# Steady State and Sign Preserving Semi-Implicit Runge-Kutta Methods for Differential Equations with Stiff Damping Term

*Alexander Kurganov*

**Tulane University**

Mathematics Department

`www.math.tulane.edu/~kurganov`

Alina Chertock, North Carolina State University, USA

Shumo Cui, Tulane University, USA

Tong Wu, Tulane University, USA

Supported by NSF and ONR

$$\boldsymbol{u}' = \boldsymbol{f}(\boldsymbol{u}, t) + G(\boldsymbol{u}, t)\boldsymbol{u}$$

$\boldsymbol{u}(t) \in \mathbb{R}^N$: unknown vector function

$\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^N$: given vector field

$G : \mathbb{R}^{N\times N} \to \mathbb{R}^{N\times N}$: diagonal non-positive definite matrix representing a (stiff) damping term

Steady States: $\boldsymbol{u}(t) \equiv \widehat{\boldsymbol{u}} \quad \text{s.t.} \quad \boldsymbol{f}(\widehat{\boldsymbol{u}}, t) \equiv -G(\widehat{\boldsymbol{u}}, t)\widehat{\boldsymbol{u}}$

Sign Preservation provided $\{\boldsymbol{u}(0) \geq 0, \quad \boldsymbol{f} \geq 0\}$ or $\{\boldsymbol{u}(0) \leq 0, \quad \boldsymbol{f} \leq 0\}$

# Explicit vs. Implicit vs. Semi-Implicit Methods

For simplicity, consider a scalar ODE

$$u' = f(u,t) + g(u,t)u, \quad g(u,t) \le 0$$

Example: **First-Order Explicit (Forward Euler) Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^n,t^n) + g(u^n,t^n)u^n\Big]$$

Example: **First-Order Implicit (Backward Euler) Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^{n+1},t^{n+1}) + g(u^{n+1},t^{n+1})u^{n+1}\Big]$$

Example: **First-Order Semi-Implicit Method**

$$u^{n+1} = u^n + \Delta t\Big[f(u^n,t^n) + g(u^n,t^n)u^{n+1}\Big]$$

# Explicit $m$-stage SSP (TVD) RK Methods

[Shu; 1988] [Shu, Osher; 1988] [Gottlieb, Shu, Tadmor; 2001]

For simplicity, consider a scalar ODE

$$u' = f(u,t) + g(u,t)u, \quad g(u,t) \leq 0$$

$f(u,t)$: nonstiff term, $g(u,t)u$: stiff damping term

A general explicit $m$-stage RK method is

$$\begin{aligned} u^{(0)} &= u^n \\ u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k} \left[ u^{(k)} + \beta_{i,k} \Delta t (f^{(k)} + g^{(k)} u^{(k)}) \right], \quad i = 1, \ldots, m \\ u^{n+1} &= u^{(m)} \end{aligned}$$

where $f^{(k)} := f(u^{(k)}, t^{(k)})$, $g^{(k)} := g(u^{(k)}, t^{(k)})$, $t^{(k)} := t^n + D_k \Delta t$, $t^{n+1} := t^n + \Delta t$ and $D_k$ are given by

$$D_0 = 0, \quad D_i = \sum_{k=0}^{i-1} \alpha_{i,k} (D_k + \beta_{i,k})$$

The RK method is fully determined by its coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$

Consistency requirements: $\sum_{k=0}^{i-1} \alpha_{i,k} = 1, \quad i = 1, \dots, m, \quad D_m = 1$

The RK method is a linear combination of the first-order FE steps:

$$u^{(i)} = \sum_{k=0}^{i-1} \alpha_{i,k} u_{i,k}^{\mathrm{FE}}$$

where

$$u_{i,k}^{\mathrm{FE}} := u^{(k)} + \beta_{i,k} \Delta t (f^{(k)} + g^{(k)} u^{(k)})$$

According to [Gottlieb, Shu, Tadmor; 2001], the RK method is SSP provided

$$\alpha_{i,k} \geq 0 \quad \text{for all} \quad i, k$$

and an appropriate time step restriction is imposed.

Negative time increments are avoided if $\beta_{i,k} \geq 0 \;\; \text{for all} \;\; i, k$

# New Semi-Implicit Methods

We first replace the FE evolution steps by the semi-implicit (SI) ones:

$$u_{i,k}^{\text{SI}} := u^{(k)} + \beta_{i,k}\Delta t(f^{(k)} + g^{(k)} u_{i,k}^{\text{SI}}) \quad \Longleftrightarrow \quad u_{i,k}^{\text{SI}} = \frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}}$$

This leads to the following SI scheme:

$$\begin{aligned}
u^{(0)} &= u^n \\
u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k} \left( \frac{u^{(k)} + \beta_{i,k}\Delta t f^{(k)}}{1 - \beta_{i,k}\Delta t g^{(k)}} \right), \quad i = 1, \dots, m \\
u^{n+1} &= u^{(m)}
\end{aligned}$$

Unfortunately, this scheme is at most first-order accurate

We, therefore, propose an order correction step:

$$u^{n+1} = \frac{u^{(m)} - C_m(\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m(\Delta t g^{(m)})^2}$$

where

$$C_0 = 0, \quad C_i = \sum_{k=0}^{i-1} \alpha_{i,k}(C_k + \beta_{i,k}^2), \quad i = 1, \dots, m$$

New class of second-order semi-implicit Runge-Kutta (SI-RK) methods:

$$
\begin{aligned}
u^{(0)} &= u^n \\
u^{(i)} &= \sum_{k=0}^{i-1} \alpha_{i,k} \left( \frac{u^{(k)} + \beta_{i,k} \Delta t f^{(k)}}{1 - \beta_{i,k} \Delta t g^{(k)}} \right), \quad i = 1, \dots, m \\
u^{n+1} &= \frac{u^{(m)} - C_m (\Delta t)^2 f^{(m)} g^{(m)}}{1 + C_m (\Delta t g^{(m)})^2}
\end{aligned}
$$

The set of coefficients $\{\alpha_{i,k}, \beta_{i,k}\}$ is taken directly from the explicit SSP-RK method of an appropriate order.

Remark. Note that in the degenerate case of $g \equiv 0$, the SI-RK methods are identical to the corresponding explicit RK methods

Theorem (Second-Order Accuracy) If the SSP-RK method is at least second-order accurate, then the corresponding SI-RK method with the same set of coefficients $\alpha_{i,k}, \beta_{i,k} \geq 0$ is second-order.

Theorem ($A(\alpha)$-Stability and Stiff Decay) Let us assume that the SI-RK methods are applied to the test equation $u' = \lambda u$, where $\lambda \in \mathbb{C}$ is a constant with $\operatorname{Re}\lambda < 0$. Then, the resulting methods, which can be written as

$$u^{n+1} = R(z)u^n, \quad z = \lambda \Delta t$$

satisfy the following two requirements:

$$|R(z)| \leq 1, \ \forall z \in \mathbb{C} \text{ s.t. } \operatorname{Re} z \leq -|\operatorname{Im} z| \quad \left(A(\alpha)\text{-stability with } \alpha = \frac{\pi}{4}\right)$$

and

$$R(z) \to 0 \text{ as } \operatorname{Re} z \to -\infty$$

provided $\alpha_{i,k} \geq 0$ and $\beta_{i,k} \geq 0$ for all $i, k$.

Theorem (Steady State Preserving Property) Let $\beta_{i,k} \geq 0$ $\forall i, k$. Then, if the computed solution is at a steady state at time $t^n$, i.e., $u^n = \widehat{u}$ such that

$$f(\widehat{u}, t) \equiv -g(\widehat{u}, t)\widehat{u}$$

it will remain at the same steady state, namely,

$$u^{n+1} = \widehat{u}$$

Theorem (Sign Preserving Property) Let the initial condition $u^0$ and function $f$ satisfy

$$\{u^0 \geq 0, \quad f \geq 0\} \quad \text{or} \quad \{u^0 \leq 0, \quad f \leq 0\}$$

Then,

$$\text{sgn}(u^n) \equiv \text{sgn}(u^0)$$

for all $n$ provided $\alpha_{i,k} \geq 0$ and $\beta_{i,k} \geq 0$ for all $i, k$

# Absolute Stability of Two SSP-Based SI-RK Methods

The first SI-RK2 method is based on the 2-order SSP-RK solver:

$$
\begin{aligned}
u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n} \\
u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}} \\
u^{n+1} &= \frac{u^{(2)} - (\Delta t)^2 f^{(2)} g^{(2)}}{1 + (\Delta t g^{(2)})^2}
\end{aligned}
$$

The second SI-RK3 method is based on the 3-order SSP-RK solver:

$$
\begin{aligned}
u^{(1)} &= \frac{u^n + \Delta t f^n}{1 - \Delta t g^n} \\
u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} \cdot \frac{u^{(1)} + \Delta t f^{(1)}}{1 - \Delta t g^{(1)}} \\
u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3} \cdot \frac{u^{(2)} + \Delta t f^{(2)}}{1 - \Delta t g^{(2)}} \\
u^{n+1} &= \frac{u^{(3)} - (\Delta t)^2 f^{(3)} g^{(3)}}{1 + (\Delta t g^{(3)})^2}
\end{aligned}
$$

To analyze the absolute stability, we consider the following test problem:

$$u' = \lambda_1 u + \lambda_2 u, \quad \lambda_1 \in \mathbb{C}, \ \mathrm{Re}(\lambda_1) \leq 0, \quad \lambda_2 \in \mathbb{R}, \ \lambda_2 \leq 0$$

$\lambda_1 u$: nonstiff part, $\lambda_2 u$: stiff part

We denote $z_1 := \lambda_1 \Delta t$ and $z_2 := \lambda_2 \Delta t$.

