PDE FOR FINANCE LECTURE NOTES (SPRING 2012)

VOLKER BETZ

1. PDE occurring in Finance

The aim of this section is to show how partial differential equations (PDE) occur naturally when considering problems in finance.

1.1. **Starting point.**

Let y_t be the state (e.g. the price) of a system (e.g. a stock) at time t. We assume that y_t is modelled by the stochastic differential equation (SDE)

$$\mathrm{d}y_t = F(y_t, t)\,\mathrm{d}t + G(y_t, t)\,\mathrm{d}\mathcal{W}_t. \tag{1.1}$$

Here $\mathcal{W}_t$ is one-dimensional standard Brownian motion. An example is geometric Brownian motion, where $F(x,t) = \mu x$ and $G(x,t) = \sigma x$, with $\sigma, \mu > 0$. We obtain

$$\mathrm{d}y_t = \mu y_t\,\mathrm{d}t + \sigma y_t\,\mathrm{d}\mathcal{W}_t.$$

This is the simplest model for the time evolution of a stock price.
Now let $\Phi : \mathbb{R} \to \mathbb{R}^n$ be a *payoff function*, which at the moment can be any function. For stock prices, which are positive, we want $\Phi : \mathbb{R}^+ \to \mathbb{R}$, but we can think of other assets that can take negative values. For example, with $\Phi(x) = x$ the payoff would be just the price of the stock, while for $\Phi(x) = (x - C)^+ \equiv \max\{0, x - C\}$ the payoff would be that of a European option with strike price C. Let us assume that at some time $t > 0$, the state of the asset is x, so this means $y_t = x$.
Question: What is the *expected payout*

$$u(x,t) = \mathbb{E}_{y_t = x}(\Phi(y_T)) \tag{1.2}$$

at a final time T?
Answer: $u(x,t)$ is the solution to the PDE

$$\partial_t u(x,t) + F(x,t)\partial_x u(x,t) + \frac{1}{2}G(x,t)^2\partial_x^2 u(x,t) = 0, \tag{1.3}$$

with *final value condition* $u(x,T) = \Phi(x)$.
(1.3) already shows the basic features of a PDE: It poses the problem to find a function where the partial derivatives balance in a certain way, at each point (x,t), and which fulfils some condition on the boundary (here: final time) of the domain. Usually, provided sufficiently many boundary conditions are given, there is at most one function that satisfies a PDE. This property of uniqueness needs to be proved in many cases, however.

Let us derive (1.3). For this, we rewrite (1.2) as

$$0 = u(x,t) - \mathbb{E}_{y_t=x}(\Phi(y_T)).$$

We want to plug the path y_t of the asset price into u. Notice that $u(x,t) = \mathbb{E}_{y_t=x}(u(y_t,t))$, and $\mathbb{E}_{y_t=x}(\Phi(y_T)) = \mathbb{E}_{y_t=x}(u(y_T,t))$. The last equality follows from $u(x,T) = \mathbb{E}_{y_T=x}(\Phi(y_T)) = \Phi(x)$. We obtain

$$0 = \mathbb{E}_{y_t=x}(u(y_T,T) - u(y_t,t)) = \mathbb{E}_{y_t=x}\Big(\int_t^T \mathrm{d}u(y_s,s)\Big). \tag{1.4}$$

If the function $s \mapsto u(y_s,s)$ would be differentiable, this would follow just from the fundamental theorem of calculus, and the meaning of $\mathrm{d}u(y_s,s)$ would be $\frac{d}{ds}u(y_s,s)\,\mathrm{d}s$. Since the non-smooth path y_s is plugged into u, we however need Itô calculus. We find

$$\begin{aligned}\mathrm{d}u(y_s,s) =& \Big(F(y_s,s)\partial_y u(y_s,s) + \frac{1}{2}G(y_s,s)^2\partial_y^2 u(y_s,s) + \partial_t u(y_s,s)\Big)\,\mathrm{d}s\\ &+ G(y_s,s)\partial_y u(y_s,s)\,\mathrm{d}\mathcal{W}_s.\end{aligned}$$

Since $\mathbb{E}\Big(\int_t^T G(y_s,s)\,\mathrm{d}\mathcal{W}_s\Big) = 0$ for every adapted process y_s (in particular our process is adapted!), it follows that in order to fulfil (1.4), it is enough that the first line of the equation above vanishes after taking the integral over time and the expectation. But the first line is just the PDE (1.3), so if u fulfils that, then also (1.4) holds. It is not difficult to see that on the other hand, if we want (1.4) hold for all t, then u needs to fulfil the PDE.

1.2. **Vector valued diffusions.** We now study several assets $\boldsymbol{y}_s = (y_s^{(1)}, \dots, y_s^{(n)})$, that solve the system of SDE

$$\mathrm{d}y_s^{(i)} = F_i(\boldsymbol{y}_s, s) + \sum_{j=1}^n G_{ij}(\boldsymbol{y}, s)\,\mathrm{d}\mathcal{W}_s^{(j)},$$

where the $\mathcal{W}_s^{(j)}$ are independent Brownian motions. Let $\Phi(\mathbb{R}^n \to \mathbb{R})$ be again a payoff function. Then

$$u(\boldsymbol{x}, t) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(\Phi(\boldsymbol{y}_T))$$

is the solution of the PDE

$$\partial_s u(\boldsymbol{x}, s) + \mathscr{L}u(\boldsymbol{x}, s) = 0 \qquad \text{for } t < s < T,$$

with final condition $u(\boldsymbol{x}, T) = \Phi(\boldsymbol{x})$, and where

$$\mathscr{L} = \sum_{i=1}^n F_i \partial_{x_i} + \frac{1}{2}\sum_{i,j,k=1}^n G_{ik}G_{kj}\partial_{x_i}\partial_{x_j} \tag{1.5}$$

is called the *generator* of the diffusion $\boldsymbol{y}_t$. The derivation is entirely parallel to the one above and uses the multi-dimensional Itô formula.

1.3. **Discounting and the Black-Scholes PDE.** We now allow for some discounting. We study

$$u(x,t) = \mathbb{E}_{y_t=x}\Big(\,\mathrm{e}^{-\int_t^T b(y_s,s)\,\mathrm{d}s}\,\Phi(y_T)\Big), \tag{1.6}$$

where $b : \mathbb{R}^2 \to \mathbb{R}$ is a discounting function, and $\Phi : \mathbb{R} \to \mathbb{R}$ is again a payoff function.

We claim that if y_t solves the SDE (1.1), then u as given above in (1.6) is the solution of the PDE

$$\partial_t u + F\partial_x u + \frac{1}{2}G^2\partial_x^2 u - bu = 0, \tag{1.7}$$

with final condition $u(x,T) = \Phi(x)$. Note that although we have not written the arguments of the functions u, F, G and b above, they are still functions and not numbers. We will use this shorter notation often.

To justify our claim, let us again first note that (1.6) is equivalent to $u(y,T) = \Phi(y)$ and

$$0 = \mathbb{E}_{y_t=x}\Big(u(y_T,T)\,\mathrm{e}^{-\int_t^T b(y_s,s)\,\mathrm{d}s} - u(y_t,t)\Big),$$

with the same justification as above. Writing $v((y_r)_{r\,\leqslant\,s}) = \mathrm{e}^{-\int_t^r b(y_r,r)\,\mathrm{d}r}$, (which depends now on more than one point in time, but is still adapted), we can transform this into

$$\begin{aligned} 0 &= \mathbb{E}_{y_t=x}\Big(u(y_T,T)v((y_r)_{r\,\leqslant\,T}) - u(y_t,t)v((y_r)_{r\,\leqslant\,T})\Big) \\ &= \mathbb{E}_{y_t=x}\Big(\int_t^T \mathrm{d}\,(u(y_s,s)v((y_r)_{r\,\leqslant\,s}))\Big). \end{aligned}$$

Now we have to apply the Itô formula as above. This is left as an exercise.

1.4. **The connection with the Black-Scholes PDE (BSPDE).** The PDE (1.7) is formally equal to the BSPDE. If we specialize (1.1) to geometric Brownian motion, then

$$\mathrm{d}y_t = \mu y_t\,\mathrm{d}t + \sigma y_t\,\mathrm{d}\mathcal{W}_t.$$

We choose a constant discounting function b (interest rate), and then (1.7) becomes

$$\partial_t u + \mu x\partial_x u + \frac{1}{2}\sigma^2 x^2\partial_x^2 u - bu = 0. \tag{1.8}$$

In comparison, the classical BSPDE would read

$$\partial_t u + \frac{1}{2}\sigma^2 x^2\partial_x^2 u + b(x\partial_x u - u) = 0. \tag{1.9}$$

The final conditions for a European call option would be $\Phi(x) = (x - C)^+$. The equations (1.8) and (1.9) are of the same structure, but the coefficients differ unless $b = \mu$. This is connected with the fact that unless we are in a risk-neutral world, the naive option pricing formula that just tries to take the expected discounted payout

$$u(x,t) = \mathbb{E}_{y_t=x}(\,\mathrm{e}^{-b(T-t)}\,\Phi(y_T))$$

as the option price gives the wrong result.

