DNP with computable mesh conditions for parabolic standard diffusion with nonlinear reaction FE PDEs

Menghis T. Bahlibi

Supervisors: János Karátson & Ferenc Izsák

Institute of Mathematics
Faculty of Science, Eötvös Loránd University

May 22, 2025

Outline of the talk

1. Model problem
2. Motivations and earlier results
3. Discrete non-negativity preservation (DNP)
4. Theoretical and experimental results
5. Conclusion

Model problem

Parabolic standard diffusion with a nonlinear reaction:

$$\begin{cases} \partial_t u(t,\mathbf{x}) - \mu_0 \Delta u(t,\mathbf{x}) + q(u(t,\mathbf{x})) = f(t,\mathbf{x}), \quad \mathbf{x} \in \Omega,\ t > 0, \\ u(t,\mathbf{x}_1) = 0, \quad \mathbf{x}_1 \in \partial\Omega,\ t > 0, \\ u(0,\mathbf{x}) = u_0(\mathbf{x}), \quad \mathbf{x} \in \Omega, \end{cases} \tag{1}$$

where

- $\Omega \subset \mathbf{R}^2$,
- $\mu_0 \in \mathbf{R}^+$ is a diffusion coefficient,
- $q \in C^1(\mathbf{R})$, $q(0) = 0$, and f is a source function.

Assume that $q(u) = r(u)u$, and there exists $\sigma_1 > 0$ such that for all $\xi \in \mathbf{R}$

$$0 \le \frac{\partial q}{\partial \xi}(\xi) \le \sigma_1, \text{ thus, } 0 \le r(\xi) \le \sigma_1. \tag{2}$$

Finite element approximation

To find the FE approximation in space for the model problem, consider a FE subspace V_h of first-order elements with the following assumptions:

(B1) $0 \leq \phi_i \leq 1, \quad (\forall i = 1, \ldots, n),$

(B2) $\sum_{i=1}^{n} \phi_i \equiv 1,$

(B3) $\phi_i(P_j) = \delta_{ij}$ for proper nodes $P_1, \ldots, P_n \in \Omega$.

- There exists a constant $\sigma_0 > 0$, independent of h, such that for any $i \neq j$ for which the interior of $\Omega_{ij} := supp\, \phi_i \cap supp\, \phi_j$ is nonempty, then the basis functions satisfy

$$\int_{\Omega_{ij}} \nabla \phi_i \cdot \nabla \phi_j \leq -\frac{\sigma_0}{h^2}\, meas(\Omega_{ij}). \tag{3}$$

Full discretizations with four versions of time-stepping

Using the matrices S(stiffness), M(mass), and matrix N obtained from nonlinear term, where δ is the time step, the full discretizations of (1):

(i) Implicit

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^{k}}{\delta}+S\mathbf{u}^{k+1}+N(\mathbf{u}^{k+1})\mathbf{u}^{k+1}=\underline{\mathbf{f}}^{k}, \tag{4}$$

(ii) Semi–implicit

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^{k}}{\delta}+S\mathbf{u}^{k+1}+N(\mathbf{u}^{k})\mathbf{u}^{k+1}=\underline{\mathbf{f}}^{k}, \tag{5}$$

(iii) Linearly implicit

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^{k}}{\delta}+S\mathbf{u}^{k+1}+N(\mathbf{u}^{k})\mathbf{u}^{k}=\underline{\mathbf{f}}^{k}, \tag{6}$$

(iv) Explicit

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^{k}}{\delta}+S\mathbf{u}^{k}+N(\mathbf{u}^{k})\mathbf{u}^{k}=\underline{\mathbf{f}}^{k}. \tag{7}$$

Full discretizations

The rearranged forms of (4)–(7) are

(i) Implicit

$$\Big(\frac{1}{\delta}M + S + N(\mathbf{u}^{k+1})\Big)\mathbf{u}^{k+1} = \underline{\mathbf{f}}^k + \frac{1}{\delta}M\mathbf{u}^k, \tag{8}$$