We denote the stability regions of the second- and third-order SSP-RK methods by $\mathcal{D}_{\mathrm{SSP2}}$ and $\mathcal{D}_{\mathrm{SSP3}}$, respectively.

We denote the corresponding time step restrictions by $\Delta t \leq \Delta t_{\mathrm{SSP2}}$ and $\Delta t \leq \Delta t_{\mathrm{SSP3}}$

Theorem (Absolute Stability of the SI-RK2 Method) The region of absolute stability of the SI-RK2 method contains $\mathcal{D}_{\mathsf{SSP2}}$, i.e., for any $z_2 \le 0$, the solution of

$$\begin{aligned} u^{(1)} &= \frac{1+z_1}{1-z_2} u^n \\ u^{(2)} &= \frac{1}{2} u^n + \frac{1}{2} \cdot \frac{1+z_1}{1-z_2} u^{(1)} \\ u^{n+1} &= \frac{1-z_1 z_2}{1+z_2^2} u^{(2)} \end{aligned}$$

satisfies $|u^{n+1}| \le |u^n|$ provided $\Delta t \le \Delta t_{\mathsf{SSP2}}$

Idea of Proof: Stability function for the second-order SSP-RK method (applied to $u' = \lambda_1 u$) is:

$$R_{\text{SSP2}}(z_1) = \frac{1}{2} + \frac{1}{2}(1 + z_1)^2$$

Stability function for the SI-RK2 methods (applied to $u' = \lambda_1 u + \lambda_2 u$) is:

$$R_{\text{SI-RK2}}(z_1, z_2) = \frac{1 - z_1 z_2}{1 + z_2^2} \cdot \left[\frac{1}{2} + \frac{1}{2}\left(\frac{1 + z_1}{1 - z_2}\right)^2\right]$$

To prove the theorem, it will be enough to show that both

$$\left|\frac{1}{2} + \frac{1}{2}\left(\frac{1 + z_1}{1 - z_2}\right)^2\right| \leq 1 \tag{1}$$

and

$$\left|\frac{1 - z_1 z_2}{1 + z_2^2}\right| \leq 1 \tag{2}$$

for all $z_1, z_2$ such that $|R_{\text{SSP2}}(z_1)| \leq 1$ and $z_2 \leq 0$

Proof of (1) is straightforward.

For fixed $z_2 < 0$, (2) is equivalent to

$$\left| z_1 + \frac{1}{|z_2|} \right| \leq |z_2| + \frac{1}{|z_2|}$$

Denoting $z_1 := x + iy$, we can write this domain as

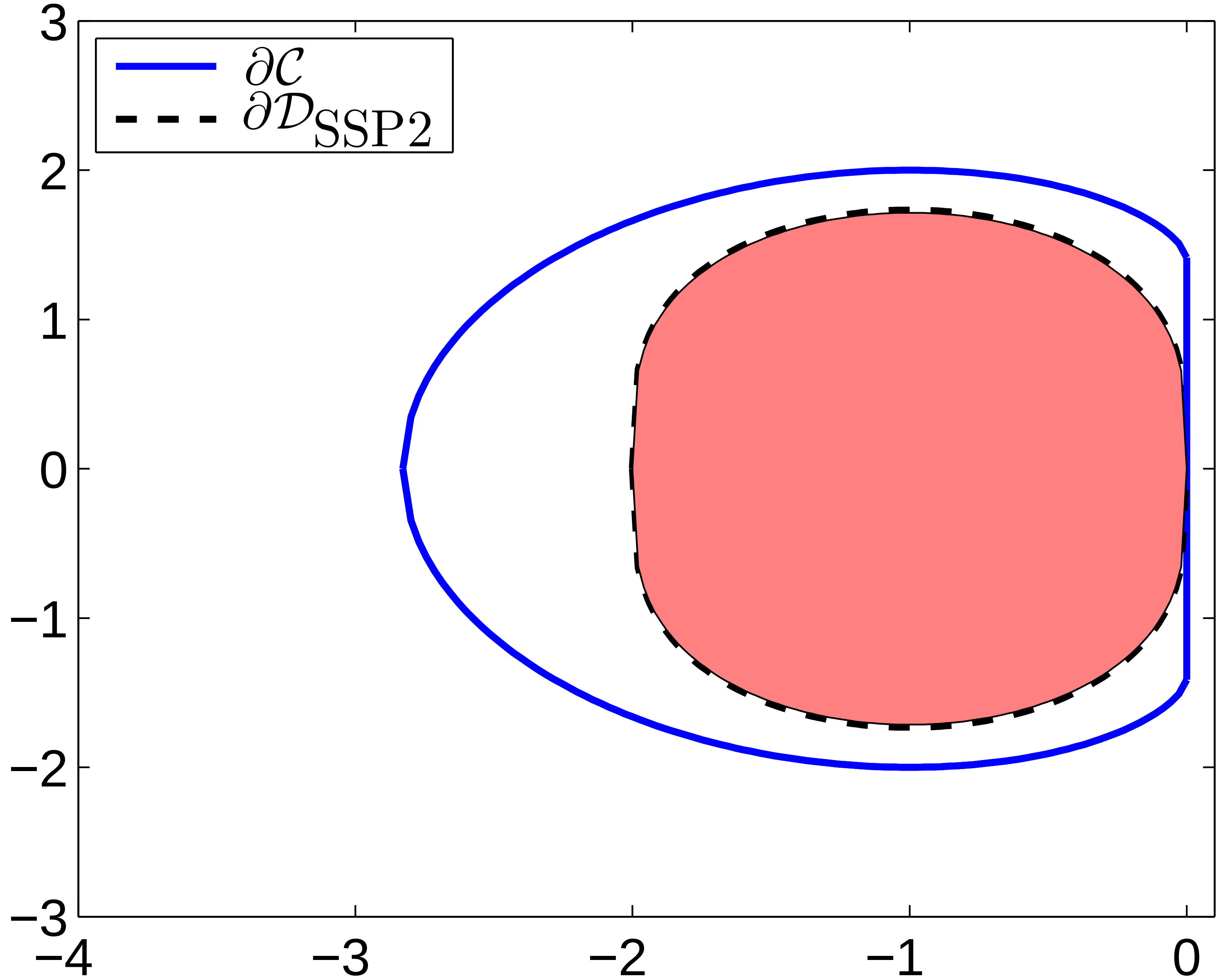
$$\mathcal{C}(z_2) := \left\{ x + iy \;\middle|\; y^2 \leq \left( z_2 + \frac{1}{z_2} \right)^2 - \left( x - \frac{1}{z_2} \right)^2 \right\}, \quad \forall z_2 < 0$$

We thus need to show that $\mathcal{D}_{\mathrm{SSP2}} \subset \mathcal{C} := \bigcap_{z_2 < 0} \mathcal{C}(z_2)$

We compute intersection of $\mathcal{C}(z_2)$'s:

$$\mathcal{C} = \left\{ x + yi \;\middle|\; y^2 \leq 2 + 3x^{2/3} - x^2,\ x \in \left[ -2\sqrt{2}, 0 \right] \right\}$$

which clearly shows that $\mathcal{D}_{\mathrm{SSP2}} \subset \mathcal{C}$



$\partial\mathcal{C}$
$\partial\mathcal{D}_{\mathrm{SSP2}}$
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0

Conjecture (Absolute Stability of the SI-RK3 Method) The region of absolute stability of the SI-RK3 method contains $\mathcal{D}_{\mathrm{SSP3}}$, i.e., for any $z_2 \le 0$, the solution of

$$
\begin{aligned}
u^{(1)} &= \frac{1+z_1}{1-z_2}\, u^n \\
u^{(2)} &= \frac{3}{4} u^n + \frac{1}{4} \cdot \frac{1+z_1}{1-z_2}\, u^{(1)} \\
u^{(3)} &= \frac{1}{3} u^n + \frac{2}{3} \cdot \frac{1+z_1}{1-z_2}\, u^{(2)} \\
u^{n+1} &= \frac{1-z_1 z_2}{1+z_2^2}\, u^{(3)}
\end{aligned}
$$

satisfies $|u^{n+1}| \le |u^n|$ provided $\Delta t \le \Delta t_{\mathrm{SSP3}}$

Idea of "Proof": Stability function for the third-order SSP-RK method (applied to $u' = \lambda_1 u$) is:

$$R_{\text{SSP3}}(z_1) = \frac{1}{3} + \frac{1}{2}(1+z_1) + \frac{1}{6}(1+z_1)^3$$