1.5. **Derivation of the Black-Scholes PDE.** We work here again with constant interest rate, which we now will call r instead of b. The no-arbitrage principle means that the price $P(t,x)$ of an option must be given by a self-financing trading strategy, for if it were not, a trader using this strategy would have an arbitrage opportunity. The strategy is determined by the amount a_t of stock and the amount b_t of risk-less bound that the trader holds at time t. It has to replicate the payout $\Phi(y_T)$ at maturity, and since at maturity the payout is equal to the option price for any pricing strategy, we have

$$a_T y_t + b_T \,\mathrm{e}^{rT} \, = P(y_T). \tag{1.10}$$

Since the strategy is self-financing, the only way the portfolio can change in value is that the stock goes either up or down, and by the growing capital in the bonds. Thus

$$\mathrm{d}(a_t y_t + b_t \,\mathrm{e}^{rt}\,) = a_t \,\mathrm{d}y_t + r b_t \,\mathrm{e}^{rt}\, \mathrm{d}t. \tag{1.11}$$

Our aim is to find the equation for $P(t,x)$ so that

$$P(t, y_t) = a_t y_t + b_t \,\mathrm{e}^{rt}\,, \tag{1.12}$$

i.e. the option price is exactly given by the value of the trading strategy portfolio for all times. Differentiating this both sides of the last equation, using the Itô formula and the SDE for $\mathrm{d}y_t$ gives for the left hand side:

$$\mathrm{d}P(t, y_t) = \partial_t P \,\mathrm{d}t + \partial_x P(F \,\mathrm{d}t + G\mathrm{d}\mathcal{W}_t) + \frac{1}{2} G^2 \partial_x^2 P \,\mathrm{d}t,$$

and for the right hand side (using (1.11) instead of the Itô formula):

$$\mathrm{d}(a_t y_t + b_t \,\mathrm{e}^{rt}\,) = a_t \,\mathrm{d}(F \,\mathrm{d}t + G\mathrm{d}\mathcal{W}_t) + r b_t \,\mathrm{e}^{rt}\, \mathrm{d}t.$$

Equating the $d\mathcal{W}_t$ terms means that we must have

$$a_t(y_t) = \partial_x P(t, y_t),$$

while equating the $\mathrm{d}t$ terms means that

$$\partial_t P + \frac{1}{2} G^2 \partial_x^2 P - r b_t \,\mathrm{e}^{rt}\, = 0$$

By (1.12), $b_t = (P(y_t, t) - a_t y_t) \,\mathrm{e}^{-rt}\, = (P(y_t, t) - \partial_x P(y_t, t) y_t) \,\mathrm{e}^{-rt}\,$, and plugging this into the last equation finally means that P must satisfy

$$\partial_t P + \frac{1}{2} G^2 \partial_x^2 P + r(x \partial_x P - P) = 0,$$

with final condition $P(x, T) = \Phi(x)$. This is the BSPDE (for general F, not only for geometric Brownian motion!). Let us compare this with the expected discounted payoff function given by (1.7), which we write slightly differently as

$$\partial_t u + \frac{1}{2} G^2 \partial_x^2 P + r \Big(\frac{1}{r} F \partial_x u - u \Big) = 0.$$

We see that the naive approach will only work when $F(x) = rx$, and not for any other F. However, this situation can be forced to happen by using the so-called Girsanov transformation. This topic is beyond the scope of the current course.

1.6. **Boundary value problems.**

Let now $D \subset \mathbb{R}^n$ be a subset of the space of possible asset values. We study the SDE

$$\mathrm{d}y_s^{(i)} = F_i(\boldsymbol{y}_s, s)\,\mathrm{d}s + \sum_{j=1}^{n} G_{ij}(\boldsymbol{y}_s, s)\,\mathrm{d}\mathcal{W}_s^{(j)}$$

whenever $\boldsymbol{y}_s \in D$, with starting value $\boldsymbol{y}_t = \boldsymbol{x} \in D$. Let $\tau(\boldsymbol{x})$ be the first time that $\boldsymbol{y}_s$ exits D, or $\tau(\boldsymbol{x}) = T$ if $\boldsymbol{y}_s$ does not leave D before T. (note that $\tau(\boldsymbol{x})$ is random!). Let

$$u(x,t) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(\Phi(\boldsymbol{y}_{\tau(\boldsymbol{x})}, \tau(\boldsymbol{x}))). \tag{1.13}$$

An example where this is useful is a knock-out option. In this case, D can be an interval $[a,b] \subset \mathbb{R}$, or a half-interval $[a,\infty)$. The option pays nothing (knock-out) if the stock price leaves D before maturity T, and pays $\Phi(\boldsymbol{y}_T, T)$ otherwise. This can be brought into the form above by having $\Phi(\boldsymbol{y}, s) = 0$ whenever $s < T$. The function u then solves the *boundary value problem*

$$\begin{aligned} \partial_t u(\boldsymbol{x}, s) + \mathscr{L}u(\boldsymbol{x}, s) &= 0 \qquad \text{for } \boldsymbol{x} \in D, \\ u(\boldsymbol{x}, t) &= \Phi(\boldsymbol{x}, t) \qquad \text{for } \boldsymbol{x} \in \partial D. \end{aligned} \tag{1.14}$$

Here, $\mathscr{L}$ is given by (1.5).

The derivation of (1.14) is similar to what we had before. (1.13) is equivalent to

$$0 = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(u(\boldsymbol{y}_{\tau(\boldsymbol{x})}, \tau(\boldsymbol{x})) - u(\boldsymbol{y}_t, t)\Big) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^{\tau(\boldsymbol{x})} \mathrm{d}u(\boldsymbol{y}_s, s)\Big).$$

with the condition $u(\boldsymbol{x}, s) = \Phi(\boldsymbol{x}, s)$ whenever $\boldsymbol{x} \in \partial D$. For $s < \tau(\boldsymbol{x})$, we compute $\mathrm{d}u(\boldsymbol{y}_s, s)$ as before using the Itô formula and obtain the PDE as we did several times already. The difference is only the boundary condition, which transfers into the PDE. There is one important caveat, which we have hidden a bit: namely, for the stochastic integrals

$$\mathbb{E}\Big(\int_t^{\tau(\boldsymbol{x})} G_{ij}(\boldsymbol{y}_s, s)\,\mathrm{d}\mathcal{W}_s^{(j)}\Big)$$

to be equal to zero, we need that $\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(\tau(\boldsymbol{x})) < \infty$. This is automatic here since we assumed $\tau(\boldsymbol{x}) \leqslant T$, but if we would study an infinite time horizon, this would cause problems, and we would have to be very careful. For more information search the internet with the keyword 'gamblers ruin'.

2. Some linear PDE theory

Let us review what the basic problem that we are treating in PDE theory is: we are looking for a function f, defined on an open subset $U \subset \mathbb{R}^n$, such that at each point $x \in U$, a certain combination of partial derivatives and values of the function itself gives a certain value. On the boundary ∂U, we may or may not prescribe values for the function or its derivatives. A (random) example would be

$$D = B_1 := \{(x,y) \in \mathbb{R}^2 : x^2 + y^2 \leqslant 1\},$$

and

$$x\partial_x^2 f(x,y) + y\partial_y^2 f(x,y) = f(x,y)^2 \quad \text{on } D, \qquad f(x,y) = xy \quad \text{on } \partial D.$$

PDE do not always have a solution, and the solution may not always be unique. Even among those that do have a unique solution, there are precious few where it is possible to find the solution in the form of a closed formula. In this section, we will study some of these few cases.

2.1. **First order PDE.** These PDE contain only first order derivatives. Recall the notation

$$\nabla f(x_1, \dots x_n) = (\partial_{x_1} f(x_1, \dots, x_n), \dots, \partial_{x_n} f(x_1, \dots, x_n))$$

for the gradient of a function, and the notation

$$\boldsymbol{b} \cdot \nabla f = b_1 \partial_{x_1} f + \ldots + b_n \partial_{x_n} f$$

for the scalar product with a vector $\boldsymbol{b} = (b_1, \dots, b_n)$.

