4.3. **Numerics for the heat equation.** We now study numerical solutions for the heat equation in more detail. Although we can solve the heat equation analytically in many cases, it is useful to study its numerical solutions, because for such a simple equation we have hope of understanding the numerical scheme really well. The equation we will always consider is the heat equation on the interval $[0,1]$:

$$\begin{aligned}&\partial_t u(x,t) = \frac{1}{2}\sigma^2\partial_x^2 u(x,t) \qquad (t>0, 0<x<1),\\ &u(x,0) = u_0(x), \qquad u(0,t) = u(1,t) = 0.\end{aligned}$$

The forward scheme for this equation is (with $t_n = hn$ and $x_j = h_x j$):

$$u_j^{n+1} = u_j^n + \frac{\sigma^2}{2}\frac{h}{h_x^2}(u_{j+1}^n + u_{j-1}^n - 2u_j^n). \tag{4.7}$$

with $u_j^0 = u_0(x_j)$, $u_0^n = 0$ and $u_J^n = 0$, where J is such that $h_x J = 1$. The ratio $\frac{1}{2}\sigma^2 h/h_x^2$ will appear a lot below, and for convenience we define $\mu = \sigma^2 h/h_x^2$.
The first question is whether the scheme (4.7) is consistent. This is easily seen to be true. Recall the definition of the truncation error. First we bring everything in (4.7) to one side, and divide by h. This gives

$$\frac{1}{h}(u_j^{n+1} - u_j^n) - \frac{\sigma^2}{2}\frac{1}{h_x^2}(u_{j+1}^n + u_{j-1}^n - 2u_j^n) = 0$$

The truncation error is the results of replacing the u_j^n above with the true solution. Thus,

$$T(x_j,t_n) := \frac{1}{h}(u(x_j,t_{n+1}) - u(x_j,t_n)) - \frac{\sigma^2}{2}\frac{1}{h_x^2}(u(x_{j+1},t_n) + u(x_{j-1},t_n) - 2u(x_j,t_n)).$$

We can now Taylor expand this expression around $u(x_j,t_n)$ and find that

$$T(x_j,t_n) = \partial_t u(x_j,t_n) + \frac{h}{2}\partial_t^2 u(x_j,t_n) + \ldots - \frac{\sigma^2}{2}\partial_x^2 u(x_j,t_n) - \frac{\sigma^2}{4!}h_x^2\partial_x^4 u(x_j,t_n) + \ldots,$$

where the terms $\ldots$ come with higher powers of h and h_x. $\partial_t u - \frac{\sigma^2}{2}\partial_x^2 u = 0$ since u is a true solution of the heat equation, and we see that $\lim_{h,h_x\to 0} T(x_j,t_n) = 0$. So the scheme is consistent.
How about convergence? There is little hope of convergence unless $\mu = \frac{1}{2}\sigma^2 h/h_x^2$ stays at least bounded as h and h_x go to zero. But even this is not enough. As one can see by experimenting with any computer implementation of the scheme, strong oscillations will build up and the numerical solution will diverge unless $\mu \leqslant 1/2$. The numerical solutions for $\mu > 1/2$ will have nothing to do with the analytical solutions, and making the step size smaller will not help here.

To see why this is so, let us take another look at our scheme (4.7). Let us define the vector $\boldsymbol{u}^n = (u_1^n, \dots, u_{J-1}^n)$. Then (4.7) reads

$$\boldsymbol{u}^{n+1} = (\mathbf{1} + \mu A)\boldsymbol{u}^n, \quad \text{with } A = \begin{pmatrix} -2 & 1 & 0 & 0 & 0 & \cdots & 0 \\ 1 & -2 & 1 & 0 & 0 & \cdots & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 \\ \vdots & \ddots & \ddots & \ddots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & -2 & 1 & 0 \\ 0 & \cdots & 0 & 0 & 1 & -2 & 1 \\ 0 & \cdots & 0 & 0 & 0 & 1 & -2 \end{pmatrix} \tag{4.8}$$

