NTNU

Kunnskap for en bedre verden

TMA4212 - Numerical solution of differential equations by difference methods

Project 1

Authors:
Vemund Aakre, Thorbjørn C. Djupvik, Oskar F. Jakobsen

28th February 2025

1 Task 1

We consider reaction-diffusion equations, which have the form

$$u_t = \mu u_{xx} + f(u) \tag{1}$$

where μ is a positive constant. We also assume that the reaction term, $f(u)$, is a linear function in u. That is, it can be written as $f(u) = au$ for some constant $a \in \mathbb{R}, a \neq 0$. Furthermore, we solve the equation on the grid $(x,t) \in [0,1] \times [0,T]$ with boundary conditions given by functions f, g_1 and g_2:

$$\begin{aligned} u(0,x) &= f(x) \\ u(t,0) &= g_1(t) \\ u(t,1) &= g_2(t) \end{aligned}$$

1.1 Discretization

We discretize the domain in both space and time, such that $x_m = mh$ and $t_n = nk$, where $m = 0, 1, \ldots, M+1$ and $n = 0, 1, \ldots, N$ for some positive integers M, N. We denote by $U_m^n = U(t_n, x_m)$ the approximation of the exact solution $u_m^n = u(t_n, x_m)$.

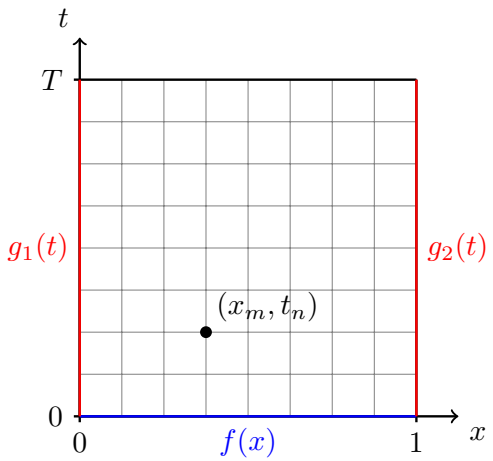

Figure 1: The domain $[0,1] \times [0,T]$, with boundary and initial conditions given by functions f, g_1 and g_2.

A scheme based on forward and backward Euler, together with a central difference in space, could be

$$\frac{1}{k}\nabla_t U_m^{n+1} = \frac{\mu}{h^2}\delta_x^2 U_m^{n+1} + f(U_m^n),$$

which can be rewritten to

$$U_m^{n+1} = U_m^n + r(U_{m+1}^{n+1} - 2U_m^{n+1} + U_{m-1}^{n+1}) + kf(U_m^n), \quad r = \mu\frac{k}{h^2}. \tag{2}$$

Time dependent PDEs with diffusion terms should be solved using implicit methods. This requires solving a nonlinear system for each step. The following scheme is based on an implicit method for the diffusion term and an explicit method for the reaction term.

$$U_m^* = U_m^n + \frac{r}{2}(\delta_x^2 U_m^* + \delta_x^2 U_m^n) + kf(U_m^n) \tag{3}$$

$$U_m^{n+1} = U_m^* + \frac{k}{2}(f(U_m^*) - f(U_m^n)) \tag{4}$$

$$U_0^n = g_1(t_n), \quad U_{M+1}^n = g_2(t_n) \tag{5}$$

1.2 Stability analysis

Define the matrix $S = \text{tridiag}(1, -2, 1) \in \text{Mat}_{M,M}(\mathbb{R})$. That is,

$$S = \begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{6}$$

Let $U^n = [U_1^n, U_2^n, \ldots, U_M^n]^T$. Furthermore, let I be the $M \times M$ identity matrix and let $\rho(C)$ denote the spectral radius of the matrix C. The scaled 2-norm will come in handy. It is defined as follows:

$$\|v\|_{2,h} = \sqrt{h}\|v\|_2$$

Before we go on with the stability analysis we need some results concerning the matrix S.

Lemma 1.1. *Let $S \in Mat_{M,M}(\mathbb{R})$ be defined as in (6). Then S is diagonalizable by an orthogonal matrix P. That is, $S = P\Lambda P^T$, for some diagonal matrix Λ. The eigenvalues of S are found on the diagonal of Λ:*

$$\lambda_m = -4\sin^2\phi_m,$$

where $\phi_m = \frac{m\pi}{2(M+1)}$ for $m = 1, \ldots, M$.