The transport equation. This is the simplest type of PDE there is.
Definition: Consider $D \subset \mathbb{R}^n \times \mathbb{R}$. We say that $u : D \to \mathbb{R}$ solves a *transport equation* (with constant coefficients) if

$$\partial_t u + \boldsymbol{b} \cdot \nabla_{\boldsymbol{x}} u = 0. \tag{2.1}$$

Above, we have $u = u(\boldsymbol{x}, t) = u(x_1, \dots, x_n, t)$.
Can we solve this equation for u? Yes, and it is easy. The key is to see that although we have separated 'time' t and 'space' $\boldsymbol{x}$, what we really have is a gradient of a function of $n+1$ variables being perpendicular to a certain vector. More explicitly, define $\boldsymbol{c} = (b_1, \dots, b_n, 1)$. Then (2.1) becomes

$$\boldsymbol{c} \cdot \nabla_{(\boldsymbol{x},t)} u = 0.$$

So, u is constant in the direction $(b_1, \dots, b_n, 1)$. In other words,

$$u(\boldsymbol{x} + s\boldsymbol{b}, t + s) = u(\boldsymbol{x}, t) \tag{2.2}$$

for all $\boldsymbol{x} \in \mathbb{R}^n$ and all $s, t \in \mathbb{R}$. Whenever we know u on one point of such a line, we know it on the whole line. To know it at one point, we need the boundary condition.
Even in this simple example, we can see clearly that both existence and uniqueness can fail easily. We have established that the solution is constant on straight lines. Now, if D is e.g. a ball, then each straight line will intersect its boundary twice. So, when prescribing values for u on ∂D, we have to be very careful in this case, otherwise we will prescribe two different values on the same straight line, and the PDE has no solution. On the other hand, we can also lose uniqueness. Before we see this, let us look at a case where we do have existence and uniqueness. This is the full space initial value problem, where (2.1) is on $D = \mathbb{R}^n \times (0, \infty)$, and $u(\boldsymbol{x}, 0) = h(\boldsymbol{x})$. This is natural, as we claim that we know u to be h at time $t = 0$ and want to see how it evolves in time. Now (2.2) gives us

$$u(\boldsymbol{x} + s\boldsymbol{b}, s) = u(\boldsymbol{x}, 0) = h(\boldsymbol{x}),$$

and so by putting $\boldsymbol{y} = \boldsymbol{x} + s\boldsymbol{b}$ we find

$$u(\boldsymbol{y}, s) = h(\boldsymbol{y} - s\boldsymbol{b}).$$

Note that the values of h are indeed 'transported' along the straight lines. Non-uniqueness now occurs if we look at more general (maybe less natural) boundary conditions. We could e.g. study (2.1) on the half space $D = \{(\boldsymbol{x}, t) : t + \boldsymbol{a} \cdot \boldsymbol{x} \geqslant 0\}$. This will still give a unique solution, except when $(\boldsymbol{b}, 1)$ is parallel to the boundary of D. Then, we will have no solution unless the boundary condition h is a constant, and infinitely many solutions if h is indeed constant, as the values along all other straight lines will not have been prescribed.
To summarize, in the very simple case of the transport equation we have found the solution by finding a coordinate direction (namely $(\boldsymbol{b}, 1)$) in which the solution is constant. Let us try this strategy for more complicated first order PDE.

Linear first order PDE. These are of the same shape as the transport equation, but now the coefficients $\boldsymbol{b}$ may depend on $\boldsymbol{x}$ and t, and we also allow a term proportional to u to appear.
Definition: Consider $D \subset \mathbb{R} \times \mathbb{R}^n$. We say that $u : D \to \mathbb{R}$ solves a *linear first order PDE* if

$$\boldsymbol{b}(\boldsymbol{x}, t) \cdot \nabla_{\boldsymbol{x},t} u(\boldsymbol{x}, t) + c(\boldsymbol{x}, t) u(\boldsymbol{x}, t) = 0 \tag{2.3}$$

on D.
Like in the transport equation, we look for a coordinate direction in which the solution is easy. Here, $\boldsymbol{b}(\boldsymbol{x}, t)$ seems promising. But note that now the 'easy' direction depends on where we are in $(\boldsymbol{x}, t)$-space! This means that we are dealing with an 'easy curve' instead of an 'easy straight line'. More precisely: Let $\gamma : \mathbb{R} \to \mathbb{R}^{n+1}$ be a curve in coordinate space. We write $\gamma(s) = (\boldsymbol{x}(s), t(s))$. Assume that γ is parallel to the vector $\boldsymbol{b}(\boldsymbol{x}, t)$ at every point $(\boldsymbol{x}, t)$ through which γ passes. In symbols, assume that

$$\dot{\gamma}(t) \equiv \frac{\mathrm{d}}{\mathrm{d}t}\gamma(t) = \boldsymbol{b}(\gamma(s)) \equiv \boldsymbol{b}(\boldsymbol{x}(x), t(s)). \tag{2.4}$$

The $\equiv$ signs mean (as always) that the two terms are the same by definition, i.e. there are just two different ways of writing them.
Let us now define $v(s) = u(\gamma(s)) \equiv u(\boldsymbol{x}(s), t(s))$. So, v is just what you get when you evaluate u along γ. The interesting point is that if u solves (1.3), then by the chain rule,

$$\frac{\mathrm{d}}{\mathrm{d}s} v(s) = \boldsymbol{b}(\gamma(s)) \cdot \nabla_{\boldsymbol{x},t} u(\gamma(s)) = -c(\gamma(s)) u(\gamma(s)) = -c(\gamma(s)) v(s).$$

This is now an ordinary differential equation for v, which is easily solved: you can check that

$$v(s) = v(0)\, \mathrm{e}^{-\int_0^s c(\gamma(r))\, \mathrm{d}r}. \tag{2.5}$$

Even though it might not seem so, we have in some sense already solved (2.1). Namely, if we need to know the solution of (2.1) at a point $(\boldsymbol{x}, t)$ in a domain D, we start a curve γ at $(\boldsymbol{x}, t)$ and make sure that it fulfils (2.4). This means we have to solve the corresponding system of ordinary differential equations, which may or may not pose a problem in itself. Assuming that this goes well, however, we then follow γ until it hits the boundary. We take the prescribed value at the boundary as $v(0)$, re-parametrize γ so that $\gamma(0)$ is on the boundary, and apply (2.5) to get the value of u along the full curve γ. In particular, since γ contains $(\boldsymbol{x}, t)$, we get the value at that point.

Of course, this is not a closed form solution. To get the latter, we need to be able to solve the ODE defining γ, and to invert the coordinates to get from the 'curve coordinates' $\gamma(s)$ to the space coordinates $(\boldsymbol{x}, t)$. This is in general hard (and has nothing whatsoever to do with PDE theory), but sometimes it can be done. Here is an example.

Example: $D = \{(x,t) : x > 0, t > 0\} \subset \mathbb{R}^2$. The initial condition is $u(x, 0) = g(x)$ for some function g, and the PDE is

$$x\partial_x u(x,t) - t\partial_x u(x,t) = u(x,t).$$

So in the framework above, we have $\boldsymbol{b}(x,t) = (-t, x)$. We seek γ with $\dot{\gamma}(s) = b(\gamma(s))$. Writing $\gamma(s) = (\gamma_1(s), \gamma_2(s))$, this means we have $\boldsymbol{b}(\gamma(s)) = (-\gamma_2(s), \gamma_1(s))$, so we want γ to fulfil

$$\dot{\gamma}_1(s) = -\gamma_2(s), \quad \dot{\gamma}_2(s) = \gamma_1(s).$$

The solution is given by

$$\gamma_1(s) = c\cos(s), \quad \gamma_2(s) = c\sin(s),$$

where c can be arbitrary and determines the points through which γ runs: namely, γ describes circles of radius c. Since in the context of (2.3), we have $c(x,t) = -1$, equation (2.5) now becomes

$$v(s) = v(0)\,\mathrm{e}^{-\int_0^s 1\,\mathrm{d}s} = v(0)\,\mathrm{e}^{-s}$$

Now we put it all together and invert the coordinates. If we want (x,t) to lie on γ, it means we want $(x,t) = (c\cos(s), c\sin(s))$. Resolving this for c and s, we find that $c = \sqrt{x^2 + t^2}$ (by squaring both components above, adding and taking the square root after using $\sin^2 + \cos^2 = 1$). Also $s = \arctan(t/x)$ (by dividing the components, and taking the arctan). We arrive at

$$u(x,t) = v(\arctan(t/x)) = g(\sqrt{x^2 + t^2})\,\mathrm{e}^{\arctan(t/x)}.$$

Characteristic curves. The method of seeking 'simple directions' can be extended to even more difficult first order PDE. In these cases, often the ODE for γ and for the solution along γ become coupled, and one has to solve the whole system at one go. While it is not conceptually much more difficult than what we had, it is considerably more messy, and we will not pursue it further.

2.2. Laplace and Poisson equations.

Motivation: Exit times from a domain. Recall the knock-out option example from the previous section. Here we study a similar problem, but with positive payout at the boundary. Consider the SDE

$$\mathrm{d}y_s^{(i)} = \mathrm{d}\mathcal{W}_i(s) \qquad \text{for } \boldsymbol{y}_s \in D, \qquad \boldsymbol{y}_0 = \boldsymbol{x} \in D,$$

with bounded $D \subset \mathbb{R}^n$, and the utility function

$$u(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}_0=\boldsymbol{x}}(\Phi(\boldsymbol{y}_{\tau(\boldsymbol{x})})),$$

where τ is the first time that $\boldsymbol{y}_t$ hits ∂D. Then in the same way as may times before, we find that u is the solution to the PDE

$$\Delta u \equiv \sum_{i=1}^{n} \partial_x^2 u = 0 \text{ on } D, \qquad u(x) = \Phi(x) \text{ on } \partial D. \tag{2.6}$$

This is the *Laplace equation.* If we include the running payoff f, i.e. if

$$u(\boldsymbol{x}) = \mathbb{E}_{\boldsymbol{y}_0=\boldsymbol{x}}\Big(\Phi(\boldsymbol{y}_{\tau(\boldsymbol{x})}) + \int_0^{\tau(\boldsymbol{x})} f(\boldsymbol{y}_s)\,\mathrm{d}s\Big),$$

then we obtain instead (by the familiar argument) that u solves the *Poisson equation*

$$\Delta u(\boldsymbol{x}) + f(\boldsymbol{x}) = 0 \text{ on } D, \qquad u(\boldsymbol{x}) = \Phi(\boldsymbol{x}) \text{ on } \partial D. \tag{2.7}$$

Now, how can we solve equations (2.6) and (2.7)? Let us start with the seemingly strange case $D = \mathbb{R}^n \setminus \{0\}$. Then it is possible (with some experience) to guess a solution. It is given by

$$F(x) = \begin{cases} -\dfrac{1}{2\pi} \ln|x| & \text{for } n = 2, \\ \dfrac{1}{n(n-2)\alpha(n)} \dfrac{1}{|x|^{n-2}} & \text{for } n \geqslant 3. \end{cases} \tag{2.8}$$

Above, $\alpha(n)$ is the volume of the unit ball, and all the x-independent prefactors are normalisations that will be useful later on. The function F in (2.8) may seem very special, but it is of great importance and is called the *fundamental solution.* We will verify as an exercise that indeed $\Delta F = 0$ on D. As a quick challenge, try to think what the situation would be in $d = 1$: What solutions to the Laplace equation are there?