Stability function for the SI-RK3 methods (applied to $u' = \lambda_1 u + \lambda_2 u$) is:

$$R_{\text{SI-RK3}}(z_1, z_2) = \frac{1 - z_1 z_2}{1 + z_2^2} \cdot \left[\frac{1}{3} + \frac{1}{2}\left(\frac{1+z_1}{1-z_2}\right) + \frac{1}{6}\left(\frac{1+z_1}{1-z_2}\right)^3\right]$$

The statement of the conjecture would be true if one could show that

$$|R_{\text{SI-RK3}}(z_1, z_2)| \leq 1 \quad \forall z_1 \text{ such that } |R_{\text{SSP3}}(z_1)| \leq 1 \quad \text{and} \quad \forall z_2 \leq 0$$

It is quite straightforward to show that

$$|R_{\text{SI-RK3}}(z_1, z_2)| \leq 1 \quad \forall z_1 \text{ such that } |R_{\text{SSP3}}(z_1)| \leq 1 \quad \text{and} \quad \forall\, z_2 \leq -3$$

To study the case $z_2 \in (-3, 0)$, we introduce a polynomial

$$P(x, y) := |R_{\text{SSP3}}(x + iy)|^2 - 1$$

and a rational function

$$Q(x, y, z_2) := |R_{\text{SI-RK3}}(x + iy, z_2)|^2 - 1$$

For fixed $z_2$, the curves $P(x, y) = 0$ and $Q(x, y, z_2) = 0$ are boundaries of the domains $\mathcal{D}_{\text{SSP3}}$ and $\mathcal{D}_{\text{SI-RK3}}(z_2)$, respectively

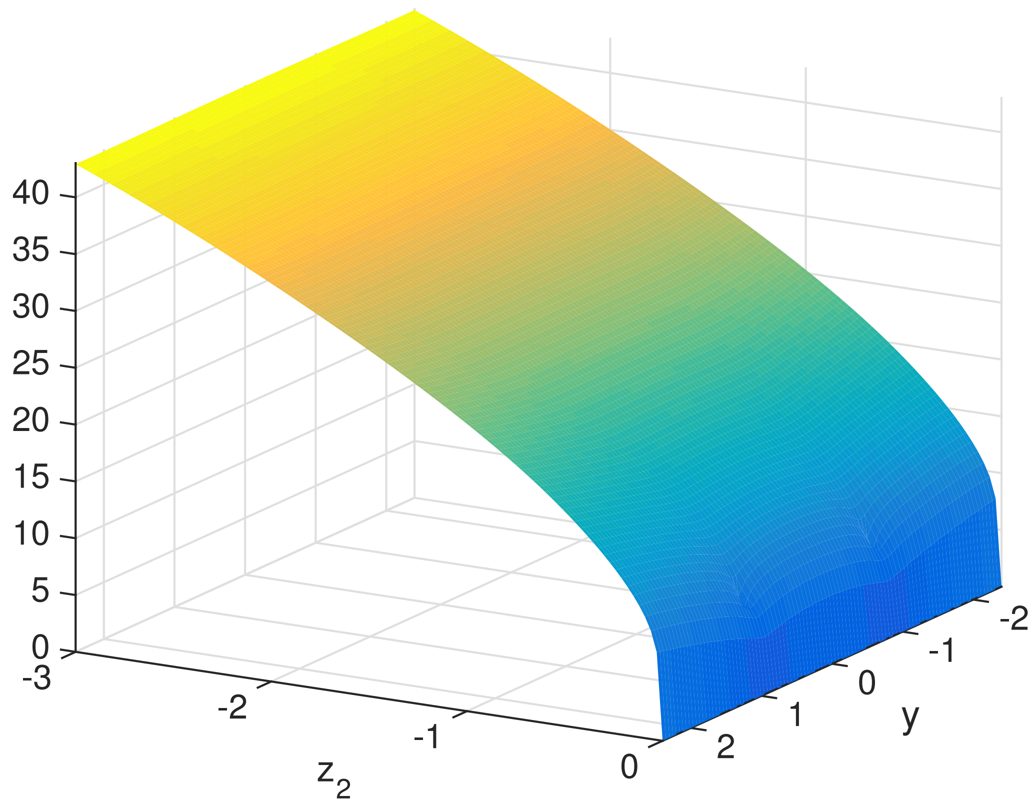
$\mathcal{D}_{\text{SI-RK3}}(z_2)$: stability domain for the SI-RK3 method for fixed $z_2$

To determine whether $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(z_2)$, we only need to verify that $\partial\mathcal{D}_{\text{SSP3}}$ is enclosed by $\partial\mathcal{D}_{\text{SI-RK3}}(z_2)$

To this end, we consider $P(x,y)$ and $Q(x,y,z_2)$ as polynomials of a single variable $x$ and compute their resultant

$$K(y,z_2) := \mathrm{res}(P,Q) = \frac{\widetilde{K}(y,z_2)}{6140942214464815497216(z_2-1)^{36}(z_2^2+1)^{12}}$$

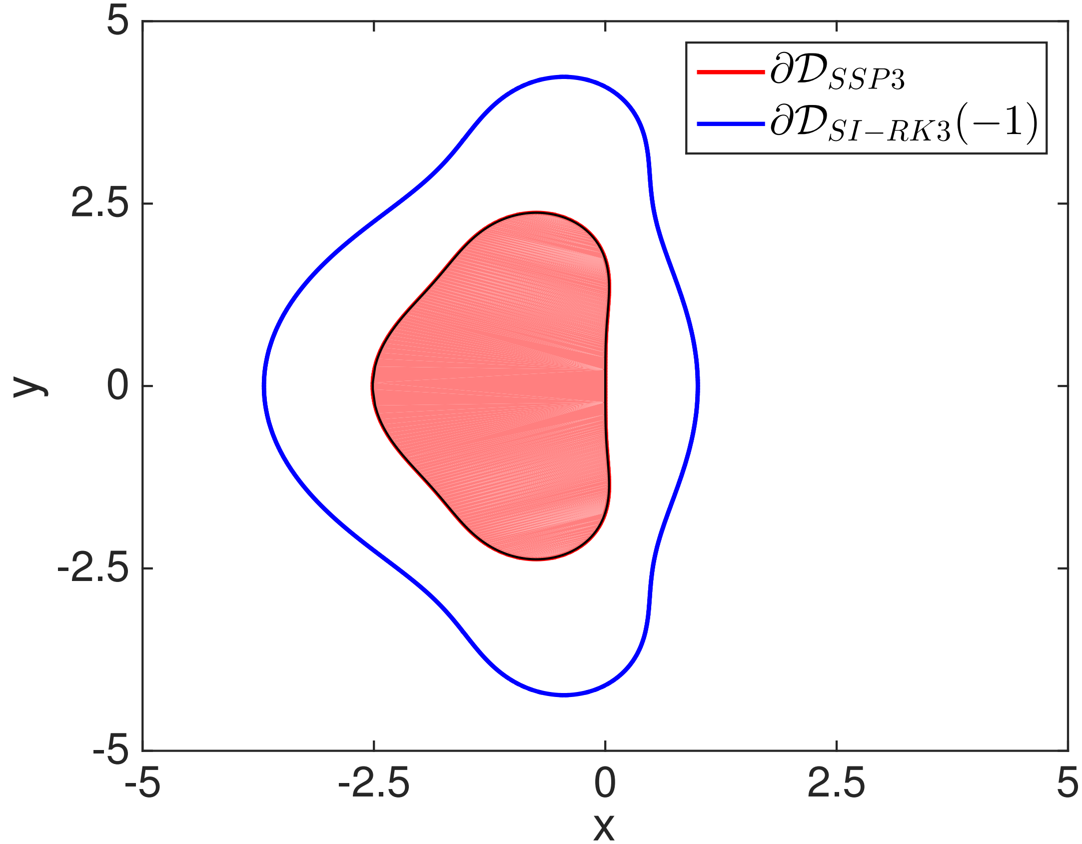
$\widetilde{K}(y,z_2)$ is explicitly given. $\log_{10}(\widetilde{K}(y,z_2)+1)$ is visualized in




which indicates that $K(y,z_2) > 0$ for all $(y,z_2) \in [-2.4, 2.4] \times (-3, 0)$

This implies that $\partial\mathcal{D}_{\text{SSP3}}$ and $\partial\mathcal{D}_{\text{SI-RK3}}(z_2)$ have no intersections when $z_2 \in (-3, 0)$.