Proof. Problem 1 of exercise set 1. □

Lemma 1.2. *Let $S \in Mat_{M,M}(\mathbb{R})$ be defined as in (6), and let r be defined as in (2). Then $I - \frac{r}{2}S$ is invertible.*

Proof. $I - \frac{r}{2}S$ is invertible iff all the eigenvalues are nonzero. From lemma 1.1 we we have a diagonalization of S. Thus

$$I - \frac{r}{2}S = PP^T - \frac{r}{2}P\Lambda P^T = P\left(I - \frac{r}{2}\Lambda\right)P^T$$

The eigenvalues of $I - \frac{r}{2}S$ are $1 - \frac{r}{2}\lambda_m$, where the set $\{\lambda_m\}_{m=1,\ldots,M}$ are the eigenvalues of S. Observe that $r > 0$ and $\lambda_m < 0$ for all $m = 1, \ldots, M$. Hence

$$1 - \frac{r}{2}\lambda_m > 1 \neq 0$$

for all $m = 1, \ldots, M$, so the matrix is invertible. □

Rewrite (3) as a matrix-vector equation and collect U^* on the left hand side.

$$\left(I - \frac{r}{2}S\right)U^* = \left(I + kaI + \frac{r}{2}S\right)U^n \tag{7}$$

$I - \frac{r}{2}S$ is invertible by lemma 1.2. We substitute for U^* in (4) and arrive at an expression for the matrix C satisfying $U^{n+1} = CU^n$.

$$\begin{aligned} U^{n+1} &= U^* + \frac{ka}{2}(U^* - U^n) \\ &= \left(1 + \frac{ka}{2}\right)\left[I - \frac{r}{2}S\right]^{-1}\left(I + kaI + \frac{r}{2}S\right)U^n - \frac{ka}{2}U^n \\ &= \left[I - \frac{r}{2}S\right]^{-1}\left(\left(1 + ka + \frac{1}{2}(ka)^2\right)I + \frac{1}{2}(1 + ka)\,rS\right)U^n \end{aligned}$$

So

$$C = \left[I - \frac{r}{2}S\right]^{-1}\left(\left(1 + ka + \frac{1}{2}(ka)^2\right)I + \frac{1}{2}(1 + ka)\,rS\right). \tag{8}$$

By exploiting lemma 1.1 once more we arrive at a diagonalization for C.

$$\begin{aligned} C &= \left[PP^T - \frac{r}{2}P\Lambda P^T\right]^{-1}\left(\left(1+ka+\frac{1}{2}(ka)^2\right)PP^T + \frac{1}{2}(1+ka)\,rP\Lambda P^T\right) \\ &= P\left[I-\frac{r}{2}\Lambda\right]^{-1}\left(\left(1+ka+\frac{1}{2}(ka)^2\right)I + \frac{1}{2}(1+ka)\,r\Lambda\right)P^T \\ &= P\Delta P^T \end{aligned}$$

where

$$\Delta = \left[I-\frac{r}{2}\Lambda\right]^{-1}\left(\left(1+ka+\frac{1}{2}(ka)^2\right)I + \frac{1}{2}(1+ka)\,r\Lambda\right) \tag{9}$$

We observe that C is symmetric since $C = P\Delta P^T = P\Delta^T P^T = \left(P\Delta P^T\right)^T = C^T$. This is nice since now the condition $\rho(C) \leq 1+\nu k$ is both necessary *and sufficient* for stability when we use $||\cdot||_{2,h}$. In particular we have that $||C||_{2,h} = \rho(C)$ for a symmetric matrix C.