The key to solving (2.7) is now to observe that not only F, but also the function $x \mapsto F(x - y)$ solves (2.6). This leads to the following

Theorem 2.1. *Let F be the fundamental solution to the Laplace equation, and assume that f in* (2.7) *is nice enough, e.g. twice continuously differentiable with compact support. Put*

$$u(x) = -\int F(x-y) f(y)\,\mathrm{d}y$$

Then u solves (2.7).

The proof goes via careful integration by parts, where it is important to pay special attention to the region of space where F diverges. Here, it is also important to use the precise normalisation for F, otherwise the theorem would not hold. The proof is given e.g. in Evans.

Solving (2.6) and (2.7) on a domain $D \subset \mathbb{R}^n$ subject to boundary conditions is more difficult, but there is a general recipe that can be followed. Here it is:

Step 1: Instead of solving (2.6) for the given boundary function Φ, solve it (in the variable y) for the special boundary function $y \mapsto F(y-x)$ for all x. I.e., solve

$$\Delta \phi^x(y) = 0 \text{ on } D, \qquad \phi^x(y) = F(y-x) \text{ on } \partial D. \tag{2.9}$$

This seems not much easier than the original problem, but for some nice D it actually is. However, this is in general the hard step.

Step 2: Define

$$G(x,y) = F(y-x) - \phi^x(y) \qquad \text{for } x \in D, x \neq y.$$

G is called *Greens function* of the Laplace equation for the domain D.

Step 3: The Poisson equation is now solved by

$$u(x) = \int_{\partial D} \Phi(y) \frac{\partial G}{\partial \nu}(x,y) \, \mathrm{d}S(y) - \int_D f(y) G(x,y) \mathrm{d}y,$$

where ν is the vector of length one that is perpendicular to ∂D at the point y and points outward, and $\mathrm{d}S$ is the surface measure on ∂D. This theorem is again proved by careful integration by parts.

Example: Greens function on the half space

Consider the half space $\mathbb{R}^n_+ = \{x = (x_1, \ldots, x_n) : x_n > 0\}$. The task of step 1 above is to find the solution to (2.9) for all $x \in \mathbb{R}^n_+$. Note that $y \mapsto F(y-x)$ trivially fulfils the boundary condition part of (2.9), but not the PDE part since $\Delta_y F(y-x) \neq 0$ for $y = x$; indeed, the function is not even defined there. But F only depends on $|y-x|$, and if we could somehow force its singularity to lie outside of the half plane, we would be in business. These two things suggest that we may try $\phi^x(y) = F(y - \bar{x})$, where $\bar{x} = (x_1, \ldots, x_{n-1}, -x_n)$. This now solves (2.9). So, $G(x,y) = F(y-x) - F(y-\bar{x})$. The derivative $\frac{\partial G}{\partial \nu}$ is just the derivative in the direction of the n-th coordinate, so on ∂D,

$$\frac{\partial G}{\partial \nu}(x,y) = -\partial_{y_n} G(x,y) = -\frac{2x_n}{n\alpha(n)|x-y|^n}.$$

Here, we have used that on ∂D, $x_n = 0$. Then,

$$u(x) = \frac{2x_n}{n\alpha(n)} \int_{\{y \in \mathbb{R}^n : y_n = 0\}} \frac{\Phi(y)}{|x-y|^n} \, \mathrm{d}y.$$

solves the Laplace equation (2.6). The solution to the Poisson equation (2.7) follows from step 3 above in the same way.

2.3. **The heat equation.** A function $u(x,t)$ solves the *heat equation* if (with $\sigma > 0$)

$$\partial_t u - \frac{1}{2}\sigma^2 \Delta u = 0. \tag{2.10}$$

In the simplest case, (2.10) is supposed to hold for all $x \in \mathbb{R}^n$, and all $t > 0$, and there is an initial condition $u(x,0) = u_0(x)$. We have seen this equation already: if in (1.3) we put $F = 0$ and $G = \sigma^2$, then we obtain (2.10). In other words, the heat equation is the Kolmogorov backward equation for Brownian motion. It is one of the most important equations in physics, as it models heat flow (hence the name), diffusion of liquids, and many more things.

From Black-Scholes to heat: We will now show that the BSPDE can be transformed into a heat equation by a change of variables. Recall the BSPDE given in (1.9):

$$\partial_t P + \frac{1}{2}\sigma^2 x^2 \partial_x^2 P + b(x\partial_x P - P) = 0, \tag{2.11}$$

with final condition $P(x,T) = \Phi(x)$. To understand how anybody could guess the variable transform that we are going to use, note that in (2.11), x need to be positive, as it is a stock price; and, that we have a final condition at T. In contrast, in (2.10), we have $x \in \mathbb{R}$ and an initial condition. So the least we would have to do to connect the two is to invert time, and to map the nonnegative x into something on all of $\mathbb{R}$. The latter is just what the logarithm does, and an additional hint for uisng it would be that geometric BM behaves like the exponential of BM itself. After these explanations, the following transformation may seem a bit less arbitrary: we put

$$y = \ln x \text{ (so } x = \mathrm{e}^y\text{)}, \quad \text{and } \tau = \frac{1}{2}\sigma^2(T-t).$$

Then we put

$$v(y,\tau) = P(\mathrm{e}^y, T - \frac{2}{\sigma^2}\tau) \qquad (= P(x,t)).$$

Let us try whether v solves the heat equation:

$$\partial_\tau v(y,\tau) = -\frac{2}{\sigma^2}\partial_2 P(\mathrm{e}^y, T - \frac{2}{\sigma^2}\tau) = -\frac{2}{\sigma^2}\partial_t P(x,t),$$

$$\partial_y v(y,\tau) = \mathrm{e}^y\, \partial_1 P(\mathrm{e}^y, T - \frac{2}{\sigma^2}\tau) = x\partial_x P(x,t),$$

$$\partial_y^2 v(y,\tau) = (\mathrm{e}^y)^2 \partial_1^2 P(\mathrm{e}^y, T - \frac{2}{\sigma^2}\tau) = \mathrm{e}^y\, \partial_1 P(\mathrm{e}^y, T - \frac{2}{\sigma^2}\tau) = x^2\partial_x^2 P(x,t) - x\partial_x P(x,t).$$

Above, $\partial_1 P$ means the function that one gets from P by differentiating with respect to the first argument. Note that this is different from $\partial_y P(\mathrm{e}^y, \ldots)$ since this would invoke a chain rule, and also better than $\partial_x P(\mathrm{e}^y, \ldots)$, where the reader is asked to guess that the first argument is somehow connected to the letter x. This example shows that the inherited notation for derivatives is not satisfactory (it uses a dummy variable explicitly), but unfortunately it is very deeply entrenched in mathematics and there is no hope to overcome it.