A is a $(J-1) \times (J-1)$ matrix, and is called the discrete Laplacian. $\mathbf{1}$ denotes the unit matrix that has 1 on the diagonal and 0 elsewhere. From iterating (4.8) we conclude

$$\boldsymbol{u}^n = (\mathbf{1} + \mu A)^n \boldsymbol{u}^0.$$

So we need to compute high powers of the matrix $(\mathbf{1} + \mu A)$. For this, we need the eigenvalues and eigenvectors of A. Let us start by assuming that we have found them, i.e. let $\boldsymbol{v}_1, \dots, \boldsymbol{v}_{J-1}$ be the eigenvectors and $\lambda_1, \dots, \lambda_{J-1}$ be the eigenvalues. We write the initial condition using the basis of eigenvectors as

$$\boldsymbol{u}^0 = \sum_{k=1}^{J-1} \alpha_k \boldsymbol{v}_k.$$

We then find

$$\boldsymbol{u}^n = \sum_{k=1}^{J-1} \alpha_k (1 + \mu\lambda_k)^n \boldsymbol{v}_k. \tag{4.9}$$

Now we can already see under which circumstances the numerical solution will explode. Namely, we must assume that none of the α_k is zero - otherwise we would restrict ourselves to very special initial conditions that are orthogonal to some eigenvector of A. Thus, the expression (4.9) will stay bounded for large n if and only if

$$|1 + \mu\lambda_k| < 1 \qquad \text{for all } k. \tag{4.10}$$

In other words, for a sensible numerical scheme we need that the matrix $\mathbf{1} + \mu A$ has no eigenvalues of absolute value greater than one.

Let us now see what the eigenvalues of A actually are. While there are systematic ways to derive them, we will here just 'guess' them. Putting $(\boldsymbol{v}_k)_j = \sin(k\pi j/J)$, some arithmetic (in particular using $\sin(x \pm y) = \sin x \cos y \pm \cos x \sin y$) gives

$$(A\boldsymbol{v}_k)_j = \sin(k\pi(j+1)/J) + \sin(k\pi(j+1)/J) - 2\sin(k\pi j/J) = \lambda_k \boldsymbol{v}_k$$

with $\lambda_k = -2(1 - \cos(k\pi/J))$. This works for $k = 1, \dots J-1$. For $k = J$, the eigenvector would be the zero vector, which is not allowed, and for larger k, we get back the (negatives of) vectors that we already had. Anyway, we have found

the $J-1$ eigenvalues and eigenvectors of the the $(J-1)\times(J-1)$-matrix A. Now we plug this into (4.10), and get the condition

$$|1-2\mu(1-\cos(k\pi/J))|<1.$$

The term in brackets is always bigger than zero, and can be up to just below 2, which happens for $k=J-1$. So in order for the absolute value above to be always smaller than one, we will need $\mu\leqslant 1/2$. Otherwise, the contribution of the eigenvalue $k=J-1$ will dominate all others, which leads to the zig-zag line that we see in the numerical simulations.

The restriction $\mu\leqslant 1/2$ is a big deal. For $\sigma^2/2=1$, it means that $h<h_x^2/2$. So if we want to discretize space into an not unreasonable grid of points that are $1/100$ apart, we are forced to move in tiny time steps of $1/20000$. What can we do about this? The answer is to use the implicit scheme. To derive it, we do the Taylor expansion around t_{n+1} and x_j, and find

$$\partial_t u(x_j,t_{n+1})\approx\frac{1}{h}\Big(u(x_j,t_{n+1})-u(x_j,t_n)\Big),$$

and

$$\partial_x^2 u(x_j,t_{n+1})\approx\frac{1}{h_x^2}\Big(u(x_{j+1},t_{n+1})+u(x_{j-1},t_{n+1})-2u(x_j,t_{n+1})\Big).$$

This leads to the scheme

$$u_j^n=u_j^{n+1}-\frac{\sigma^2}{2}\frac{h}{h_x^2}(u_{j+1}^{n+1}+u_{j-1}^{n+1}-2u_j^{n+1}).$$