The eigenvalues for C are found on the diagonal of Δ.

$$\Delta_m = \frac{1+ka+\frac{1}{2}(ka)^2+\frac{1}{2}(1+ka)\,r\lambda_m}{1-\frac{r}{2}\lambda_m} \tag{10}$$

Bounding $\rho(C) = \max_m|\Delta_m|$.

$$\begin{aligned} |\Delta_m| &= \left|\frac{1+ka+\frac{1}{2}(ka)^2+\frac{1}{2}(1+ka)\,r\lambda_m}{1-\frac{r}{2}\lambda_m}\right| \\ &= \left|(1+ka)\,\frac{1+\frac{1}{2}r\lambda_m}{1-\frac{1}{2}r\lambda_m} + \frac{1}{2}(ka)^2\frac{1}{1-\frac{1}{2}r\lambda_m}\right| \\ &\leq |1+ka|\left|\frac{1+\frac{1}{2}r\lambda_m}{1-\frac{1}{2}r\lambda_m}\right| + \frac{1}{2}(ka)^2\left|\frac{1}{1-\frac{1}{2}r\lambda_m}\right| \end{aligned}$$

Note that $-1 < \lambda_m < 0$ for all $m = 1,\ldots,M$. In addition, the step size k is bounded by the length of the time domain, T. We simplify further.

$$|\Delta_m| \leq |1+ka| + \frac{1}{2}(ka)^2 = 1 + (|a| + \frac{1}{2}ka^2)k \leq 1 + (|a| + \frac{1}{2}Ta^2)k = 1+\nu k$$

Finally, we have arrived at the expression needed for stability without imposing any conditions on the scheme. We summarize our discussion in a theorem.

Theorem 1.3. *The scheme given in equations* (3) *and* (4) *is unconditionally stable with respect to* $\|\cdot\|_{2,h}$.

1.3 Consistency

Theorem 1.4. *Given that u is sufficiently smooth and $k \neq \frac{-2}{a}$, the local truncation error of the method is of order $\mathcal{O}(k^2+h^2)$.*

Proof. For the boundaries we get $\tau_0 = \tau_{M+1} = 0$ since we have Dirichlet boundary conditions and the approximations are exact. For the inner points we rewrite (4) to an explicit equation for U_m^*,

$$U_m^* = \frac{U_m^{n+1} + \frac{ka}{2}U_m^n}{1+\frac{ka}{2}}.$$

This is then substituted into (3) to remove U_m^* from the equation. Multiply by $(1+\frac{ka}{2})$ and rearrange the terms.

$$U_m^{n+1} = \left(1+ka+\frac{1}{2}(ka)^2\right)U_m^n + \frac{r}{2}\left((1+ka)\delta_x^2 U_m^n + \delta_x^2 U_m^{n+1}\right)$$

Then the approximations are replaced by u and the local truncation error is therefore introduced.

$$k\tau_m^n + u_m^{n+1} = \left(1 + ka + \frac{1}{2}(ka)^2\right) u_m^n + \frac{r}{2}\left((1+ka)\delta_x^2 u_m^n + \delta_x^2 u_m^{n+1}\right)$$

Taylor expand around u_m^n and substitute ∂_x^2 for δ_x^2, picking up some more error terms.

$$\begin{aligned} k\tau_m^n &= \left(1 + ka + \frac{1}{2}(ka)^2\right) u_m^n - u_m^{n+1} + \frac{\mu k}{2h^2}\left(h^2\partial_x^2 u_m^{n+1} + (1+ka)\, h^2 \partial_x^2 u_m^n + \mathcal{O}(h^4)\right) \\ &= \left(1 + ka + \frac{1}{2}(ka)^2\right) u_m^n - u_m^n - k\partial_t u_m^n - \frac{k^2}{2}\partial_t^2 u_m^n + \mathcal{O}(k^3) \\ &\quad + \frac{\mu k}{2}\left(\partial_x^2\left(u_m^n + k\partial_t u_m^n + \mathcal{O}(k^2)\right) + (1+ka)\,\partial_x^2 u_m^n\right) + \mathcal{O}(kh^2) \end{aligned}$$

This expression is then divided by k and rearranged to

$$\tau_m^n = a u_m^n - \partial_t u_m^n + 2\frac{\mu}{2}\left(\partial_x^2 u_m^n\right) + \frac{ka^2}{2}u_m^n - \frac{k}{2}\partial_t^2 u_m^n + \frac{\mu}{2}\left(k\partial_x^2\partial_t u_m^n + ka\partial_x^2 u_m^n\right) + \mathcal{O}(k^2 + h^2).$$