We take $z_2 = -1$ and illustrate that $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(-1)$:




Since $K(y, z_2)$ is continuous, we conclude that $\mathcal{D}_{\text{SSP3}} \subset \mathcal{D}_{\text{SI-RK3}}(z_2)$ for all $z_2 \in (-3, 0)$

# Numerical Examples

We test the second-order SI-RK3 method and compare the results with the ones obtained using the second-order IMEX-SSP3(3,3,2) method of Pareschi and Russo.

The obtained results clearly demonstrate that the new SI-RK3 method outperforms the IMEX-SSP3(3,3,2) when a large time step and/or coarse grid are used.

**Example — Scalar ODE**

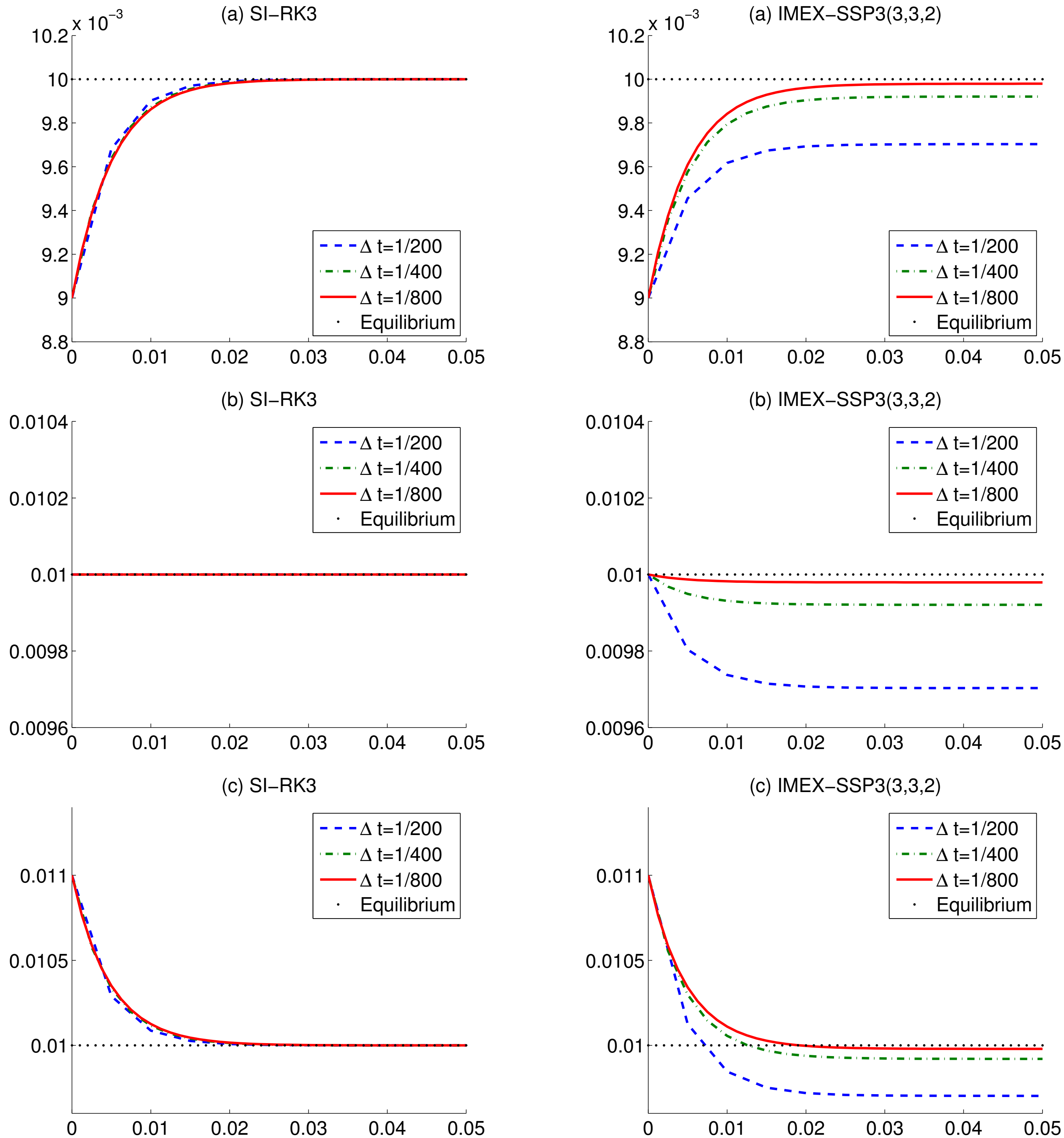
$$u' = 1 - k|u|u, \quad k > 0$$

It has one equilibrium point $u^* = 1/\sqrt{k}$

## Steady State Preserving Test

We take $k = 10000$ with the corresponding equilibrium point $u^* = 0.01$. We consider three different initial values:

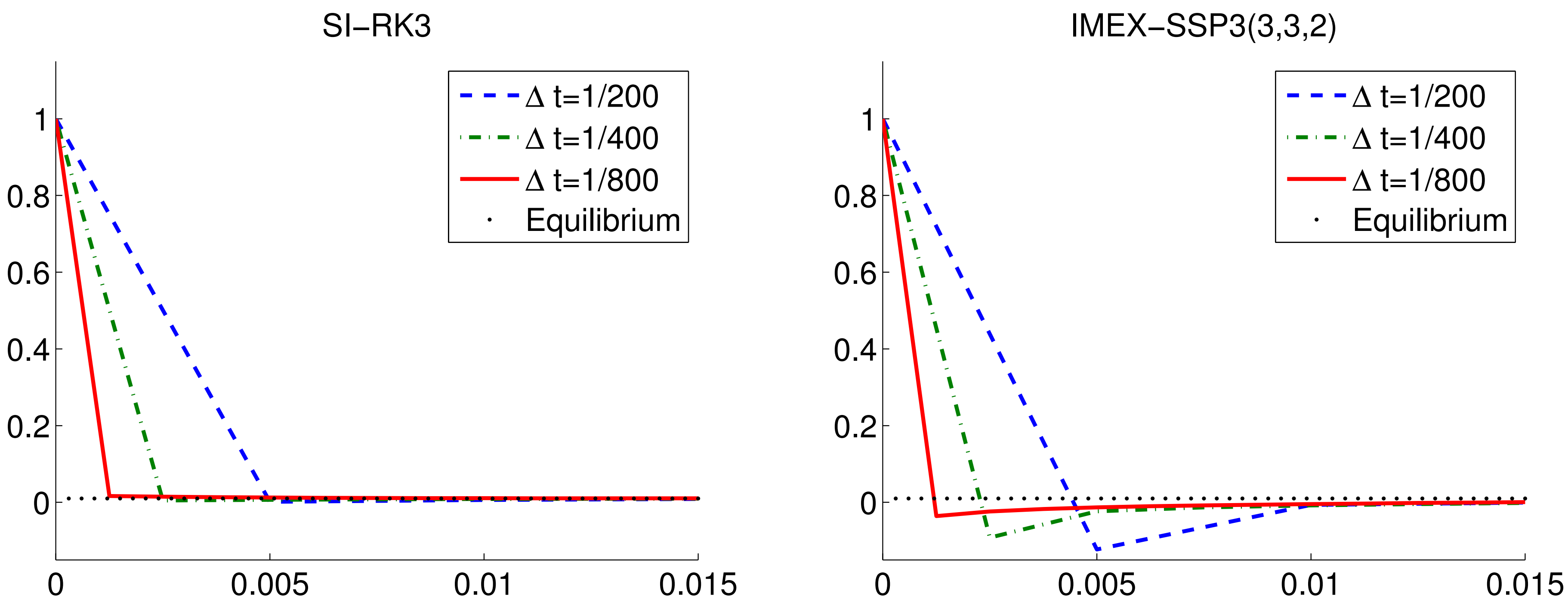
$$\text{(a) } u(0) = 0.9u^*, \quad \text{(b) } u(0) = u^*, \quad \text{(c) } u(0) = 1.1u^*$$


(a) SI–RK3
(a) IMEX–SSP3(3,3,2)
x 10−3
Δ t=1/200
Δ t=1/400
Δ t=1/800
Equilibrium
(b) SI–RK3
(b) IMEX–SSP3(3,3,2)
(c) SI–RK3
(c) IMEX–SSP3(3,3,2)