(ii) Semi-implicit

$$\Big(\frac{1}{\delta}M + S + N(\mathbf{u}^{k})\Big)\mathbf{u}^{k+1} = \underline{\mathbf{f}}^k + \frac{1}{\delta}M\mathbf{u}^k, \tag{9}$$

(iii) Linearly implicit

$$\Big(\frac{1}{\delta}M + S\Big)\mathbf{u}^{k+1} = \underline{\mathbf{f}}^k + \Big(\frac{1}{\delta}M - N(\mathbf{u}^{k})\Big)\mathbf{u}^k, \tag{10}$$

(iv) Explicit

$$\Big(\frac{1}{\delta}M\Big)\mathbf{u}^{k+1} = \underline{\mathbf{f}}^k + \Big(\frac{1}{\delta}M - S - N(\mathbf{u}^{k})\Big)\mathbf{u}^k. \tag{11}$$

Motivations and earlier results

- The general idea for the conditions of δ and h is taken from Faragó I., Horváth R., and Karátson J. (2010, 2012) (θ-schemes for $0 < \theta \leq 1$ were studied).
- We introduce different types of discretizations by treating the diffusion and reaction terms in various ways, like (9)-(11), which have not yet been studied.
  We can determine the bounds of δ and h in all cases to ensure non-negativity.
- The implicit discretization in (8) is a special case of the approaches discussed in the works of Faragó I., Horváth R., and Karátson J. (2010, 2012) for $\theta = 1$.
  However, our bounds calculations for the semi-implicit case yielded the same result as the implicit case.
  Therefore, we determine the bounds in a combined form.

Entries of the matrix $A(\underline{\mathbf{v}})$

Let $A(\underline{\mathbf{v}}) = S + N(\underline{\mathbf{v}})$, then $A(\underline{\mathbf{v}})\underline{\mathbf{v}} = S\underline{\mathbf{v}} + N(\underline{\mathbf{v}})\underline{\mathbf{v}}$ for a vector $\underline{\mathbf{v}}$. Then, the entries of the matrices are

$$a_{ij}(\underline{\mathbf{v}}) = s_{ij} + n_{ij}(\underline{\mathbf{v}}) = \int_\Omega \nabla\phi_i \cdot \nabla\phi_j + \int_\Omega r(v_h)\phi_i\phi_j \, dx.$$

(i)-(ii): Implicit and semi-implicit Euler discretizations

Using the stiffness matrices $A(\mathbf{v})$ and the mass matrix M, the schemes (8)–(9) can be expressed in combined form as follows:

$$M\frac{\mathbf{u}^{k+1}-\mathbf{u}^{k}}{\delta}+A(\mathbf{u}^{k+w})\mathbf{u}^{k+1}=\underline{\mathbf{f}}^{k}, \tag{12}$$

where $w=0$ or 1, and $\mathbf{u}^{k+w}=(u_0^w,\ldots,u_n^{k+w})$ is the coordinate vector of FE solution.

Note that $w=0$ and $w=1$ in (12) correspond to semi-implicit and implicit time discretization, respectively.

The scheme (12) can be rephrased as

$$(A(\mathbf{u}^{k+w})+\frac{1}{\delta}M)\mathbf{u}^{k+1}=:B(\mathbf{u}^{k+w})\mathbf{u}^{k+1}=\underline{\mathbf{f}}^{k}+\frac{1}{\delta}M\mathbf{u}^{k}. \tag{13}$$

(i)-(ii) Implicit and semi-implicit time stepping

We wish to ensure the preservation of non-negativity in the numerical solution.

Theorem

Assume the conditions $f \geq 0$, $u_0 \geq 0$, and that for all $k \in N$, $B(\mathbf{u}^{k+w})$ is a Stieltjes matrix. Then the implicit and semi-implicit time stepping (13) *preserves non-negativity.*

Proof: The proof is established using the principle of induction.