The first three terms are simply (1) with all terms moved to the right hand side. Some terms can also be rewritten using (1), $\mu\partial_x^2 u_m^n = \partial_t u_m^n - a u_m^n$.

$$\begin{aligned} \tau_m^n &= \frac{ka^2}{2}u_m^n - \frac{k}{2}\partial_t^2 u_m^n + \frac{k}{2}\partial_t\left(\partial_t u_m^n - a u_m^n\right) + \frac{ka}{2}\left(\partial_t u_m^n - a u_m^n\right) + \mathcal{O}(k^2 + h^2) \\ &= \frac{ka^2}{2}u_m^n - \frac{ka^2}{2}u_m^n - \frac{k}{2}\partial_t^2 u_m^n + \frac{k}{2}\partial_t^2 u_m^n - \frac{ka}{2}\partial_t u_m^n + \frac{ka}{2}\partial_t u_m^n + \mathcal{O}(k^2 + h^2) \end{aligned}$$

Many terms cancel and we are left with $\tau_m^n = \mathcal{O}(k^2 + h^2)$. □

Corollary 1.4.1. *$\|\tau^s\|_{2,h} = \mathcal{O}(k^2 + h^2)$ for all $s = 0, \ldots, N$.*

Proof. We use theorem 1.4 for an upper bound of $|\tau_m^n|$.

$$\|\tau^s\|_{2,h} = \sqrt{h}\left(\sum_{i=0}^{M+1} |\tau_i^s|^2\right)^{\frac{1}{2}} \leq \sqrt{h}\left(\sum_{i=0}^{M+1} R(k^2+h^2)\right)^{\frac{1}{2}} = \sqrt{h}\left[(M+2)R^2(k^2+h^2)^2\right]^{\frac{1}{2}}.$$

Furthermore, $h = \frac{1}{M+1}$, so

$$\|\tau^s\|_{2,h} \leq \frac{\sqrt{M+2}}{\sqrt{M+1}} R(k^2+h^2) \leq \sqrt{\frac{3}{2}} R(k^2+h^2)$$

□

Theorem 1.5. *Given that u is sufficiently smooth and $k \neq \frac{-2}{a}$, the scheme is convergent of order $\mathcal{O}(k^2+h^2)$ with respect to $\|\cdot\|_{2,h}$.*

Proof. The method is stable, theorem 1.3, and consistent, theorem 1.4, and therefore convergent by Lax' equivalence theorem. We also need to prove the order of convergence. The scheme can be written in the form

$$\left(I - \frac{r}{2}S\right) U_m^{n+1} = \left(\left(1 + ka + \frac{1}{2}(ka)^2\right) I + \frac{1}{2}(1+ka)\, rS\right) U_m^n$$

or using $A = \left(I - \frac{r}{2}S\right)$ and $B = \left(\left(1 + ka + \frac{1}{2}(ka)^2\right) I + \frac{1}{2}(1+ka)\, rS\right)$ as

$$AU_m^{n+1} = BU_m^n.$$

If we place $U_m^n - u_m^n$ into the equation instead, we get

$$AE^{n+1} = BE^n - k\tau^n.$$

Exchanging $n+1$ with n, multiplying by A^{-1}, which exists by lemma 1.2

$$E^{n+1} = A^{-1}BE^n - kA^{-1}\tau^n$$

we set $q^n = -kA^{-1}\tau^n$ and $C = A^{-1}B$.

$$E^{n+1} = CE^n + q^n$$

Using the term recursively, we get

$$E^{n+1} = C^{n+1}E^0 + C^n q^0 + C^{n-1}q^1 + \cdots + Cq^{n-1} + q^n.$$

Taking the norm on both sides and using stability from theorem 1.3 to give a bound L for $\|C^{n+1}\|$.

$$\|E^{n+1}\|_{2,h} \leq L\sum_{s=0}^{n}\|q^s\|_{2,h}$$

We the use that

$$\|q^s\|_{2,h} \leq k\|A^{-1}\|_{2,h}\|\tau^s\|_{2,h} \leq k\tilde{K}\|\tau^s\|_{2,h}$$

and, using $(n+1)k \leq (N+1)k = T + k \leq 2T = \tilde{T}$, obtain

$$\|E^{n+1}\|_{2,h} \leq L\tilde{K}(n+1)k \max_{0\leq s\leq n}\|\tau^s\|_{2,h} \leq L\tilde{K}\tilde{T}\max_{0\leq s\leq n}\|\tau^s\|_{2,h}.$$