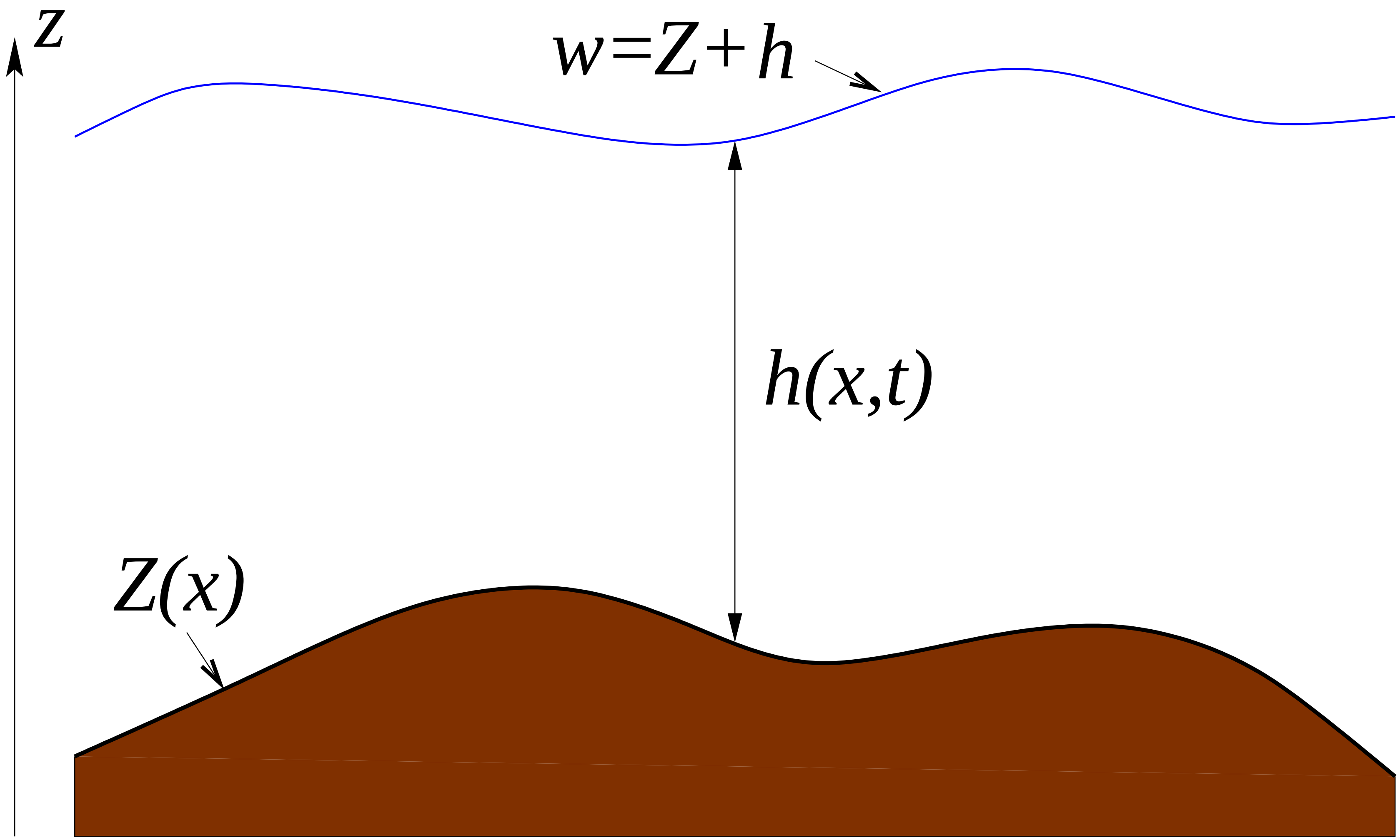
## Sign Preserving Test

We take $k = 10000$ with the corresponding equilibrium point $u^* = 0.01$. We consider large initial value:

$$u(0) = 1$$



# Shallow Water Equations



# 1-D Saint-Venant System

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x \end{cases}$$

This is a system of hyperbolic balance laws

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U}, Z)_x = \boldsymbol{S}(\boldsymbol{U}, Z), \quad \boldsymbol{U} := (h, q)$$

$h$: depth

$u$: velocity

$q := hu$: discharge

$Z$: bottom topography

$g$: gravitational constant

# Finite-Volume Methods

**1-D System:** $\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U})_x = \boldsymbol{0}$

$\overline{\boldsymbol{U}}_j(t) \approx \dfrac{1}{\Delta x} \displaystyle\int\limits_{C_j} \boldsymbol{U}(x,t)\,dx$ : cell averages over $C_j := (x_{j-\frac{1}{2}}, x_{j+\frac{1}{2}})$

This solution is approximated by a piecewise polynomial (conservative, high-order accurate, non-oscillatory) reconstruction:

$\widetilde{\boldsymbol{U}}(x) = \boldsymbol{P}_j(x)$ for $x \in C_j$

Second-order schemes employ piecewise linear reconstructions:

$\widetilde{\boldsymbol{U}}(x) = \overline{\boldsymbol{U}}_j + (\boldsymbol{U}_x)_j(x - x_j)$ for $x \in C_j$

For example,

$$\boxed{(\boldsymbol{U}_x)_j = \text{minmod}\left(\theta\frac{\overline{\boldsymbol{U}}_j - \overline{\boldsymbol{U}}_{j-1}}{\Delta x}, \frac{\overline{\boldsymbol{U}}_{j+1} - \overline{\boldsymbol{U}}_{j-1}}{2\Delta x}, \theta\frac{\overline{\boldsymbol{U}}_{j+1} - \overline{\boldsymbol{U}}_j}{\Delta x}\right)} \quad \theta \in [1, 2]$$

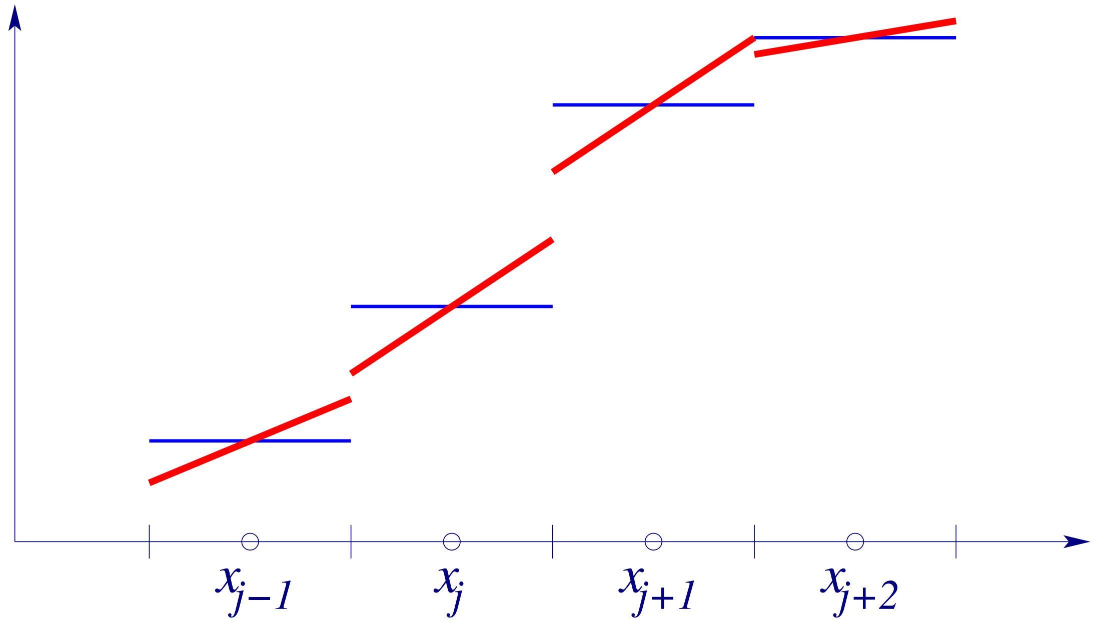
where the minmod function is defined as:

$$\boxed{\text{minmod}(z_1, z_2, ...) := \begin{cases} \min_j\{z_j\}, & \text{if } \ z_j > 0 \ \ \forall j, \\ \max_j\{z_j\}, & \text{if } \ z_j < 0 \ \ \forall j, \\ 0, & \text{otherwise.} \end{cases}}$$

The reconstructed point values at cell interfaces are:

$$\boldsymbol{U}^-_{j+\frac{1}{2}} := \boldsymbol{P}_j(x_{j+\frac{1}{2}}) = \overline{\boldsymbol{U}}_j + \frac{\Delta x}{2}(\boldsymbol{U}_x)_j$$

$$\boldsymbol{U}^+_{j+\frac{1}{2}} := \boldsymbol{P}_{j+1}(x_{j+\frac{1}{2}}) = \overline{\boldsymbol{U}}_{j+1} - \frac{\Delta x}{2}(\boldsymbol{U}_x)_{j+1}$$




The discontinuities appearing at the reconstruction step at the interface points $\{x_{j+\frac{1}{2}}\}$ propagate at finite speeds estimated by:

$$a^+_{j+\frac{1}{2}} := \max\left\{\lambda_N\Big(A(\boldsymbol{U}^-_{j+\frac{1}{2}})\Big), \lambda_N\Big(A(\boldsymbol{U}^+_{j+\frac{1}{2}})\Big), 0\right\}$$

$$a^-_{j+\frac{1}{2}} := \min\left\{\lambda_1\Big(A(\boldsymbol{U}^-_{j+\frac{1}{2}})\Big), \lambda_1\Big(A(\boldsymbol{U}^+_{j+\frac{1}{2}})\Big), 0\right\}$$