Computable condition for implicit and semi–implicit case

Theorem

Let the assumptions (2), (3) *hold and assume that* $C_m > 0$ *such that*

$$\int_{\Omega_{ij}} \phi_i \phi_j \, dx \leq C_m \mathrm{meas}\,(\Omega_{ij}), \quad (\forall\, i,j = 1,\ldots,n).$$

For a given time-step $\delta > 0$, *let the mesh size* h *in the space satisfy*

$$0 < h \leq h_0 = \left(\frac{\sigma_0 \mu_0 \delta}{(\sigma_1 \delta + 1) C_m} \right)^{\frac{1}{2}}. \tag{14}$$

Then the implicit and semi-implicit time stepping (13) *satisfies*

$$b_{ij}(\mathbf{u}^{k+w}) \leq 0, \quad i = 1,...,n, \ j = 1,...,n \quad (i \neq j)$$

and, consequently, preserves non-negativity.

(iii) Linearly implicit discretization

We wish to ensure the preservation of non-negativity in the numerical solution.

Theorem

Assume the conditions $f \geq 0$, $u_0 \geq 0$, $\delta \leq \frac{1}{\sigma_1}$ *and that* $\frac{1}{\delta} M + S$ *is a Stieltjes matrix. Then, the linearly implicit time stepping* (10) *preserves non-negativity.*

Proof: The proof is established using the principle of induction.

Computable condition of the linearly implicit discretization.

Theorem

Let the assumption (3) *hold and assume that* $C_m > 0$ *such that*

$$\int_{\Omega_{ij}} \phi_i \phi_j \, dx \leq C_m \mathrm{meas}\,(\Omega_{ij}) \quad (\forall\, i,j = 1,\ldots,n).$$

For a given time-step $\delta > 0$*, let the mesh size* h *in the space discretization of* (10) *satisfy*

$$0 < h \leq h_0 = \left(\frac{\sigma_0 \mu_0 \delta}{C_m} \right)^{\frac{1}{2}}. \tag{15}$$

Then the linearly implicit time stepping (10) *satisfies*

$$b_{ij} \leq 0, \quad i = 1,\ldots,n,\ j = 1,\ldots,n \quad (i \neq j),$$

where b_{ij} *are entries of the matrix* $B = \frac{1}{\delta} M + S$*, and consequently, non-negativity* $\mathbf{u}^k \geq 0$ *holds for all* $k = 0,\ldots,n$.

(iv) Explicit scheme

The explicit scheme (11) must be modified to ensure non-negativity, since $\frac{1}{\delta}M$ is not a Stieltjes matrix, which implies that $\left(\frac{1}{\delta}M\right)^{-1}$ is not necessarily element-wise nonnegative.

- We utilize the mass lumping techniques described in Thomée V. (2007) for the linear case.
- We adopt the methods outlined in Frittelli M., Madzvamuse A., Sgura I., and Venkataraman C. (2017) for the non-linear reaction case.

Definition

A lumped mass matrix is a diagonal matrix whose elements are obtained by adding the corresponding rows of the mass matrix.

Lumped matrix scheme

The explicit scheme presented in (11) can be modified as follows:

$$\left(\frac{1}{\delta}\tilde{M}\right)\mathbf{u}^{k+1} = \underline{\mathbf{f}} + \left(\frac{1}{\delta}\tilde{M} - S - \tilde{N}(\mathbf{u}^k)\right)\mathbf{u}^k, \tag{16}$$

where $\tilde{M}$ and $\tilde{N}$ are the lumped matrix for M and N respectively. Using the condition $\sum_{j=1}^{n} \phi_j = 1$, and for all $i, j = 1, \ldots, n$

$$\tilde{M}_{ij} = \begin{cases} \sum_{j=1}^{n} m_{ij} = \sum_{j=1}^{n} \int\limits_{\Omega} \phi_i \phi_j = \int\limits_{\Omega} \phi_i \sum_{j=1}^{n} \phi_j = \int\limits_{\Omega} \phi_i, & i = j \\ 0, & i \neq j. \end{cases}, \tag{17}$$

Similarly,

$$\tilde{N}(\mathbf{u}^k)_{ij} = \begin{cases} \sum_{j=1}^{n} \int\limits_{\Omega} r(u_h^k)\phi_i \phi_j = \int\limits_{\Omega} r(u_h^k)\phi_i \sum_{j=1}^{n} \phi_j = \int\limits_{\Omega} r(u_h^k)\phi_i, & i = j, \\ 0, & i \neq j. \end{cases} \tag{18}$$