By corollary 1.4.1

$$\|E^{n+1}\|_{2,h} \leq L\tilde{K}\tilde{T}\tilde{R}(k^2+h^2).$$

Hence, $\|E^{n+1}\|_{2,h} = \mathcal{O}(k^2+h^2)$. □

1.4 Numerical experiments

We use the following solution to construct a test equation:

$$\tilde{u}(t,x) = \mathrm{e}^{-(\mu b+a)t}\sin(bx+\varphi) \tag{11}$$

μ and a are the constants from the original problem. Then we set $\mu = 1/5, b = 3\pi/2, a = 1$ and $\varphi = \pi/4$, and solve it on the domain $[0,1]\times[0,1]$. The convergence plots in figure 2 show that the order of convergence is quadratic both in time and space. This aligns with theorem 1.5.

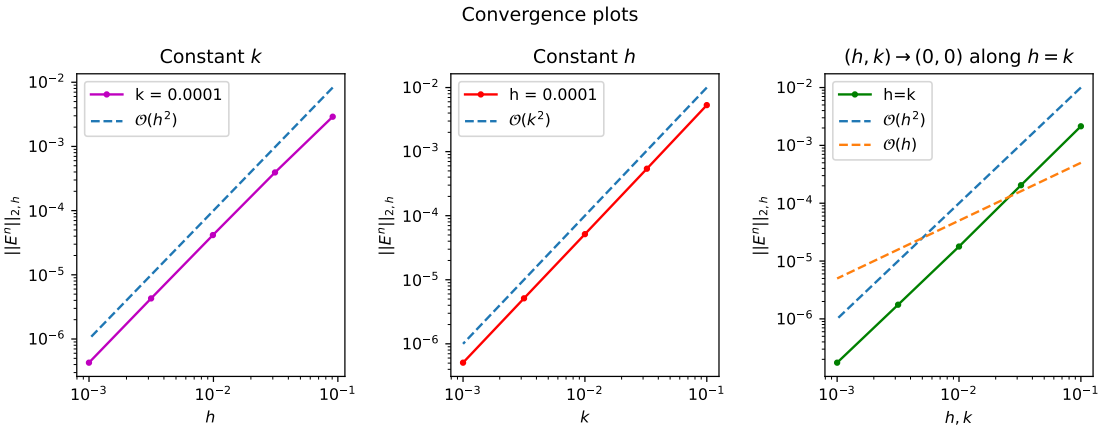

Figure 2: The global error decreases quadratically with the step sizes k and h. When both k and h tends to zero simultaneously, the global error also decreases quadratically.

2 Task 2

A well known model for the spread of infectious disease is the SIR-model. The model describes the population at a time t as

- **Susceptible** S: A person is susceptible if they can get the disease.
- **Infected and infectious** I: A person is infective if they have the disease and can transfer it to others.
- **Recovered** R: A person is recovered if they can not be infected or infect others.

A simple version of SIR-model is

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS \\ \frac{dI}{dt} &= \beta IS - \gamma I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{12}$$

Where β describes the infectiousness of the model, γ describes how quickly the infected become removed. However, this does not describe the spatial spread of a disease, which can be accommodated for by modifying the equations slightly. One proposed modification is given by Murray [1], chapter 13, which is adding two dispersion terms to account for the spread in space, which gives the following equations

$$\begin{aligned} \frac{dS}{dt} &= -\beta IS + \mu_S \Delta S \\ \frac{dI}{dt} &= \beta IS - \gamma I + \mu_I \Delta I \\ \frac{dR}{dt} &= \gamma I. \end{aligned} \tag{13}$$

Where μ_S and μ_I describe how much the susceptible and infected move in space. Combined with initial values and boundary conditions, this can be solved with numerical schemes.