Back to the calculation. We find

$$\begin{aligned}\partial_\tau v - \partial_y^2 v &= -\frac{2}{\sigma^2}\partial_t P - x^2\partial_x^2 P - x\partial_x P \\ &= -\frac{2}{\sigma^2}\left(\partial_t P + \frac{\sigma^2}{2}x^2\partial_x^2 P + \frac{\sigma^2}{2}x\partial_x P\right) \\ &= -\frac{2}{\sigma^2}\left(-b(x\partial_x P - P) + \frac{\sigma^2}{2}x\partial_x P\right) \\ &= -\frac{2}{\sigma^2}\big((-b+\sigma^2/2)\partial_y v + bv\big).\end{aligned}$$

So v solves

$$\partial_\tau v - \partial_y^2 v + (1-\frac{2b}{\sigma^2})\partial_y v + \frac{2b}{\sigma^2}v = 0. \tag{2.12}$$

This is not quite yet the heat equation. To proceed, let us put $k = 2b/\sigma^2$, and

$$u(y,\tau) = \mathrm{e}^{-\alpha y - \beta\tau}\, v(y,\tau),$$

thus $v(y,\tau) = \mathrm{e}^{\alpha y + \beta\tau}\, u(y,\tau)$. Then (2.12) becomes

$$(\beta u + \partial_\tau u) - (\alpha^2 u + 2\alpha\partial_y u + \partial_y^2 u) + (1-k)(\alpha u + \partial_y u) + ku = 0.$$

Then $\partial_y u$ terms vanish if $-2\alpha + (1-k) = 0$, and the u terms vanish if $\beta - \alpha^2 + (1-k)\alpha + k = 0$. This gives

$$\alpha = \frac{1-k}{2}, \qquad \beta = -\frac{(k+1)^2}{4}.$$

With this choice of α, β, the function u indeed solves the heat equation

$$\partial_\tau u - \partial_y^2 u = 0, \qquad u(y,0) = \mathrm{e}^{\frac{1}{2}(k-1)y}\,\Phi(\mathrm{e}^y). \tag{2.13}$$

So to solve the BS-PDE, we have to solve (2.13) to get u, then get v from u, and then undo the change of variables to find $P(x,t) = v(\ln x, \frac{1}{2}\sigma^2(T-t))$. This will work for any payoff-function Φ, provided we can solve the heat equation with the corresponding initial condition. This we can indeed do:

Solution for the whole-space heat equation:

The function

$$f(\boldsymbol{x},t) = \frac{1}{(2\pi\sigma^2 t)^{n/2}}\int_{\mathbb{R}^n} \mathrm{e}^{-\frac{|\boldsymbol{x}-\boldsymbol{y}|^2}{2\sigma^2 t}}\, f_0(\boldsymbol{y})\,\mathrm{d}\boldsymbol{y} \tag{2.14}$$

solves the heat equation (2.10) with initial condition $f(\boldsymbol{x},0) = f_0(\boldsymbol{x})$. The function

$$F(\boldsymbol{x},t) = \frac{1}{(2\pi\sigma^2 t)^{n/2}}\,\mathrm{e}^{-\frac{|\boldsymbol{x}|^2}{2\sigma^2 t}}$$

is called *fundamental solution* of the heat equation. It actually also solves the heat equation for $t > 0$, except when $\boldsymbol{x} = 0$.

Remarks: F is the transition density of Brownian motion, i.e. $\mathbb{P}_0(\mathcal{W}_t \in A) = \int_A F(\boldsymbol{x},t)\,\mathrm{d}\boldsymbol{x}$. This is no accident, but is related to the time reversibility of Brownian motion and the Kolmogorov equations. We will not discuss this further in the present lecture.

Also, we need to place some restrictions on the initial condition f_0 for the solution to make sense. In fact, f_0 need not be continuous, but it must not grow too fast at infinity. If $f_0(\boldsymbol{x}) \leqslant M\,\mathrm{e}^{c|\boldsymbol{x}|^2}$, then the solution (2.14) at least exists for finite time. If $f_0(\boldsymbol{x}) \leqslant M\,\mathrm{e}^{c|\boldsymbol{x}|^{2-\delta}}$, for some small $\delta > 0$, then the solution (2.14) exists for all times. Translating the latter condition back to the Black-Scholes coodinates gives

$$u(y,0) = \mathrm{e}^{-\frac{1}{2}(k-1)y}\,\Phi(\mathrm{e}^y) \leqslant M\,\mathrm{e}^{c|y|^{2-\delta}} \Leftrightarrow x^{-\frac{1}{2}(k-1)}\Phi(x) \leqslant M\,\mathrm{e}^{c(\ln x)^{2-\delta}}.$$

This works fine if $\Phi(x) < |x|^r$ for any $r > 0$, but will fail if Φ grows exponentially at infinity.

2.4. **Solution of the heat equation on a half space.** This solution will be useful for the pricing of barrier options. We want to solve

$$\begin{aligned} &\partial_t u = \frac{1}{2}\partial_x^2 u = 0 \quad \text{for } t > 0, x > 0. \\ &u(x,0) = g(0), \quad u(0,t) = \phi(t). \end{aligned} \tag{2.15}$$

It seems at first that the boundary data needs to fit together for this equation to make sense; more precisely, when $\lim_{t\to 0}\phi(t) \neq \lim_{x\to 0} g(x)$, then it seems that we want on the one hand a function that is twice differentiable in x and differentiable in t (for the PDE to make sense), but on the other hand is discontinuous at the boundary. We will however see that this is not a problem. To solve (2.15), let us split it into two easier problems.

Proposition: Assume that v solves

$$\partial_t v = \frac{1}{2}\partial_x^2 v, \qquad v(x,0) = g(x), v(0,t) = 0, \tag{2.16}$$

and that w solves

$$\partial_t w = \frac{1}{2}\partial_x^2 w, \qquad w(x,0) = 0, w(0,t) = \phi(t). \tag{2.17}$$

Then $u = v + w$ solves (2.15).

Proof. It is clear that u fulfils the boundary conditions, and that it solves the PDE follows from the fact that the derivatives can be dsitributed onto v and w, who individually solve the heat equation. □

It will turn out to be an advantage if in (2.17), we have $\lim_{t\to 0}\phi(t) = 0$; then the formula for the solution will be easier to make sense of. This can be easily achieved: we replace $\phi(t)$ with $\phi(t) - \phi(0)$ in (2.17), and $g(x)$ with $g(x) - \phi(0)$ in (2.16). To the solution $\tilde{u}$ that we obtain from this we only have to add $\phi(0)$, which then solves the original equation.

Solution of (2.16):

We use a reflection trick: Let us put

$$\tilde{g}(x) = \begin{cases} g(x) & \text{if } x > 0, \\ -g(-x) & \text{if } x < 0. \end{cases}$$

We now solve the whole space problem with initial condition $\tilde{g}$, using (2.14). The result is

$$v(x,t) = \frac{1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{|x-y|^2}{2t}}\, \tilde{g}(y)\,\mathrm{d}y = (*).$$

We now change integration variables from y to $-y$, with the result

$$(*) = \frac{1}{\sqrt{2\pi t}} \int_{\infty}^{-\infty} \mathrm{e}^{-\frac{|x+y|^2}{2t}}\, \tilde{g}(-y)\,\mathrm{d}y == \frac{-1}{\sqrt{2\pi t}} \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{|(-x)-y|^2}{2t}}\, \tilde{g}(y)\,\mathrm{d}y = -v(-x,t)$$

So, the solution has the same symmetry as the boundary condition for all times! In particular, $v(0,t) = -v(0,t)$, which only leaves the possibility $v(0,t) = 0$. Thus v restricted to $x > 0$ indeed solves (2.16).

The above solution can be written in terms of the fundamental solution $F(x,t) = \frac{1}{\sqrt{2\pi t}} \exp(-\frac{x^2}{2t})$. Namely, simple manipulations show that

$$v(x,t) = \int_0^{\infty} G(x,y,t)g(y)\,\mathrm{d}y, \qquad \text{with } G(x,y,t) = F(x-y,t) - F(x+y,t).$$

Note the striking similarity to the Greens function we found in the solution to (2.6). G is indeed the Greens function for the heat equation. This will become clear when we consider the

Solution to (2.17)**:**

As (2.17) is a bit like the boundary value problem (2.6), we can expect a similar solution formula, and there is indeed one: the function

$$w(x,t) = \int_0^t \partial_y G(x,y,t-s)|_{y=0}\phi(s)\,\mathrm{d}s \tag{2.18}$$

solves (2.17). The partial derivative of G can of course be computed, leading to

$$w(x,t) = \int_0^t \frac{x}{(t-s)\sqrt{2\pi(t-s)}}\, \mathrm{e}^{-\frac{x^2}{2(t-s)}}\, \phi(s)\,\mathrm{d}s.$$

The proof of this formula goes roughly as follows: Since F solves the heat equation, so does $(x,t) \mapsto G(x,y,t)$ for all y, and also $(x,t) \mapsto \partial_y G(x,y,t-s)$ for all s (just exchange the order of derivatives). But when two or more functions solve the heat equation, (or, any linear equation), then all weighted sums of these function solve the same equation (just distribute the derivatives), and this even applies to convergent sums of infinitely many terms, and even integrals. So w as given by (2.18) does solve the heat equation. For the boundary conditions: naively, $w(x,0) = 0$ as the range of the integration is zero. However, we have to approach this limit coming from positive t, which makes it less trivial. Likewise, the limit of w as $x \to 0$ needs to be studied carefully, and it needs to be shown that it converges to $\phi(t)$. This is beyond the scope of the present lecture and will not be done here.

Pricing a barrier option

A barrier option changes its value suddenly when the asset process $\boldsymbol{y}_t$ hits a predefined barrier. For example, a down-and-out call with barrier X will be worthless if the stock falls below X before maturity. Otherwise, it behaves like a normal call.

Of course, such options are very little different from gambling in a casino, and encourage massive market manipulation to temporarily suppress a stock price, and should not be legal. But this is not out concern here, we are trying to price them, assuming that no manipulation takes place. In that case, interestingly, the Black-Scholes PDE gives a fair price, so they are not fundamentally different from vanilla options. The procedure goes like this: We start with the BSPDE with boundary condition zero at asset value X. We do the variable transform to turn this into a heat equation with zero boundary condition at a suitably modified place. We then solve this heat equation using the theory above. Finally we transform back to the Black-Scholes coordinates. The result for a down-and-out call with barrier X is

$$V(x,t) = V_0(\ln x, \frac{1}{2}\sigma^2(T-t)) - \Big(\frac{x}{X}\Big)^{1-k} V_0(\ln\frac{X^2}{x}, \frac{1}{2}\sigma^2(T-t)),$$

with $k = 2r/\sigma^2$, and where V_0 is the value of a vanilla option with the same strike price. The details will be worked out on an exercise sheet. Just notice that the formula makes sense: when the stock price x is much larger than the barrier price X, the price is almost that of a vanilla option, while it is almost zero if the stock price x is close to X.