Why is this scheme better than the one we had before? Let us write the scheme in vector notation:

$$\boldsymbol{u}^n=\boldsymbol{u}^{n+1}-\mu A\boldsymbol{u}^{n+1}=(\mathbf{1}-\mu A)\boldsymbol{u}^{n+1},$$

where μ is the same as before. Therefore,

$$\boldsymbol{u}^{n+1}=(\mathbf{1}-\mu A)^{-1}\boldsymbol{u}^n.$$

We can easily find the eigenvalues of $(\mathbf{1}-\mu A)^{-1}$. Recall that the eigenvalues of A are $-2(1-\cos(k\pi/J))$ with $1\leqslant k\leqslant J-1$. Thus, the eigenvalues of $\mathbf{1}-\mu A$ are $1+2\mu(1-\cos(k\pi/J))$, with the same eigenvectors as A has. Finally, the eigenvalues of $(\mathbf{1}-\mu A)^{-1}$ are given by $\frac{1}{1+2\mu(1-\cos(k\pi/J))}$, with the same eigenvectors that A has (you should check this!). These eigenvalues are smaller than 1 for any μ. So, in this case the scheme is sensible for all μ, and we can e.g. take $\mu=1000$ if we want to do a fine space discretisation. This will then still lead to a reasonably large time step.

Let us look at a consequence of the fact that the matrix $(\mathbf{1}-\mu A)^{-1}$ has only eigenvalues of absolute value smaller than 1. Consider two initial conditions for the PDE that are very similar (you can think of one as the true initial condition, and the other one as some approximation to the true initial condition). Let us thus

assume that we have $\boldsymbol{u}^0$ and $\boldsymbol{w}^0$ with $\|\boldsymbol{u}^0 - \boldsymbol{w}^0\| < \varepsilon$. When we write both $\boldsymbol{u}^0$ and $\boldsymbol{w}^0$ in terms of the eigenvectors $\boldsymbol{v}_k$, this means that

$$\varepsilon^2 > \|\boldsymbol{u}^0 - \boldsymbol{w}^0\|^2 = \|\sum_{k=1}^{J-1} \alpha_k \boldsymbol{v}_k - \sum_{k=1}^{J-1} \beta_k \boldsymbol{v}_k\|^2 = \sum_{k=1}^{J-1} (\alpha_k - \beta_k)^2.$$

The last equality follows from the fact that the $\boldsymbol{v}_k$ are orthogonal and normalized. If $\boldsymbol{u}^n$ and $\boldsymbol{w}^n$ are the solutions obtained with the implicit numerical scheme, then we find

$$\|\boldsymbol{u}^n - \boldsymbol{w}^n\|^2 = \|\sum_{k=1}^{J-1} (\alpha_k - \beta_k)\lambda_k^n \boldsymbol{v}_k\|^2 = \sum_{k=1}^{J-1} (\alpha_k - \beta_k)^2 |\lambda_k|^{2n} < \varepsilon^2,$$

where in the last step we have used that all the λ_k are in norm smaller than 1. We conclude that small errors in the initial conditions do *not* become larger as we take many steps of the scheme; small difference of initial conditions leads to small difference in the solutions at any time in the future. This property is called *stability* of the scheme. We will see it again in the next subsection below.

4.4. **The Lax Equivalence Theorem.** We have seen for the heat equation that although the forward difference scheme is consistent, it is not necessarily convergent. This is different for the backwards scheme, which is both consistent and convergent no matter what parameter μ we used. We now generalize this to arbitrary linear PDE and state and prove one of the fundamental theorems of numerics of PDE, the Lax Equivalence Theorem. We consider a domain $D \subset \mathbb{R}^d$, and the linear PDE

$$\begin{cases} \partial_t u(x,t) = Lu(x,t), & (x \in \mathbb{R}^d, t \in (0,T]), \\ u(x,0) = u_0(x) & \text{(initial condition)}, \\ u(x,t) = u_{\mathrm{b}}(x) \quad \text{for } x \in \partial D, & \text{(boundary condition)}. \end{cases} \tag{4.11}$$