2.1 Numerical scheme

In the following, we further assume that $S(t) + I(t) + R(t) = 1$, so that we are always working with percentages of the total population instead of the actual population. We consider the case with two spatial dimensions. To solve the equations we will use the Forward-Euler method, which for (13) results in the following scheme

$$\begin{aligned} S_{n,m}^{l+1} &= S_{n,m}^{l} + k(-\beta I_{n,m}^{l} S_{n,m}^{l} + \mu_S \Delta_h S_{n,m}^{l}) \\ I_{n,m}^{l+1} &= I_{n,m}^{l} + k(\beta I_{n,m}^{l} S_{n,m}^{l} - \gamma I_{n,m}^{l} + \mu_I \Delta_h I_{n,m}^{l}) \\ R_{n,m}^{l+1} &= R_{n,m}^{l} + k\gamma I_{n,m}^{l}, \end{aligned} \tag{14}$$

where

$$\Delta_h X_{n,m}^{l} = \frac{1}{h^2}(X_{n+1,m}^{l} + X_{n-1,m}^{l} + X_{n,m+1}^{l} + X_{n,m-1}^{l} - 4X_{n,m}^{l}).$$

Here k is the stepsize in time, h the stepsize in both spatial directions and X denotes one of the variables, S, I or R. $X_{n,m}^{l}$ corresponds to the numerical solution at (x_n, y_m, t_l). It should be noted that the Forward-Euler scheme is not stable for all combinations of h and k. For the equation $\frac{\partial u}{\partial t} = \alpha \Delta u$, it is stable for all (h, k) such that $\frac{\alpha k}{h^2} < \frac{1}{4}$. This gives a rough estimate for when it is stable for our scheme as well.

The initial values describe the initial distributions of susceptible, infected and recovered people, while the boundary conditions describe how the different groups move across the boundary of the

problem. Since we want the total population to be constant on the domain, we set the boundary conditions to be

$$\begin{aligned} \nabla S(\partial\Omega) &= \vec{0} \\ \nabla I(\partial\Omega) &= \vec{0} \\ \nabla R(\partial\Omega) &= \vec{0}, \end{aligned} \tag{15}$$

where $\partial\Omega$ is the edge of the domain. To discretize this in order $\mathcal{O}(h^2)$, it becomes

$$\frac{\Delta_h X^l_{n,m} - \nabla_h X^l_{n,m}}{2h} = 0 \tag{16}$$

along the boundaries. Here ∇_h and Δ_h operate in x if $n = 0$ or $n = N + 1$ and in y if $m = 0$ or $m = N + 1$, and in both directions where both conditions apply. To implement this, an extra false boundary is created such that the total grid in space becomes $[-1, N + 2] \times [-1, N + 2]$. This false boundary will not be part of the model and only used for the numerical solution. The boundary conditions, equations (16), are then

$$\begin{aligned} X^l_{n,-1} &= X^l_{n,1} \\ X^l_{n,N+2} &= X^l_{n,N} \\ X^l_{-1,m} &= X^l_{1,m} \\ X^l_{N+2,m} &= X^l_{N,m} \end{aligned} \tag{17}$$

on a square grid for $n \in [1, N]$.

2.2 Results

As our main exploration we varied both population density and β over the grid so we effectively got four areas with different properties. The top right corner has a high number of infected people, while both the population and β is somewhat small. This is the only area where there are infected people in the beginning. The bottom right corner has a low β but a high population, the bottom left has both a high β and a high population, and the top left has a high β but a low population. In figure (3) we can see the time evolution of the infected. As expected the area with both a high population and high β gets infected pretty quickly, while the area with a low β remains unaffected for a longer time. The area with both low population and high β gets infected pretty fast, although one can barely see it in the plot. If we look at (4), one interesting thing to note is that some susceptible seem to migrate towards the upper right corner and thereby avoiding the infected somewhat.

Another thing to note is the population size. Although the boundary conditions are set to enforce a constant population size, we still observe a change in population size because of errors in the numerical scheme. As (5) shows, the population increases slightly during the simulation. A possible solution would be to not solve for $R(t)$ by equation (14), but rather to let $R(t) = 1 - S(t) - I(t)$ such that the total population is constant. This will effectively just hide the error in the $R(t)$ term. For this reason we solve the full model as is, and accept an error in the total population size.

