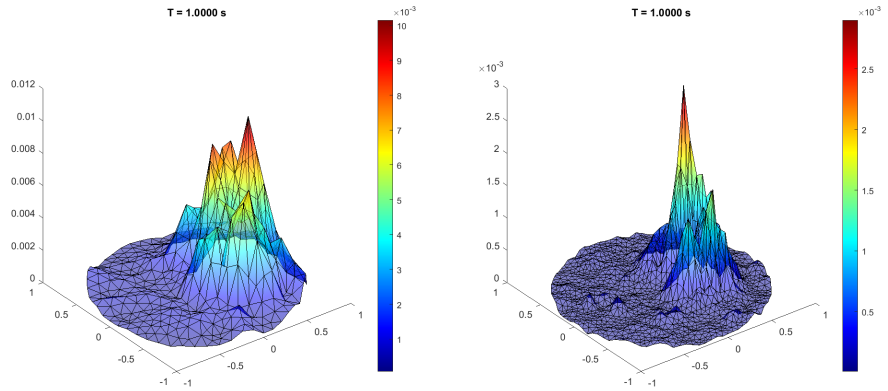


Figure 8: Loglog plot of error vs mesh size for RV with continuous initial condition



(a) Residual of RV method with max mesh size = 1/8

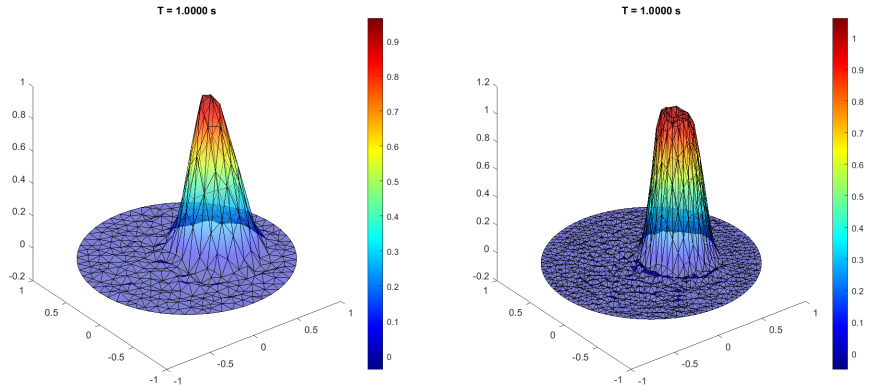
(b) Residual of RV method with max mesh size = 1/16

Figure 9: Computed residuals for Artificial Residual Viscosity method at T = 1 at different mesh sizes with continuous initial condition

For the artificial residual viscosity the computed residuals can also be plotted, as done in figure 9. Here we can see the residual is largest around where the solution peak is located, but is also present over the entire domain to dampen fluctuations.

Problem 2.3

Discontinuous features in initial conditions as in 14 is somewhat problematic for respective methods that incorporates FEM in some way, as the gradient term ∇u_h of the residual 21 becomes very large at the discontinuity. Convergence is therefore slower since a finer mesh doesn't resolve this issue. Actually, when mesh size is smaller, the gradient will be even larger. The solution from the GLS method 12 has a convergence rate of $\alpha_{GLS} = 0.43$ 13 and similarly RV 10 has $\alpha_{RV} = 0.49$, see 11. In spite of the lower convergence rate and higher error, both methods does perform significantly better than regular GFEM 3. The residual viscosity method does compute a smoother solution, but all the while, its error is similar to GLS in convergence and absolute error.



(a) RV method with max mesh size = 1/8 (b) RV method with max mesh size = 1/16

Figure 10: Solutions for Artificial Residual Viscosity method at T = 1 at different mesh sizes with discontinuous initial condition

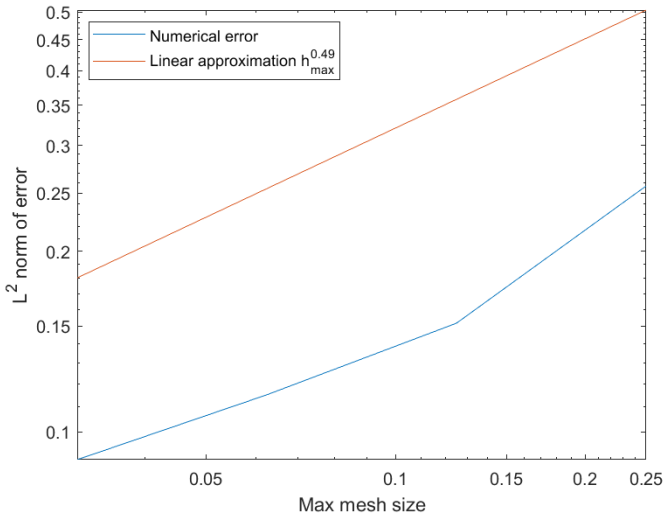
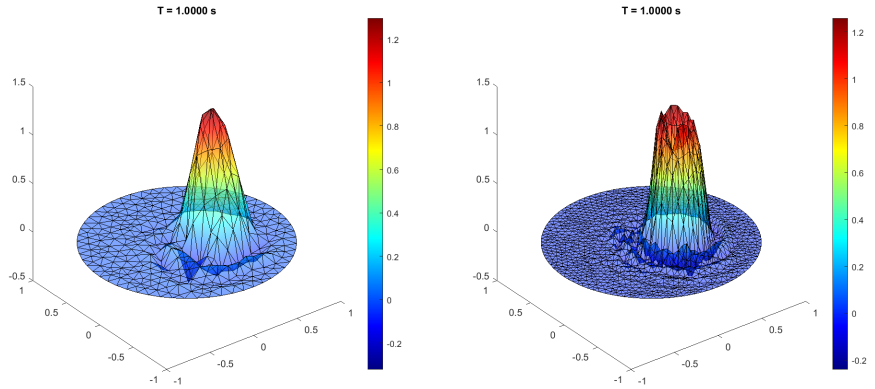


Figure 11: Loglog plot of error vs mesh size for RV with discontinuous initial condition



(a) GLS method with max mesh size = 1/8 (b) GLS method with max mesh size = 1/16

Figure 12: Solutions for Galerkin Least Squares method at T = 1 at different mesh sizes with discontinuous initial condition

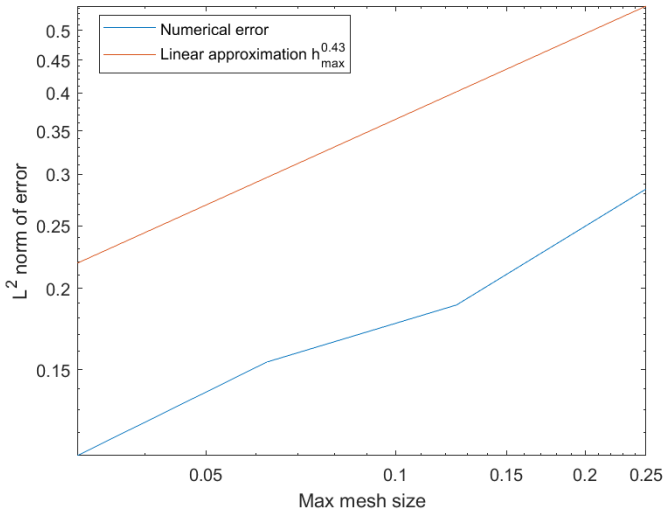


Figure 13: Loglog plot of error vs mesh size for GLS with discontinuous initial condition