Above L is a differential operator, such as the F that we have seen in (4.5). However, we also demand that L is linear, i.e. that only the partial derivatives $\partial_x^n u$ appear in L (possibly with prefactors that depend on x), but no squares (or higher powers, or any nonlinear functions) of them appear, and the same u itself. This is e.g. not the case with the HJB equation. Furthermore, we demand that L does not explicitly depend on t.
We have little hope of finding a well-behaved numerical scheme if the equation itself is not well-behaved. What exactly constitutes a well-behaved equation is the content of the following definition.

Definition: The PDE (4.11) is *well-posed* if
(i): for all bounded initial conditions u_0 a solution exists.
(ii): There exists a constant $C > 0$ such that for any two bounded initial conditions u_0 and $\tilde{u}_0$, we have

$$|u(x,t) - \tilde{u}(x,t)| \leqslant C|u_0(x) - \tilde{u}_0(x)|,$$

for all $x \in D$, and all $t \in [0,T]$. Here $\tilde{u}$ is the solution of the PDE with initial condition $\tilde{u}_0$.
The condition (ii) is called *continuous dependence on the data*. It is very important for predicting solutions in situations where the initial data may be only approximately known (i.e. almost all situations arising in practice). However, there are many PDE that do not have this property, and lead to so-called chaotic behaviour.
Let us now consider a numerical scheme for (4.11). We will not treat the most general case, see the book by Morton and Mayers, Chapter 5, for more generality. Our procedure is:
1) Discretize the time in steps of size h. Put $t_n = hn$.
2) Discretize the space with a grid of points that are h_x apart, i.e. $|x_j - x_l| = h_x$ for two neighbouring grid points. On the boundary, we may have to introduce additional points and may then have $|x_j - x_l| < h_x$ if one of the two points is on the boundary.
3) Write u_j^n for the approximate solution, i.e. u_j^n is supposed to approximate $u(x_j, t_n)$.

We only study schemes of the form

$$u_j^{n+1} = u_j^n + \sum_{i=1}^{M} B_{ij} u_i^n + F_j, \tag{4.12}$$

where $B = (B_{ij})_{1 \leqslant i,j \leqslant M}$ is a $M \times M$ matrix that approximates L, F_j may come from an inhomogeniety or from boundary conditions, and M is the number of all spatial grid points. In vector notation, we have

$$\boldsymbol{u}^{n+1} = \boldsymbol{u}^n + B\boldsymbol{u}^n + \boldsymbol{F}. \tag{4.13}$$

Note that both the explicit and the implicit finite difference schemes for the heat equation are of this form. In the latter, B already involves the inverse of the discrete Laplacian.
To be precise, we need another notion.
Definition: A *refinement path* is a map $h \mapsto h_x(h)$ such that $\lim_{h\to 0} h_x(h) = 0$. In words, it is a way of making both time steps and spatial grid points get closer and closer together in some sort of coupled way.
We will henceforth assume that some refinement path is given and not talk about it much more. For example, a refinement path is implicitly present in the following definition.
Definition: The scheme (4.13) is *consistent* if for all $n \in \mathbb{N}$ with $hn \leqslant T$, and all $j \leqslant M$ we have

$$T_j^n = \frac{1}{h}\left(u(x_j, t_{n+1}) - u(x_j, t_n) - \sum_{i=1}^{M} B_{ij} u(x_i, t_n) + F_j \right) \to 0 \tag{4.14}$$

as $h \to 0$, uniformly in j and n such that the points x_j and t_n lie in the space-time domain.

Note that the matrix B will contain h_x in some way, and h and h_x are coupled through a refinement path. The idea of the above definition is again that $\frac{1}{h}(\boldsymbol{u}^{n+1} = \boldsymbol{u}^n) \approx \partial_t u$, and that $\frac{1}{h}(B\boldsymbol{u}^n - \boldsymbol{F}) \approx Lu$, and so the equations converge to each other. $\boldsymbol{T}$ is called the *truncation error*.