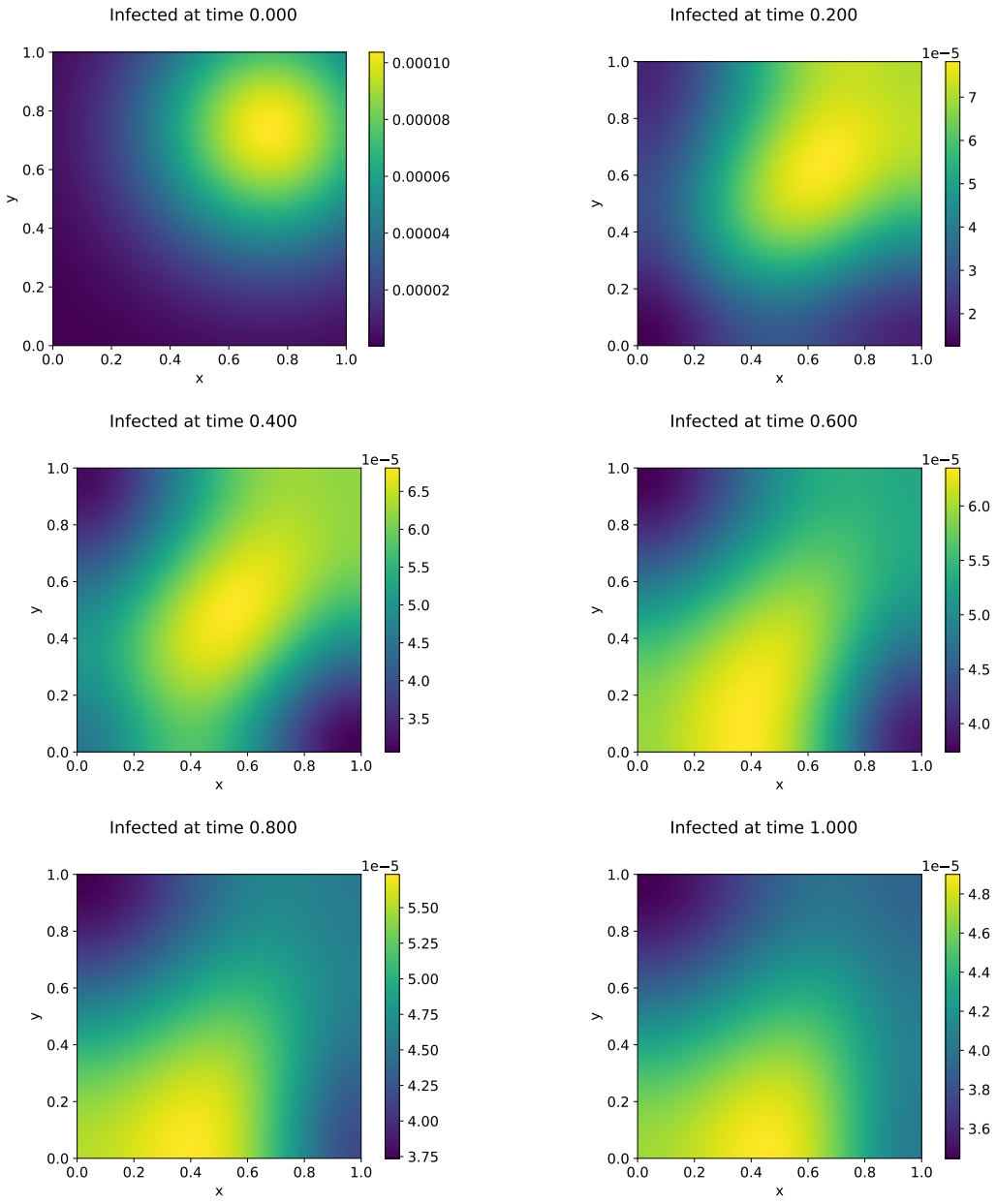

Figure 3: Movement of infection areas. The movement of the areas of infection is interesting, not the variation in scaling. $\gamma = 1$, $\mu_I = 0.1$, $\mu_S = 0.1$, $\beta = 100000e^{-5((x-0.25)^2+\frac{(y-0.5)^2}{10})}$.