Modified explicit scheme: Non-negativity

Definition

Let $\phi_i, \phi_2, \ldots, \phi_n$ be a basis. Then,

$$R(h) := \max_{i=1,\ldots,n} \left(\frac{\int_\Omega (\nabla \phi_i)^2}{\int_\Omega \phi_i^2} \right).$$

Theorem

Assume the conditions $f \geq 0$, $u_0 \geq 0$,

$$\delta \leq \frac{1}{R(h) + \sigma_1} \tag{19}$$

hold, and (3) *satisfied for the FE mesh. Then, the modified explicit time stepping in* (16) *preserves non-negativity.*

Remarks: Implicit and semi-implicit time stepping

Now we can discuss the space mesh and time stepping restrictions from equations (14) to ensure non-negativity.

The implicit and semi-implicit time stepping (13)

- If δ is fixed, then the upper bound on h is

$$h \leq \left(\frac{\sigma_0 \mu_0 \delta}{(\sigma_1 \delta + 1) C_m} \right)^{\frac{1}{2}} =: h_0^{(\delta)} \tag{20}$$

  Note that here

$$h \leq O(\delta^{\frac{1}{2}}) \quad \text{as} \quad \delta \to 0.$$

- If h is fixed, then it is necessary to determine a condition for δ. Then, using equation (20):

$$C_m h^2 \leq \delta(\sigma_0 \mu_0 - h^2 \sigma_1 C_m). \tag{21}$$

Lower bounds of δ

- If $\sigma_0\mu_0 - h^2\sigma_1 C_m > 0$ in (21) , then we can determine the lower bound for δ from (21) as follows:

$$\delta \geq \frac{h^2 C_m}{\sigma_0\mu_0 - h^2\sigma_1 C_m} = \frac{h^2 C_m}{C_2 - h^2 C_3} =: \delta_0^{(h)}, \tag{22}$$

  where $C_2 = \sigma_0\mu_0$ and $C_3 = \sigma_1 C_m$.
- If $\sigma_0\mu_0 - h^2\sigma_1 C_m \leq 0$ then (21) is not possible since $C_m h^2$ is a positive quantity.
  Then, from the first argument, we can determine a bound for the admissible values of h, which is independent of δ, as follows:

$$h < h_1 = \sqrt{\frac{\sigma_0\mu_0}{\sigma_1 C_m}}. \tag{23}$$

Summary when h is fixed

In summary, when h is fixed we have two possibilities:

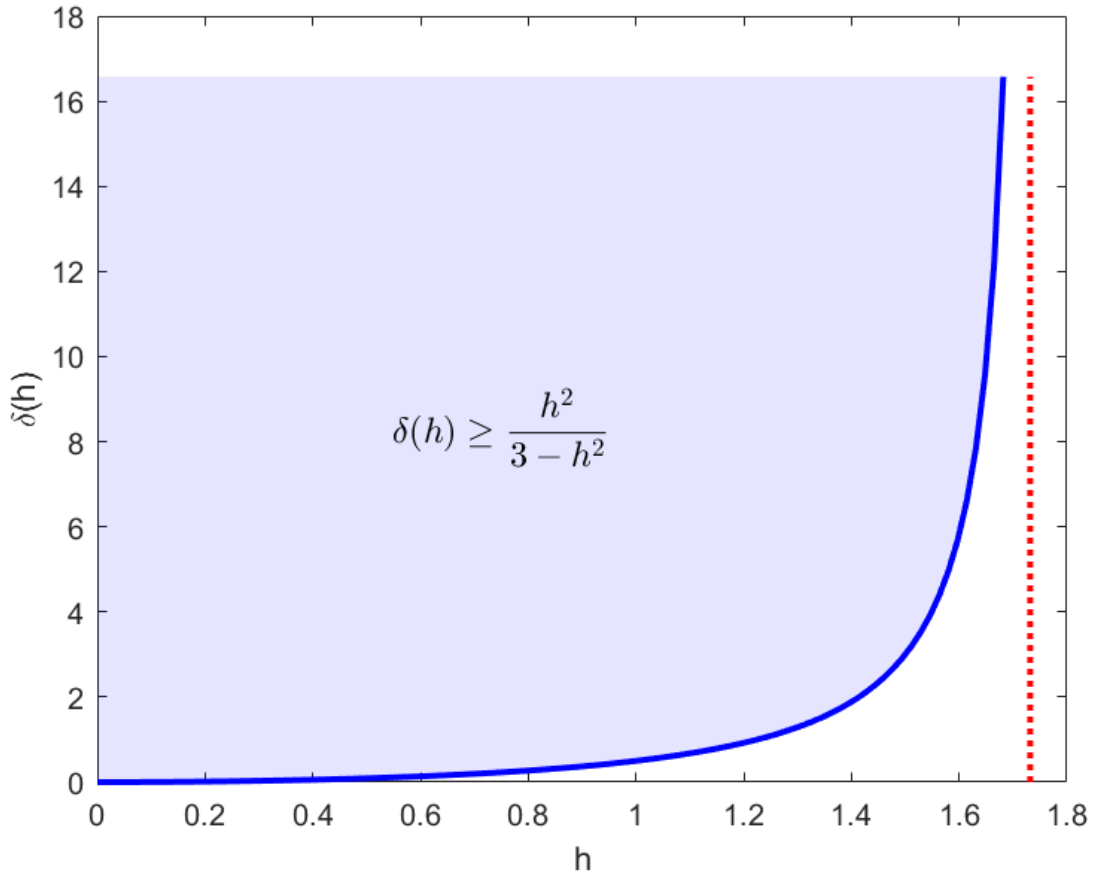
- If $h < h_1$, then $\delta \geq \delta_0^{(h)}$ implies non-negativity and