As we have seen, consistency alone is not enough for convergence. We need stability, which is the numerical equivalent for well-posedness.

Definition: The scheme (4.13) is *stable* if there exists $K > 0$ such that for all $h > 0$, all $n \in \mathbb{N}$ with $hn \leqslant T$, and all bounded numerical initial conditions $\boldsymbol{u}^0, \boldsymbol{w}^0$, we have

$$|u_j^n - w_j^n| \leqslant K|u_j^0 - w_j^0|,$$

for all $j \leqslant M$. (Note that the number of spatial grid points M will grow when h gets smaller - this is where the refinement path is hidden in this definition). Of course, there $\boldsymbol{u}^n$ is the numerical solution with initial condition $\boldsymbol{u}^0$, and $\boldsymbol{w}^n$ is the numerical solution with initial condition $\boldsymbol{w}^0$.

Let us now define what it means for a numerical scheme to be convergent:

Definition: The scheme (4.13) is *convergent* if for all x, t in the space-time domain, and all x_j, t_n such that $x_t \to x$ and $t_n \to t$ as $h \to 0$, we have $|u(x_j, t_n) - u_j^n| \to 0$ as $h \to 0$. Here u_j^n is the solution of the numerical scheme, and $u(x_j, t_n)$ is the true solution evaluated at x_j and t_n.

The main result now is:

Therorem (Lax Equivalence Theorem): Assume that (4.11) is linear and well-posed. Assume that (4.13) is consistent. Then (4.13) is convergent if and only if it is stable.

Proof. We only prove the direction that stability implies convergence. This direction is more important in practice, and the proof of the other direction requires tools from functional analysis that we do not have. We calculate

$$\begin{aligned}
&|u_j^{n+1} - u(x_j, t_{n+1})| \\
&= |u_j^n + \sum_{i=1}^{M} B_{ij} u_i^n + F_j - u(x_j, t_n) - \sum_{i=1}^{M} B_{ij} u(x_i, t_n) - F_j - hT_j^n|.
\end{aligned}$$

In matrix notation this means

$$\|\boldsymbol{u}^{n+1} - \boldsymbol{u}_{\text{true}}^{n+1}\| = \|(1 + B)(\boldsymbol{u}^n - \boldsymbol{u}_{\text{true}}^n) - h\boldsymbol{T}^n\| = (*).$$

Here, we defined $\boldsymbol{u}_{\text{true}}^n = (u(x_1, t_n), \dots u(x_M, t_n))$. We further have

$$\begin{aligned}
(*) &= (1 + B)^2(\boldsymbol{u}^{n-1} - \boldsymbol{u}_{\text{true}}^{n-1}) - h(1 + B)\boldsymbol{T}^n - h\boldsymbol{T}^{n-1} = \dots = \\
&= (1 + B)^n(\boldsymbol{u}^0 - \boldsymbol{u}_{\text{true}}^0) + h\sum_{k=1}^{n}(1 + B)^{n-k}\boldsymbol{T}^k = (**).
\end{aligned}$$

Now by stability, $\|(1-B)^{n-k}\| \leqslant K$ for all $n-k$ (otherwise we could find a vector such that $\|(1 - B)^{n-k}\boldsymbol{u}_0\| \geqslant K\|\boldsymbol{u}_0\|$. But this would mean that the difference

between the numerical solution with zero initial condition and the one with initial condition $\boldsymbol{u}_0$ is greater than $K\|\boldsymbol{u}_0\|$, which contradicts stability.) So,

$$(**) \leqslant hK \sum_{k=1}^{n} \boldsymbol{T}^n \leqslant hn \sup_k \|\boldsymbol{T}^k\|.$$

Now $hn \leqslant T$, and the $\sup_k \|\boldsymbol{T}^k\| \to 0$ as $h \to 0$ by consistency. Thus we have shown convergence. $\square$