2.5. **The heat equation on an interval.** Recall that vanilla options correspond to the heat equation on the full space, while barrier options correspond to the heat equation on the half space. We will now study the heat equation on an interval, which corresponds to double barrier options. Surprisingly, this will turn out to be easier than the cases considered before. We will need the following fact:

Theorem: Let $f : [0,1] \to \mathbb{R}$ be a continuous function with $f(0) = f(1) = 0$. Then for all $x \in [0,1]$,

$$f(x) = \sum_{k=1}^{\infty} a_k \sin(k\pi x) \qquad \text{with} \qquad a_k = 2\int_0^1 f(x)\sin(k\pi x)\,\mathrm{d}x. \tag{2.19}$$

We will not prove this Theorem, but make some remarks:

(i): The theorem is even valid for discontinuous function, but then the sum does not converge for all $x \in [0,1]$. Instead, the (squared) area between the function f and the partial sums becomes arbitrarily small (convergence in L^2).

(ii): Equation (2.19) is called *Fourier series* (more precisely: Fourier-sine-series) of f, and the a_n are the *Fourier coefficients.* A more common variant of (2.19) is

$$f(x) = \sum_{k=0}^{\infty} \tilde{a}_k\, \mathrm{e}^{2\pi \mathrm{i} kx} \qquad \text{with} \qquad \tilde{a}_k = \int_0^1 f(x)\, \mathrm{e}^{-2\pi \mathrm{i} kx}\, \mathrm{d}x.$$

This does not need the condition that $f(0) = f(1) = 0$; however, if we do have that condition, (2.19) is simpler and we will stick with that.

(iii): The proof uses orthogonality of the set $\{\sin(k\pi x) : k \geqslant 1\}$ in the space of square integrable functions. More concretely, you can check that

$$2\int_0^1 \sin(k\pi x)^2\,\mathrm{d}x = 1, \qquad \text{and} \qquad \int_0^1 \sin(k\pi x)\sin(m\pi x)\,\mathrm{d}x = 0,$$

for $0 < k < m$. This already shows that (2.19) works for the sine functions themselves and eventually leads to the proof.
Let us now use the theorem to solve

$$\begin{aligned} &\partial_t u = \frac{1}{2}\partial_x^2 u \qquad \text{for } t > 0, 0 < x < 1, \\ &u(x,0) = g(x), \quad u(0,t) = \phi_0(t), \quad u(1,t) = \phi_1(t). \end{aligned} \tag{2.20}$$

Let us start with the easiest case $\phi_0 = \phi_1 = 0$. Then for each $t > 0$, we can express the solution $u(x,t)$ of (2.20) using (2.19). Thus,

$$u(x,t) = \sum_{k=1}^{\infty} a_k(t) \sin(k\pi x) \tag{2.21}$$

for some coefficients $a_k(t)$. By the initial condition, we must have

$$a_k(0) = 2 \int g(x) \sin(k\pi x) \, \mathrm{d}x,$$

for all k. By the differential equation, we must have

$$\sum_{k=1}^{\infty} \partial_t a_k(t) \sin(k\pi x) = \sum_{k=1}^{\infty} (-\frac{1}{2} k^2 \pi^2 \sin(k\pi x)) a(t).$$

Now the orthogonality of the sine functions that we mentioned above really means that we cannot build one of them out of a finite or infinite number of different ones. So, for the equality above to hold, the individual terms for each k have to agree, and we find $\partial_t a_k = -k^2\pi^2 a_k$. with solution

$$a_k(t) = a_k(0) \, \mathrm{e}^{-\frac{1}{2}k^2\pi^2 t}.$$

Wrapping up, we find

$$u(x,t) = \sum_{k=1}^{\infty} g_k \, \mathrm{e}^{-\frac{1}{2}k^2\pi^2 t} \sin(k\pi x), \qquad \text{with } g_k = 2 \int_0^1 g(y) \sin(k\pi y) \, \mathrm{d}y. \tag{2.22}$$

Note that the term $\mathrm{e}^{-k^2\pi^2 t}$ decays extremely quickly as t and k get large. This means that for large t (i.e. for maturities far in the future), the first few terms of the sum give a very good approximation of the solution.
Let us now write the solution in form of a Green's function: we write

$$G(x,y,t) = 2 \sum_{k=1}^{\infty} \mathrm{e}^{-\frac{1}{2}k^2\pi^2 t} \sin(k\pi x) \sin(k\pi y),$$

and then have

$$u(x,t) = \int_0^1 G(x,y,t) g(y) \, \mathrm{d}y.$$

G is the Green's function for the heat equation on the interval, and in the same way as for the half space, one can see that the solution of (2.20) with nonzero boundary conditions ϕ_0 and ϕ_1 and initial condition $g(x) = 0$ is given by

$$u(x,t) = \int_0^t \partial_y G(x,y,t)|_{y=0}\phi_0(s)\,\mathrm{d}s - \int_0^t \partial_y G(x,y,t)|_{y=1}\phi_1(s)\,\mathrm{d}s.$$

The general solution for nonzero initial and boundary conditions can then be found, as in the half-space case, by adding the two types of solutions that we have found above.

Remark: The backwards heat equation

Let us close with a remark about the backwards heat equation, given by $\partial_t u = -\partial_x^2 u$. Why do we never investigate it? The reason can be seen from equation (2.22): This would change to

$$u(x,t) = \sum_{k=1}^{\infty} g_k\,\mathrm{e}^{+\frac{1}{2}k^2\pi^2 t}\,\sin(k\pi x).$$

But the exponential term now grows when k grows, which means that unless g_k decays very quickly as $k \to \infty$, the 'solution' will be infinite for every positive time. This is related to the fact that the heat equation has such good smoothing properties: it converts e.g. any bounded function at time $t = 0$ into an infinitely differentiable function at $t > 0$ (just check on the solution formula); conversely, the backwards heat equation will be un-solvable for any initial condition that is not infinitely often differentiable, and will be in fact un-solvable for most other initial conditions. So we don't treat it.

2.6. **Uniqueness and the Maximum Principle.** So far, we have focused on finding solutions to the heat equation by giving explicit formulae. Here, we will investigate whether these solutions are the only ones there are. This question is of practical importance, as we can see when we recall how we approached the problem of option pricing: we found that

1) the option price must solve a PDE;

2) we can find a solution to the PDE.

But who guarantees that the solution we found actually has any connection to the option price? Without uniqueness, it might well be that the option price is given by a solution to the PDE that we did not find. To be sure that by finding a solution to the PDE we have found the correct option price, we therefore need uniqueness. Fortunately, it holds. A first step to proving it is the following result:

Proposition: Consider the heat equation on a bounded open domain $D \subset \mathbb{R}^n$:

$$(*)\begin{cases} \partial_t u = \Delta u \text{ for } x \in D, t > 0; \\ u(x,0) = g(x), \qquad u(y,t) = \phi(y) \text{ for } y \in \partial D. \end{cases}$$

Assume that

(A) the only solution to $(*)$ with $g = 0$ and $\phi = 0$ is the constant solution $u = 0$.

Then, the solution of $(*)$ is unique for all g and all ϕ.

Proof. Assume that u and $\tilde{u}$ solve $(*)$ with the same initial and boundary conditions. Then $u - \tilde{u}$ solves $(*)$ with zero initial and boundary conditions, by linearity. By assumption (A), this means that $u - \tilde{u} = 0$, and thus $u = \tilde{u}$. □

It remains to show that assumption (A) is valid. For this we use the **Maximum Principle**:
Theorem: Let $D \subset \mathbb{R}^n$ be a bounded open set. Let $f : D \times [0, T] \to \mathbb{R}$ satisfy the inequality

$$\partial_t f \leqslant \Delta f(x,t) \qquad \text{for all } x \in D, 0 < t < T.$$

Then the maximum of f over the set $\overline{D} \times [0, T]$ (overline means closure) is taken either on the spacial boundary $\partial D \times [0, T]$, or at the initial boundary $D \times \{0\}$.
Remarks: a) it follows that the maximum can neither be in the interior of the space-time domain, nor at the final time (except when it is also on the boundary of the space domain).
b) From the maximum principle, (A) follows easily: since u solves $\partial_t u = \Delta u \leqslant \Delta u$, and since $u = 0$ on the boundary, we find that $u(x,t) \leqslant 0 =$ the value at the boundary. And, since $\partial_t(-u) = \Delta(-u) \leqslant \Delta(-u)$, and $-u = 0$ on the boundary, we find that $-u(x,t) \leqslant 0$. So we must have $u(x,t) = 0$.