$$\delta \geq O(h^2) \quad \text{as} \quad h \to 0.$$

- If $h \geq h_1$, then there is no δ which could ensure non-negativity.

Graphical illustration for lower bound of δ in (22)

Using the constant for the bilinear elements $C_m = \frac{1}{18}$, and other constants which help to simplify easily such that $C_2 = \frac{3}{18}$ and $C_3 = \frac{1}{18}$, the lower bound of δ in terms of h is illustrated in the following figure.



Remarks: Linearly implicit time stepping

We can discuss the space mesh and time-stepping restrictions from equations (15).

- If $\delta \leq \frac{1}{\sigma_1}$, which has already been assumed, is fixed, then the upper bound on h is

$$h \leq \left(\frac{\sigma_0 \mu_0 \delta}{C_m} \right)^{\frac{1}{2}} =: h_0^{(\delta)}.$$

  We observe in this context $h \leq O(\delta^{\frac{1}{2}}) \quad \text{as} \quad \delta \to 0.$

- If h is fixed, then it is necessary to determine a condition for δ. Then, using equation (15):

$$\delta \geq \frac{C_m h^2}{\sigma_0 \mu_0} = C h^2 =: \delta_0^{(h)},$$

  where $C = \frac{C_m}{\sigma_0 \mu_0}$. That is, $\delta \geq O(h^2) \quad \text{as} \quad h \to 0.$
  In addition, δ is also bounded from above by $O(1)$, since $\delta \leq \frac{1}{\sigma_1}$, that is,

$$O(h^2) \leq \delta \leq O(1).$$

Remarks: Explicit time stepping

We can discuss the space mesh and time-stepping restrictions from equations (19). For the explicit case, based on Faragó I., Horváth R., and Karátson J. (2010), page 11, Remark 5.5, we obtain $R(h) = O(h^{-2})$.

Then, assuming a constant $K \in \mathbf{R}^+$ for the general form of (19), we obtain

$$\delta \leq \frac{1}{R(h) + \sigma_1} \leq \frac{1}{\frac{K}{h^2} + \sigma_1} = \frac{h^2}{K + \sigma_1 h^2} \leq \frac{h^2}{K} = O(h^2).$$

Thus,

$$\delta \leq O(h^2).$$

Example: Express the upper bound of δ in terms of h

Consider a uniform mesh size h for a two-dimensional domain.