$\lambda_1 < \lambda_2 < \ldots < \lambda_N$: $N$ eigenvalues of the Jacobian $A(\boldsymbol{U}) := \dfrac{\partial \boldsymbol{F}}{\partial \boldsymbol{U}}$

# Central-Upwind Schemes

Godunov-type central schemes with a built-in upwind nature

[Kurganov, Tadmor; 2000]

[Kurganov, Petrova; 2000, 2001]

[Kurganov, Noelle, Petrova; 2001]

[Kurganov, Lin; 2007]

# 1-D Semi-Discrete Central-Upwind Scheme

$$\frac{d}{dt}\overline{\boldsymbol{U}}_j(t) = -\frac{\boldsymbol{H}_{j+\frac{1}{2}}(t) - \boldsymbol{H}_{j-\frac{1}{2}}(t)}{\Delta x}$$

The central-upwind numerical flux is:

$$\boldsymbol{H}_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}}\boldsymbol{F}(\boldsymbol{U}^-_{j+\frac{1}{2}}) - a^-_{j+\frac{1}{2}}\boldsymbol{F}(\boldsymbol{U}^+_{j+\frac{1}{2}})}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} + a^+_{j+\frac{1}{2}}a^-_{j+\frac{1}{2}}\left[\frac{\boldsymbol{U}^+_{j+\frac{1}{2}} - \boldsymbol{U}^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}} - \boxed{\boldsymbol{d}_{j+\frac{1}{2}}}\right]$$

The built-in "anti-diffusion" term is:

$$\boldsymbol{d}_{j+\frac{1}{2}} = \text{minmod}\left(\frac{\boldsymbol{U}^+_{j+\frac{1}{2}} - \boldsymbol{U}^*_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}, \frac{\boldsymbol{U}^*_{j+\frac{1}{2}} - \boldsymbol{U}^-_{j+\frac{1}{2}}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}\right)$$

The intermediate values $\boldsymbol{U}^*_{j+\frac{1}{2}}$ are:

$$\boldsymbol{U}^*_{j+\frac{1}{2}} = \frac{a^+_{j+\frac{1}{2}}\boldsymbol{U}^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}\boldsymbol{U}^-_{j+\frac{1}{2}} - \left\{\boldsymbol{F}(\boldsymbol{U}^+_{j+\frac{1}{2}}) - \boldsymbol{F}(\boldsymbol{U}^-_{j+\frac{1}{2}})\right\}}{a^+_{j+\frac{1}{2}} - a^-_{j+\frac{1}{2}}}$$

## Remarks

1. $\boldsymbol{d}_{j+\frac{1}{2}} \equiv 0$ corresponds to the central-upwind scheme from [Kurganov, Noelle, Petrova; 2001]

2. For the system of balance laws

$$\boldsymbol{U}_t + \boldsymbol{F}(\boldsymbol{U})_x = \boldsymbol{S}$$

the central-upwind scheme is:

$$\frac{d}{dt}\bar{\boldsymbol{U}}_j(t) = -\frac{\boldsymbol{H}_{j+\frac{1}{2}}(t) - \boldsymbol{H}_{j-\frac{1}{2}}(t)}{\Delta x} + \boxed{\boxed{\bar{\boldsymbol{S}}_j(t)}}$$

where

$$\bar{\boldsymbol{S}}_j(t) \approx \frac{1}{\Delta x} \int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} \boldsymbol{S}(x,t)\,dx$$

## Saint-Venant System — Numerical Challenges

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x \end{cases}$$

- Steady-state solutions:

$$q = \text{Const}, \quad \frac{u^2}{2} + g(h+Z) = \text{Const}$$

- "Lake at rest" steady-state solutions:

$$u = 0, \quad h + Z = \text{Const}$$

- Dry ($h = 0$) or near dry ($h \sim 0$) states

# Well-Balanced Positivity Preserving Central-Upwind Scheme

[Kurganov, Petrova; 2007]

- $w = h + Z$: water surface $\Longrightarrow$ "Lake at rest" states: $\boxed{q \equiv 0, \;\; w \equiv \text{Const}}$

$\Longrightarrow$ Reconstruct the equilibrium variables $w$ and $q$ rather than $h$ and $q$

- Use the well-balanced quadrature

$$\int\limits_{x_{j-\frac{1}{2}}}^{x_{j+\frac{1}{2}}} hZ_x\,dx = \left(\overline{w}_j - \frac{Z(x_{j+\frac{1}{2}}) + Z(x_{j-\frac{1}{2}})}{2}\right) \cdot \left(Z(x_{j+\frac{1}{2}}) - Z(x_{j-\frac{1}{2}})\right)$$

- Make positivity preserving correction of the reconstruction of $w$

- Desingularize the computation of $u = \dfrac{q}{h}$ for small $h$

# Shallow Water System with Friction Terms

[Chertock, Cui, Kurganov, Wu; 2015]

$$\begin{cases} h_t + q_x = 0 \\ q_t + \left(hu^2 + \frac{g}{2}h^2\right)_x = -ghZ_x - g\frac{n^2}{h^{1/3}}|u|u \end{cases}$$

$n$: Manning coefficient

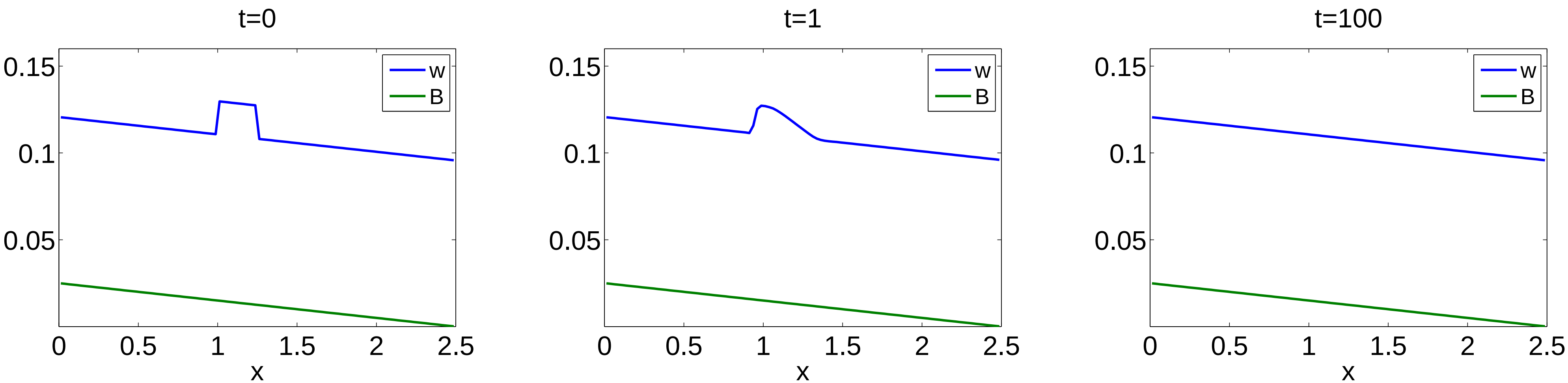
Special Steady-State Solutions

$$q \equiv \text{Const}, \quad h \equiv \text{Const}, \quad Z_x \equiv \text{Const}$$

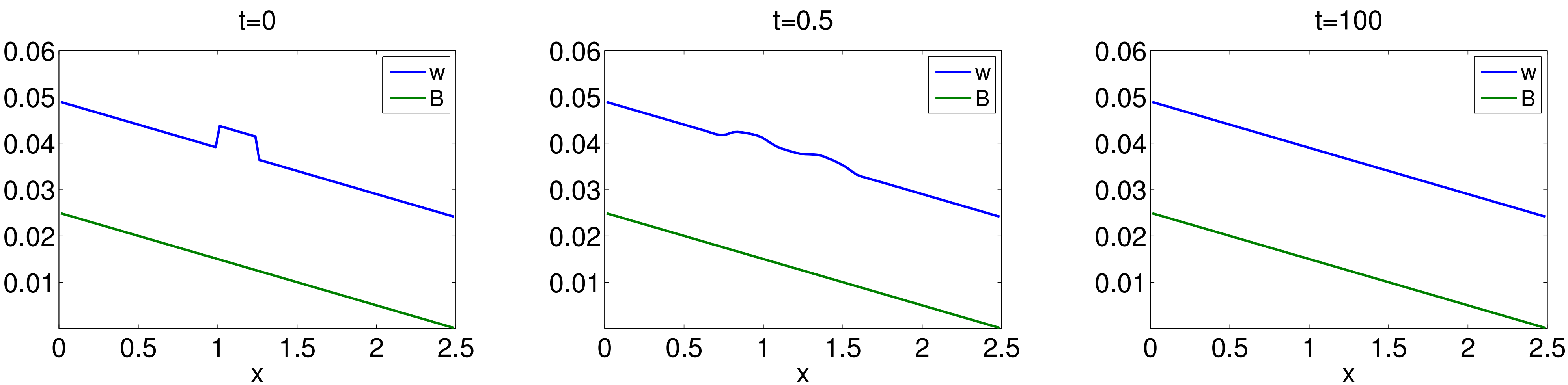
correspond to the situation when the water flows over a slanted infinitely long surface with a constant slope.