Proof of the Theorem. Fist let us see what happens when f fulfils the strict inequality $\partial_t f < \Delta f$, i.e. $\partial_t f - \Delta f < 0$. If f is maximal, inside $D \times [0, T]$, say at (y, t), then $\partial_t f(y,t) = 0$, since at the maximum all partial derivatives vanish. Also, $\partial_{y_i}^2 f(y,t) \leqslant 0$, since at the maximum all second partial derivatives are negative or zero. So, $\partial_t f - \Delta f \geqslant 0$, which contradicts the strict inequality. Thus f cannot have a maximum in the interior. Now for the final time, if we have a maximum we must have $\partial_t f(y, T) \geqslant 0$, since otherwise we could follow the slope into the interior of the domain and find points where f is even larger. Thus the same argument as above shows that f cannot have a maximum here, either.
Having sorted the case where the strict inequality holds, let us now consider f with only $\partial_t f - \Delta f \leqslant 0$. We define $f_\varepsilon(x,t) = f(x,t) - \varepsilon t$. Then

$$\partial_t f_\varepsilon(x,t) = \partial_t f(x,t) - \varepsilon \leqslant \Delta f(x,t) - \varepsilon < \Delta f(x,t) = \Delta f_\varepsilon(x,t).$$

The last equality is because the term εt that we added does not depend on x. We see that f_ε fulfils the strict inequality and therefore takes its maximum at the boundary. On the other hand, $f_\varepsilon(x,t) \leqslant f(x,t) \leqslant f_\varepsilon(x,t) + \varepsilon T$, and thus

$$\max_{\text{boundary}} f(x,t) \geqslant \max_{\text{boundary}} f_\varepsilon(x,t) = \max_{\text{whole set}} f_\varepsilon(x,t) \geqslant \max_{\text{whole set}} f(x,t) - \varepsilon T.$$

As this is valid for any ε, we can take the limit $\varepsilon \to 0$ in the above inequality and show the claim. □

The maximum principle also holds for unbounded domains, but it does need extra assumptions. Indeed, it can be shown that there are solutions to the whole space heat equation with initial condition $u(x, 0) = 0$ that are nonzero for positive times. These are called 'Tikhonov's great blast of heat from infinity', after the Russian

mathematician Andrey Nikolayevich Tikhonov who found them. However, some mild growth conditions on the solution at spatial infinity can restore uniqueness.
Theorem: If f solves $\partial_t f = \Delta f$ for $t > 0$, $f(x,0) = 0$ for all $x \in \mathbb{R}^n$, and if $f(x,t) \leqslant M\,\mathrm{e}^{c|x|^2}$ for some $M, c > 0$ and all $x \in \mathbb{R}^n, t > 0$, then we must have $f(x,t) = 0$ for all x, t.

Proof. We first prove it for $t < t_0$ for some t_0 to be fixed shortly. By induction we then know it to hold for $t < 2t_0$, $t < t_0$ et cetera, and thus eventually for all t. So let us assume that f fulfils all the conditions of the theorem. We define

$$g(x,t) = f(x,t) - \frac{\delta}{(t-t_1)^{n/2}}\,\mathrm{e}^{\frac{|x|^2}{4(t_1-t)}}\,,$$

where t_1 and δ are for the moment arbitrary. It can be checked directly that g solves the heat equation. Now write $D_R = \{x \in \mathbb{R}^n : |x| < R\}$ for the ball of radius R. We first establish that when we take R large enough, and t_1 small enough, then on the boundary of the space-time domain $D_R \times [0, t_1]$ we have $g \leqslant 0$. Indeed, $g(x,0) < f(x,0) = 0$, and for $x \in \partial D$ we have

$$\begin{aligned} g(x,t) = f(x,t) - \frac{\delta}{(t-t_1)^{n/2}}\,\mathrm{e}^{\frac{R^2}{4(t_1-t)}} &\leqslant M\,\mathrm{e}^{c|x|^2} - \frac{\delta}{(t-t_1)^{n/2}}\,\mathrm{e}^{\frac{R^2}{4(t_1-t)}} \leqslant \\ &\leqslant M\,\mathrm{e}^{cR^2} - \frac{\delta}{(t-t_1)^{n/2}}\,\mathrm{e}^{\frac{R^2}{4(t_1-t)}}\,. \end{aligned}$$

For t_1 such that $\frac{1}{4t_1} > C$, the second function grows faster than the first, and thus for R large enough, indeed $g(x,t) < 0$ on ∂D. What's more, this will continue to hold for any $\tilde{R} > R$. So for all $\tilde{R} > R$, and all $t_0 < t_1$, we can apply the finite domain maximum principle to find $g(x,t) \leqslant 0$ on $D_{\tilde{R}} \times [0, t_0]$. Therefore indeed $g(x,t) < 0$ on $\mathbb{R}^d \times [0, t_1]$ (each point of $\mathbb{R}^d$ is in a sufficiently large ball). Translating back to f, we find

$$f(x,t) \leqslant \frac{\delta}{(t_1-t)^{n/2}}\,\mathrm{e}^{\frac{|x|^2}{4(t_1-t)}}\,.$$

We have not yet used the parameter δ. Now we see that the above inequality holds foe any $\delta > 0$, and so by taking the limit $\delta \to 0$, we must have $f(x,t) \leqslant 0$ for all $t \leqslant t_0$. The same argument can now be applied to $-f$, yielding $f(x,t) \geqslant 0$ on the same set. In conclusion, we find $f(x,t) = 0$. □

3. Optimal Control

3.1. **The Problem, and Heuristics.** Let $y(t)$ be a quantity (e.g. production rate at a factory, portfolio value, etc.), that satisfies an ordinary differential equation:

$$\partial_t y(t) = F(y(t), \alpha_q(t), \alpha_2(t), \ldots).$$

Above, $\alpha_1, \alpha_2, \ldots$ are parameters that may depend on time and which influence F.

Example:

$$\partial_t y(t) = ry(t) - \alpha(t), \qquad y(0) = x. \tag{3.1}$$

This can of course be solved, and the solution

$$y(t) = \mathrm{e}^{rt}\left(x - \int_0^t \mathrm{e}^{-rs}\,\alpha(s)\,\mathrm{d}s\right)$$

depends on $\alpha(s)$. Let now h be a utility function, which can depend on y and α; for example, the utility function could be lower when α is large (cost of controlling something), and higher when y is high.

The *problem of optimal control* is this: find *control parameters* $\alpha_1(s), \alpha_2(s), \ldots$ such that h is maximal when averaged over time. In other words, we have to find

$$u(x,t) = \max_\alpha \int_t^T h(y(s), \alpha(s))\,\mathrm{d}s, \tag{3.2}$$

where initially (at time t), the value of $y(t) = x$. Importantly, $y(s)$ itself depends both on x and on the control $\alpha(s)$ through the ODE, e.g. (3.1).

There are many variants of the optimal control problem. One can let h depend explicitly on time. The most important instance of this is *discounting*, when we have $\mathrm{e}^{-rs}\,h(y(s), \alpha(s))$ under the integral. Or, we can have a final time utility $g(y(T))$ added to u.

The example (3.1) corresponds to an optimal consumption problem. There $y(t)$ is the wealth at time t, $\alpha(t)$ is the rate of consumption, y_0 is the initial wealth; wealth that has not been consumed earns interest at rate r. The control problem is to maximize

$$u(x,t) = \int_0^T \mathrm{e}^{-\rho s}\,h(\alpha(s))\,\mathrm{d}s. \tag{3.3}$$

h is usually a concave function, such as $h(\alpha) = \alpha^{1/2}$ or, more generally, $h(\alpha) = \alpha^\gamma$ with $0 < \gamma < 1$. The concavity has a nice interpretation: while it may make you happier to spend more money per day, it will not make you twice as happy; or, the less you are used to get by with, the more you will appreciate a bit more money to spend. T is the final time by which you should have spent it all, wealth that is left at that time will not benefit you.

It is a bit confusing that (3.3) seems not to depend on y at all. But in fact it does. Firstly, x is the starting point of $y(s), s \geqslant t$, and secondly, we have the condition that $y(s) \geqslant 0$ for all s, i.e. we adhere to the old-fashioned strategy that you cannot spend more than you have. This in turn restricts the controls α, since a control that is too large at the beginning will have to be zero after $y(s)$ hits zero, which happens sooner if α is large. Also, of course $\alpha \geqslant 0$, i.e. the consumption should be non-negative.

In the following we will give a general strategy to solve optimal control problems. The problem of optimal consumption will then be done in an exercise.