Triangular (Courant) element: $\int_\Omega \phi_i^2 = \frac{h^2}{6}$ and $\int_\Omega |\nabla \phi_i|^2 = 4$,

bilinear element (H=h): $\int_\Omega \phi_i^2 = \frac{4h^2}{9}$ and $\int_\Omega |\nabla \phi_i|^2 = \frac{8}{3}$.

Therefore, for Courant element, $R(h) = 24h^{-2}$ and for bilinear element, $R(h) = 6h^{-2}$.

For the bilinear element, setting $\sigma_1 = 1$ in (19),

$$\delta \leq \frac{1}{6h^{-2}+1} = \frac{h^2}{6+h^2} \leq \frac{h^2}{6}. \tag{24}$$

Graphical illustration for example in (24)

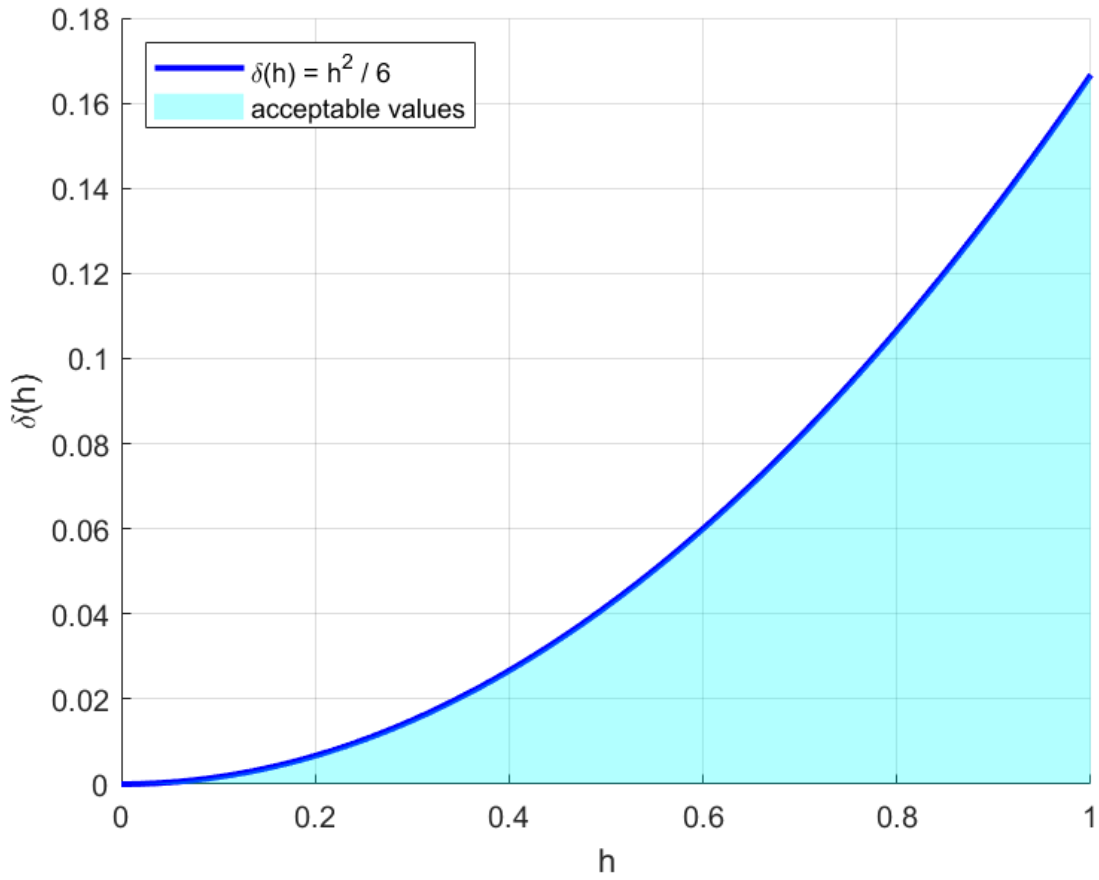


Figure: The upper bound of δ in terms of h for the explicit time stepping.