A straightforward midpoint discretization of the friction term leads to the well-balanced positivity preserving **semi-discrete** central-upwind scheme

# Example — Small Perturbation of a Steady Flow Over a Slanted Surface




Supercritical case




Subcritical case

## Example — Infinite Slanted Surface with a Periodic Flow

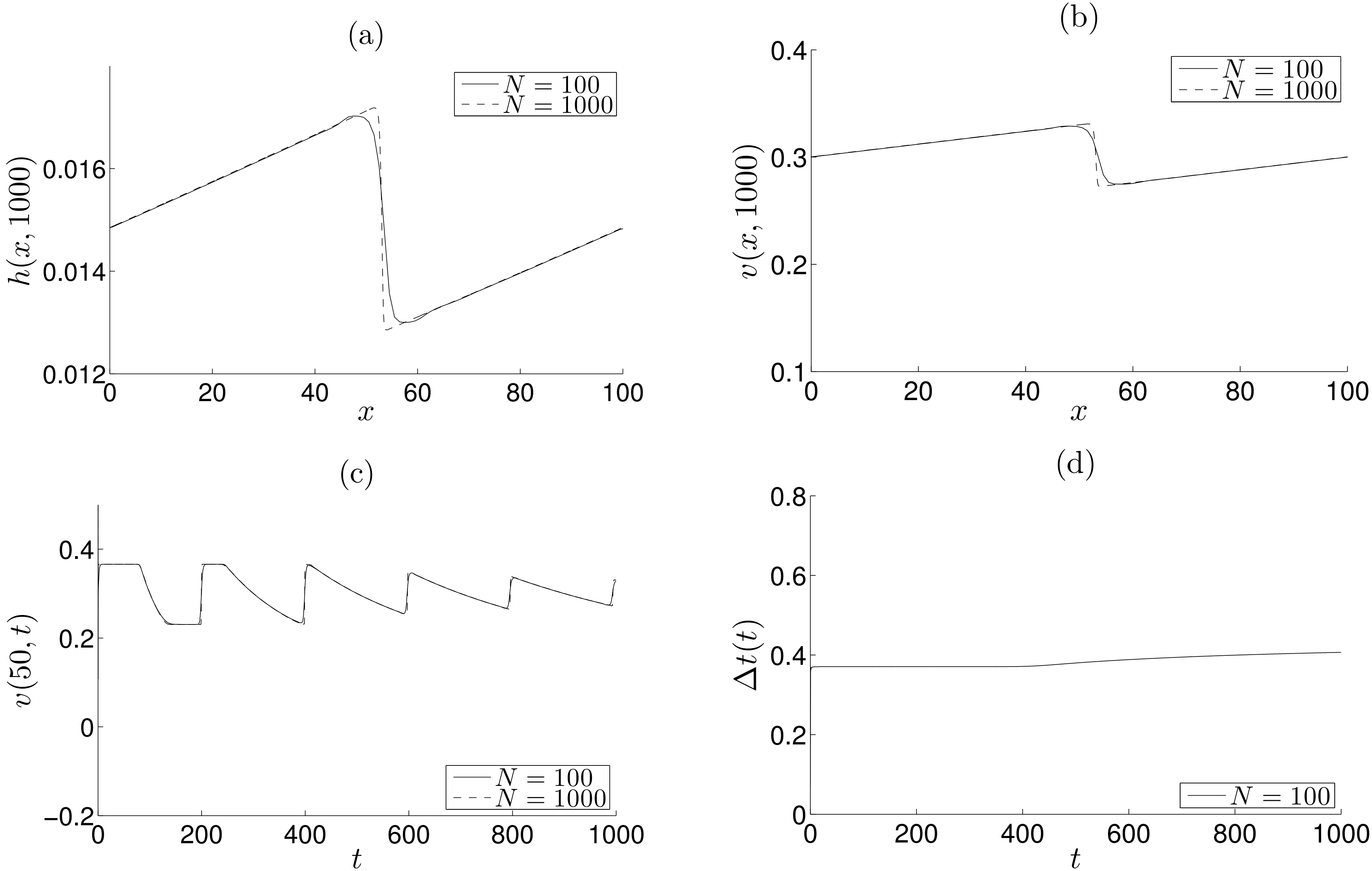
We take $Z_x \equiv -0.2$, $n = 0.09$ and the following initial conditions:

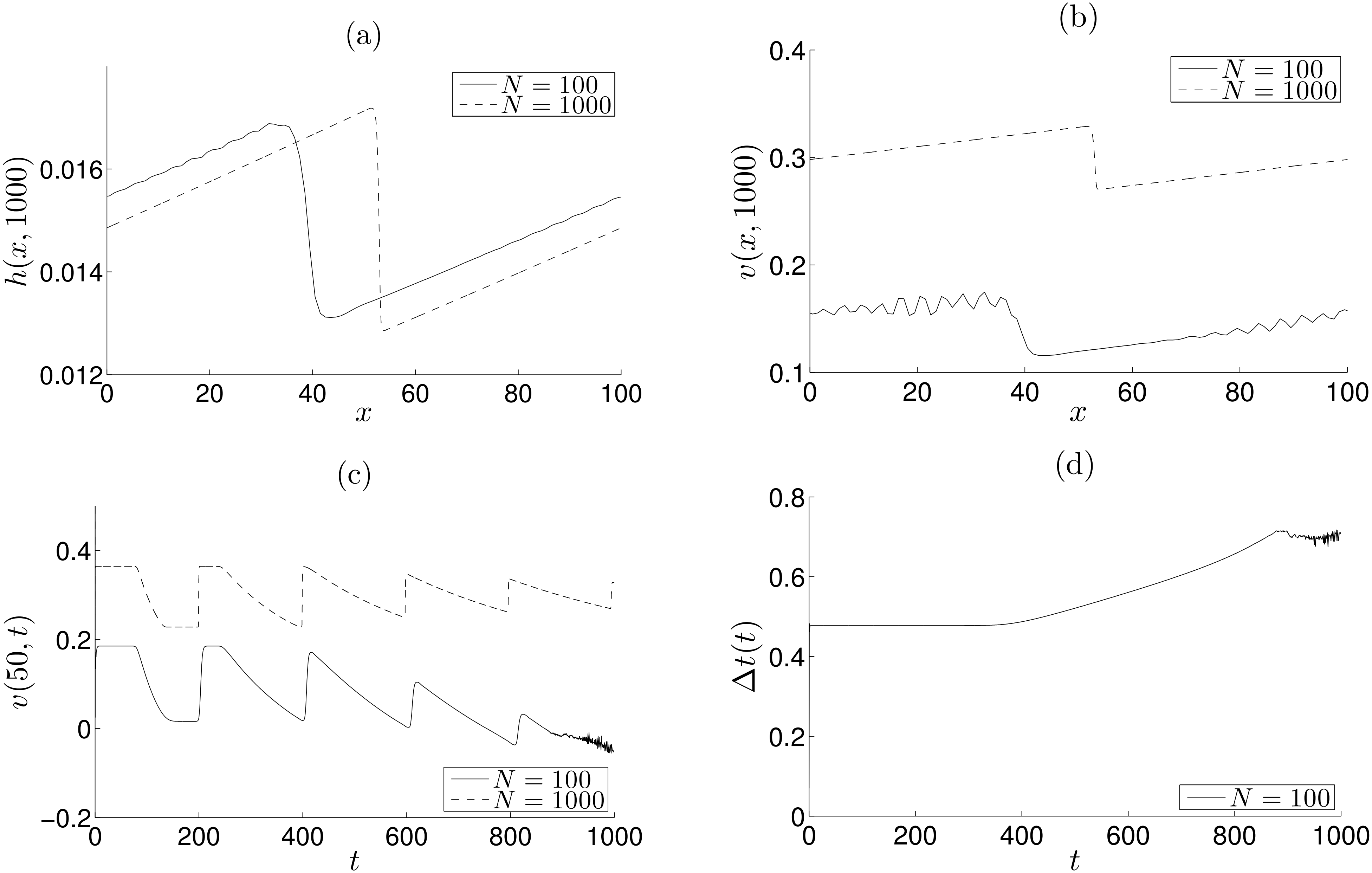
$$h(x,0) = \begin{cases} 0.02, & x < 50 \\ 0.01, & x > 50 \end{cases} \qquad q(x,0) = \begin{cases} 0, & x < 50 \\ 0.04, & x > 50 \end{cases}$$

We restrict the computational domain to $[0, 100]$, which is divided into $N$ uniform cells, and impose the periodic boundary conditions.