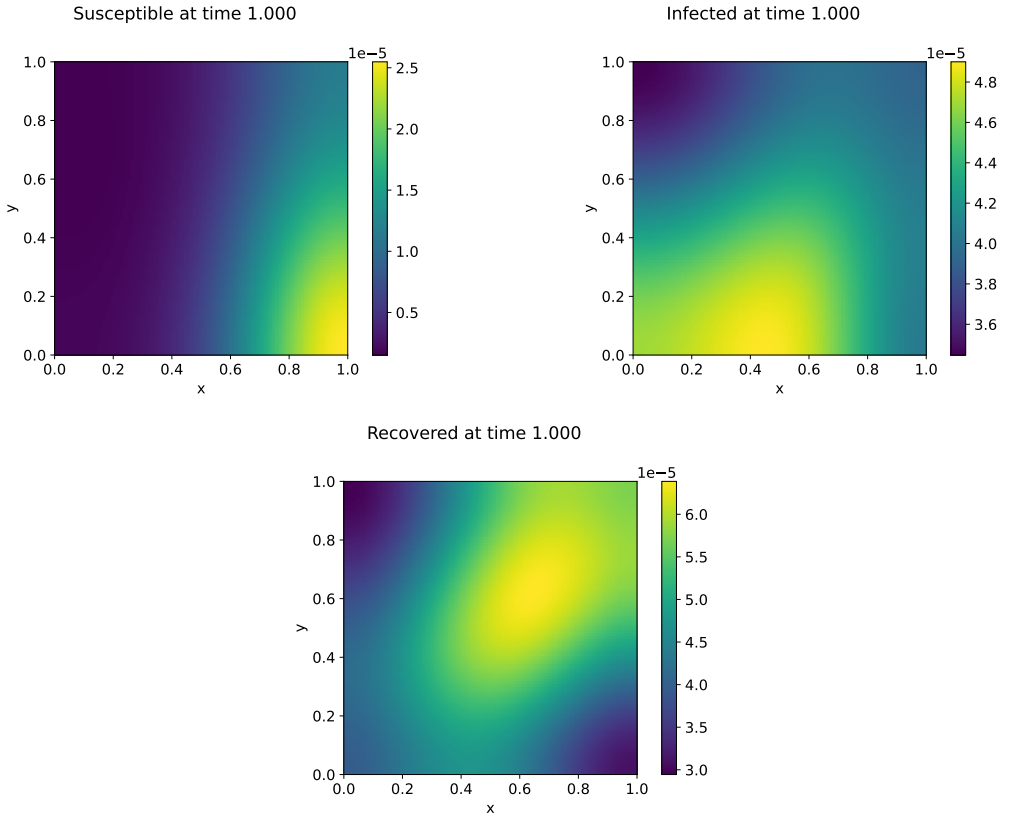

Figure 4: Final state of the SIR model. $\gamma = 1, \mu_I = 0.1, \mu_S = 0.1, \beta = 100000e^{-5((x-0.25)^2+\frac{(y-0.5)^2}{10})}$.

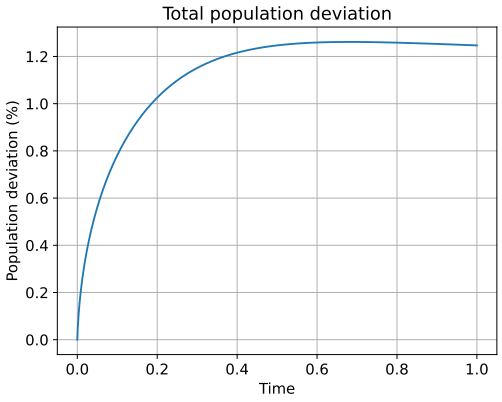

Figure 5: The deviation from the standard population we begin with as a function of time.

The last thing to note is the effect of the parameters on the equation. The size of the spatial domain greatly influences the other parameters. The larger domain, the larger the step size in that domain in order to get the same runtime, but that includes larger errors. In addition, for small stepsizes, the constants such as β, μ_I, μ_S need to be larger in order to get the same results.

Bibliography

[1] J. D. Murray. *Mathematical Biology II: Spatial Models and Biomedical Applications*. Vol. 18. Interdisciplinary Applied Mathematics. Springer New York, 2003. DOI: 10.1007/b98869.