Numerical experiments: necessity of the bounds of h and δ

We illustrate the above theoretical results with an experiment for the bilinear FE solution of a 2D problem (Michaelis-Menten nonlinearity):

$$\begin{cases} \partial_t u - \mu_0 \Delta u + \frac{u}{1+\epsilon u} = f \quad \text{in } \Omega := [0,1]^2, \quad t \in (0,1] \\ u = 0 \quad \text{on } \partial\Omega. \end{cases} \tag{25}$$

, where the initial condition $(t = 0)$ $u_0 = 0$.

Numerical experiments for semi-implicit discretization

- In the experiment $\mu_0 = 10^{-5}$ and $\epsilon = 10^{-3}$ are constants given by Keller (1969).

- $f(x, y) := (2x - 1)^6 \geq 0$ describes a source function mostly concentrated near two sides of the square domain.

The table below illustrates the numerical solutions for five different meshes h and fixed $\delta = 0.1$ (non-negativity can fail for too coarse mesh).

Non-negative minima hold:

- theoretical results from (14): $h \leq h_0 = 0.00165$,
- experimental results: $h \leq 0.002$.

This indicates that the estimation magnitude is reasonable.

Numerical experiments: δ is fixed and h is decreased

h	0.1	0.01	0.0025	0.002	0.001
min u_h for $t_1 = 0.1$	-0.004	$-1.1e-12$	$-9.6e-16$	0	0
min u_h for $t_2 = 0.2$	-0.01	$-1.1e-12$	$-5.4e-16$	0	0
min u_h for $t_3 = 0.3$	-0.01	$-2.6e-13$	$-1.5e-16$	0	0
min u_h for $t_4 = 0.4$	-0.02	$-2.3e-12$	0	0	0
min u_h for $t_5 = 0.5$	-0.02	$-6.1-12$	0	0	0
min u_h for $t_6 = 0.6$	-0.02	$-1.1e-11$	0	0	0
min u_h for $t_7 = 0.7$	-0.02	$-1.6e-11$	0	0	0
min u_h for $t_8 = 0.8$	-0.02	$-2.1e-11$	0	0	0
min u_h for $t_9 = 0.9$	-0.03	$-2.7e-11$	0	0	0
min u_h for $t_{10} = 1$	-0.03	$-3.2e-12$	0	0	0

Table: Space size h and time levels $t_1, \ldots, t_{10}$ with the corresponding min u_h.

Numerical experiments: h is fixed and δ is increased

The table below illustrates the numerical solutions for five different time stepping δ and fixed $h = 0.001$ (non-negativity can fail for too fine time-stepping).

Non-negative minima hold:

- theoretical results from (22): $\delta \geq \delta_0 = 0.035$,
- experimental results: $\delta \geq 0.02$.

This indicates that the estimation magnitude is reasonable.

Numerical experiments: h is fixed, δ is increased

δ	0.001	0.01	0.02	0.035	0.1
min u_h for $t_1 = 0.1$	-1.1e-16	0	0	0	0
min u_h for $t_2 = 0.2$	-8.7e-15	0	0	0	0
min u_h for $t_3 = 0.3$	-1.3e-15	0	0	0	0
min u_h for $t_4 = 0.4$	-5.9e-14	0	0	0	0
min u_h for $t_5 = 0.5$	-1.3e-14	0	0	0	0
min u_h for $t_6 = 0.6$	-2.9e-15	-7.9e-16	0	0	0
min u_h for $t_7 = 0.7$	-4.7e-13	-6.5e-14	0	0	0
min u_h for $t_8 = 0.8$	-6.8e-14	-9.4e-15	0	0	0
min u_h for $t_9 = 0.9$	-3.6e-12	-5.3e-14	0	0	0
min u_h for $t_{10} = 1$	-5.1e-11	-6.2e-13	0	0	0

Table: Time step δ and time levels $t_1, \ldots, t_{10}$ with the corresponding min u_h.

Conclusion

- We have determined threshold mesh sizes for h and the upper and lower bounds of δ using computable conditions based on the geometric characteristics of widely studied FE shapes for spatial discretization (triangles and rectangles) and four cases of time-discretization approaches.

  This ensures the discrete non-negativity preservation (DNP) for parabolic standard diffusion with nonlinear reaction PDEs.

Thank you for your attention!