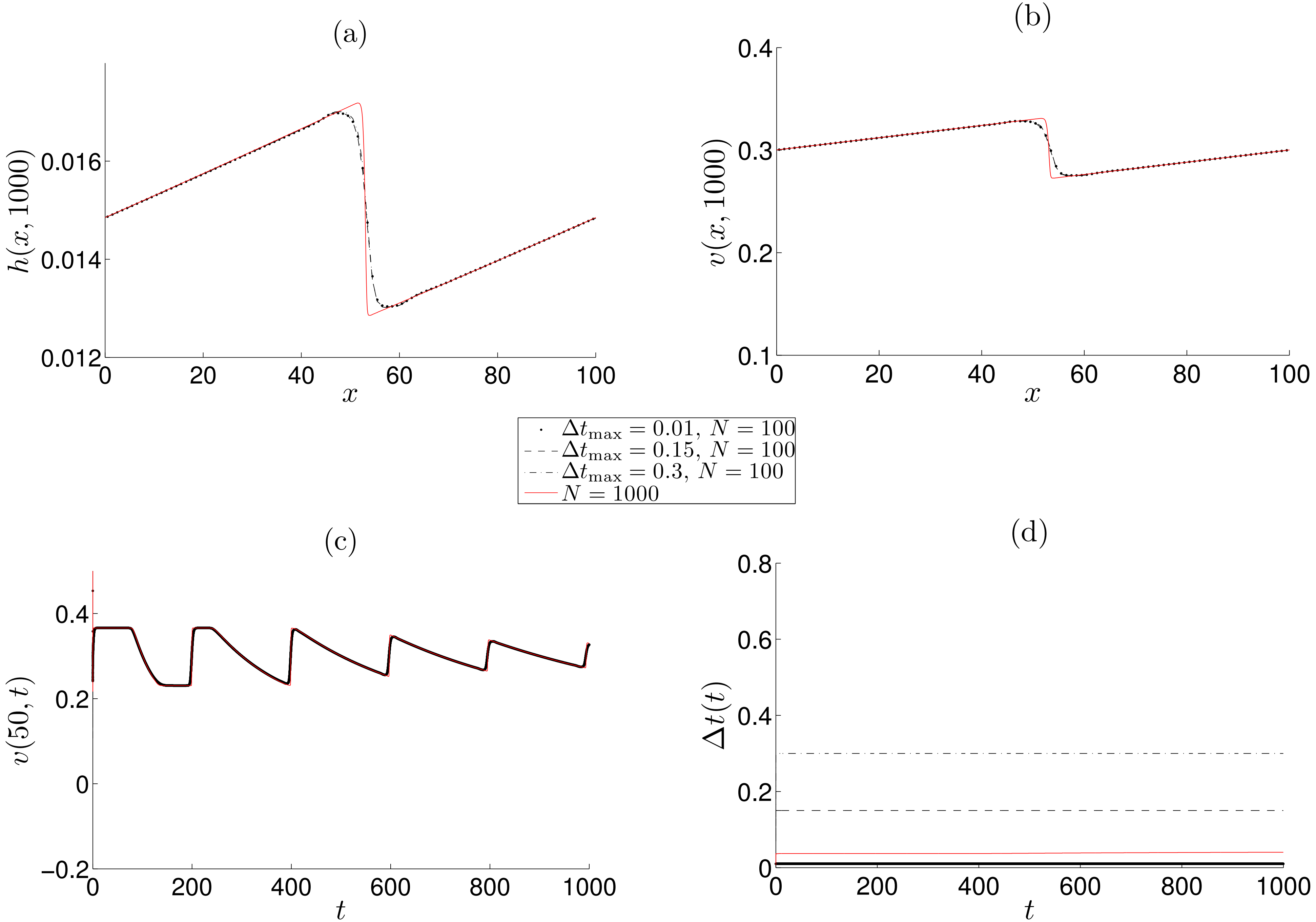
In this example, the friction term is very stiff and we compare the results obtained by the proposed second-order SI-RK3 method with the ones obtained using the second-order IMEX-SSP3(3,3,2) method.

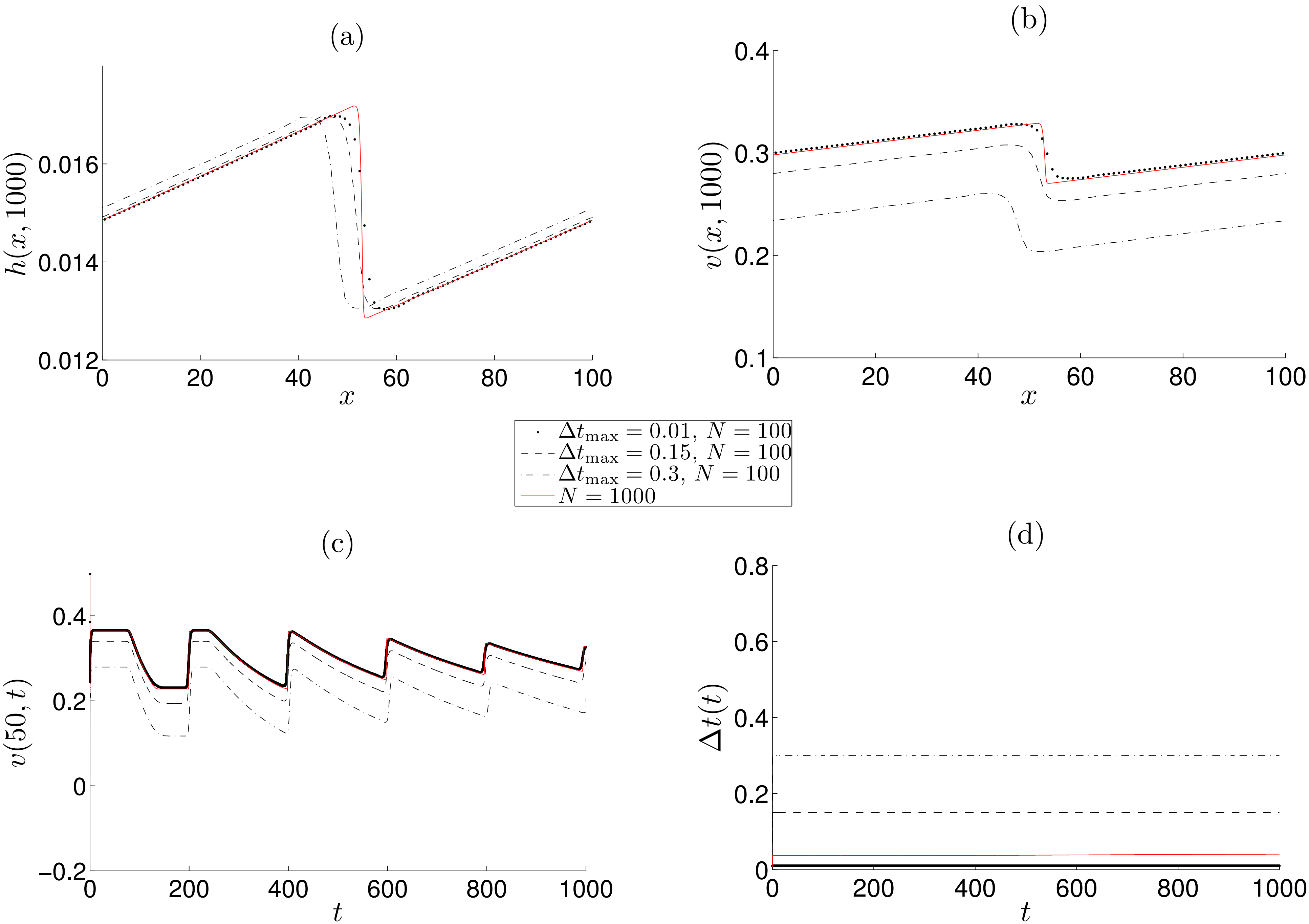
# Time Steps Restricted by the CFL Condition (the CFL number is 0.3)




(a)
$h(x, 1000)$
$x$
$N = 100$
$N = 1000$
0.012
0.014
0.016
0
20
40
60
80
100
(b)
$v(x, 1000)$
$x$
$N = 100$
$N = 1000$
0.1
0.2
0.3
0.4
0
20
40
60
80
100
(c)
$v(50, t)$
$t$
$N = 100$
$N = 1000$
−0.2
0
0.2
0.4
0
200
400
600
800
1000
(d)
$\Delta t(t)$
$t$
$N = 100$
0
0.2
0.4
0.6
0.8
0
200
400
600
800
1000

# Fixed Time Step Restriction ($\Delta t = \min\{\Delta t_{\mathrm{CFL}}, \Delta t_{\max}\}$)





(a)
h(x, 1000)
0.016
0.014
0.012
0
20
40
60
80
100
x
(b)
v(x, 1000)
0.4
0.3
0.2
0.1
0
20
40
60
80
100
x
Δt_max = 0.01, N = 100
Δt_max = 0.15, N = 100
Δt_max = 0.3, N = 100
N = 1000
(c)
v(50, t)
0.4
0.2
0
−0.2
0
200
400
600
800
1000
t
(d)
Δt(t)
0.8
0.6
0.4
0.2
0
0
200
400
600
800
1000
t