3.2. **Solving the optimal control problem: dynamic programming.** A general optimal control problem consists of the following parts:
The ODE:

$$\partial_s \boldsymbol{y}(s) = F(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)), \qquad \boldsymbol{y}(t) = \boldsymbol{x},$$

with $\boldsymbol{y} : \mathbb{R} \to \mathbb{R}^n$, $\boldsymbol{\alpha} : \mathbb{R} \to \mathbb{R}^m$ and $F : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^n$.
The control constraint: $\boldsymbol{\alpha}(s) \in A \subset \mathbb{R}^m$ for all s. This constraint may not always be present, and A may even depend on time.
The state constraint: $y(t) \in Y \subset \mathbb{R}^n$ for all t. This is usually a difficult constraint. The way we deal with it in the optimal consumption problem is to ignore it and verify afterwards that the solution fulfils it. This is not always possible.
The control problem: Let $h : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ be the running utility function, and $g : \mathbb{R}^n \to \mathbb{R}$ be the final utility function. The control problem is to find the optimal value function

$$u(\boldsymbol{x}, t) = \max_{\boldsymbol{\alpha} \in A, \boldsymbol{y} \in Y} \left(\int_t^T h(\boldsymbol{y}_{\boldsymbol{\alpha},\boldsymbol{x}}(s), \boldsymbol{\alpha}(s)) \,\mathrm{d}s + g(\boldsymbol{y}_{\boldsymbol{\alpha},\boldsymbol{x}}(T)) \right). \tag{3.4}$$

The subscript on the function $\boldsymbol{y}$ is there to remind us that it actually depends on the starting point and the control, but it will be often omitted in the notation.
The idea when solving this is to work backwards from the final time T. We ignore the state constraint for now.
1) For $t = T$, it is clear that there is nothing to control, and $u(\boldsymbol{x}, T) = g(\boldsymbol{x})$.
2) Let δt be a very small time step. Let us consider $t = T - \delta t$. We try to optimize (3.4) by using a control $\boldsymbol{\alpha}$ that is constant on $[T - \delta t, T]$. In the limit when $\delta t \to 0$ there is hope that this is good enough. We approximate the solution of the ODE by Taylor expansion:

$$\boldsymbol{y}(t + s) \approx \boldsymbol{y}(t) + \partial_t \boldsymbol{y}(t) s = \boldsymbol{y}(t) + F(\boldsymbol{y}(t), \boldsymbol{\alpha}(t)) s \tag{3.5}$$

for $s \in [0, \delta t]$. We also approximate $h(\boldsymbol{y}(t+s), \boldsymbol{\alpha}(t+s)) \approx h(\boldsymbol{y}(t)\boldsymbol{\alpha}(t))$ (zero order Taylor), and

$$\int_{t=T-\delta t}^T h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)) \,\mathrm{d}s \approx (T - t) h(\boldsymbol{y}(t), \boldsymbol{\alpha}(t)) = \delta t h(\boldsymbol{y}(t), \boldsymbol{\alpha}(t)). \tag{3.6}$$

Thus

$$\begin{aligned} u(\boldsymbol{x}, T - \delta t) &= \max_{\boldsymbol{\alpha}(s)} \left(\int_{T-\delta t}^T h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)) \,\mathrm{d}s + g(\boldsymbol{y}(T)) \right) \\ &\approx \max_{\alpha} \left(h(\boldsymbol{x}, \boldsymbol{\alpha}) \delta t + g(\boldsymbol{x} + F(\boldsymbol{x}, \boldsymbol{\alpha})) \delta t \right) \\ &= \max_{\alpha} \Big(h(\boldsymbol{x}, \boldsymbol{\alpha}) \delta t + u\Big(\boldsymbol{x} + F(\boldsymbol{x}, \boldsymbol{\alpha}), T\Big) \delta t \Big). \end{aligned} \tag{3.7}$$

So, given that we know the optimal value function at time T (which we do, it is g), we also know it at time $T - \delta t$. Indeed, we get it by maximizing the known function on the right hand side of (3.7) over $\boldsymbol{\alpha}$, for all $\boldsymbol{x}$, e.g. by differentiating and looking for zeros. This even gives a numerical scheme for finding the optimal value

function and the optimal control, but that scheme is very impractical in higher dimensions. However, the above reasoning will soon lead to a PDE that one can treat easier, at least numerically.
3) Step 2 is now repeated: we know the value function at time $T - \delta t$, so we get it (by step 2) for time $T - 2\delta t$, and so on until we reach the initial time.
The insight that knowing the optimal utility at a time t_1 helps us determine it at an earlier time is important. Indeed, the equation

$$u(\boldsymbol{x}, t) = \max_{\boldsymbol{\alpha}(s)} \int_t^{t_1} h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)) \, \mathrm{d}s + u(\boldsymbol{y}(t_1), t_1),$$

for $t_1 > t$, can be be seen to be true in a similar way as above. This equation is called *dynamic programming principle* because it is at the basis of the algorithm given above.

3.3. **Hamilton-Jacobi-Bellman (HJB) equation.** Let us replace T by $s + \delta t$ in (3.7). We then have

$$u(\boldsymbol{x}, s) = \max_{\boldsymbol{\alpha}} \Big(h(\boldsymbol{x}, \boldsymbol{\alpha}) \, \delta t + u\Big(\boldsymbol{x} + F(\boldsymbol{x}, \boldsymbol{\alpha})\delta t, s + \delta t\Big)\Big).$$

We now do a Taylor expansion of the function $\delta t \mapsto u(\boldsymbol{x} + F(\boldsymbol{x}, \boldsymbol{\alpha})\delta t, s + \delta t)$, and find

$$u(\boldsymbol{x}, s) \approx \max_{\boldsymbol{\alpha}} \Big(h(\boldsymbol{x}, \alpha)\delta t + u(\boldsymbol{x}, s) + \nabla u(\boldsymbol{x}, s) \cdot F(\boldsymbol{x}, \boldsymbol{\alpha})\delta t + \partial_t u(\boldsymbol{x}, s) \, \delta t\Big).$$

We can now subtract $u(\boldsymbol{x}, s)$ from both sides. The $\approx$ sign really means that there are terms of order $(\delta t)^2$ that we ignored. Dividing by δt and sending $\delta t \to 0$ gives the *Hamilton-Jacobi-Bellman equation* (HJB equation):

$$\partial_s u(\boldsymbol{x}, s) + \max_{\boldsymbol{\alpha} \in A} (\nabla u(\boldsymbol{x}, s) \cdot F(\boldsymbol{x}, \boldsymbol{\alpha}) + h(\boldsymbol{x}, \boldsymbol{\alpha})) = 0. \tag{3.8}$$

There is a special notation that is often used for the HJB equation: we define the *Hamiltonian* $H(\boldsymbol{p}, \boldsymbol{x})$ as

$$H(\boldsymbol{p}, \boldsymbol{x}) = \max_{\boldsymbol{\alpha} \in A} (\boldsymbol{p} \cdot F(\boldsymbol{x}, \boldsymbol{\alpha}) + h(\boldsymbol{x}, \boldsymbol{\alpha})), \tag{3.9}$$

with $\boldsymbol{x}, \boldsymbol{p} \in \mathbb{R}^n$. Then (3.8) reads

$$\partial_s u + H(\nabla u, \boldsymbol{x}) = 0. \tag{3.10}$$

It is clear from (3.12) that for solving a control problem, we have to maximize over $\boldsymbol{\alpha}$ in (3.11) first, and then solve (3.12). In particular, to find the optimal control at space point $\boldsymbol{x}$ and 'momentum' $\boldsymbol{p}$, we need not know the solution u of (3.12). However, of course since we then replace $\boldsymbol{p}$ by ∇u, the solution will be fed back into H.
We have now found the equation for the optimal value function: It is the solution of (3.8) with final data $u(\boldsymbol{x}, T) = g(\boldsymbol{x})$. To know the value function is good enough for e.g. option pricing, when we allow the buyer to change some control parameter: we now know how much an investor can make at most, if they play the game in an optimal way. This should then be the fair price of the option.

But what about actually finding the optimal strategy $\boldsymbol{\alpha}(s)$ for $t \leqslant s \leqslant T$? For this, we have to plug the solution $u(\boldsymbol{x}, s)$ back into the second term of (3.8), and find the argmax for each time. More precisely: from (3.11) we obtain both $H(\boldsymbol{p}, \boldsymbol{x})$ for all $\boldsymbol{p}$ and all $\boldsymbol{x}$, and $\alpha_*(\boldsymbol{p}, \boldsymbol{x})$ as the argmax of the right hand side. Now we solve (3.12) with the H that we just obtained, and the correct final condition. Once this is done, we know $u(\boldsymbol{x}, s)$ for all $\boldsymbol{x}$ and all s, and thus also $\nabla u(\boldsymbol{x}, s)$. The ODE for the optimally controlled $\boldsymbol{y}$ is then

$$\partial_s \boldsymbol{y}(s) = F(\alpha_*(\nabla u(\boldsymbol{y}(s), s), \boldsymbol{y}(s)), s),$$

where now the right hand side only contains known functions of s and $\boldsymbol{y}(s)$. Once we have obtained the solution $\boldsymbol{y}_*(s)$ to this ODE (with initial condition $\boldsymbol{y}(t) = \boldsymbol{x}$), we can finally determine the purely time dependent optimal control for the case when $\boldsymbol{y}(t) = \boldsymbol{x}$. It is given by $\boldsymbol{\alpha}_*(\nabla u(\boldsymbol{y}_*(s), s), \boldsymbol{y}_*(s))$.

The thing we now have to check is that $\boldsymbol{y}(t)$ actually fulfils the state constraints. If it does, everything is fine. But if it does not, we have to go back to the dynamic programming approach and incorporate this at each step. There is no nice theory for this, and we will not do it.

We have derived the HJB equation using a lot of heuristic and non-rigorous steps. But once we have it, we can actually prove that it gives the optimal value function. This is the content of the following

Theorem: *Consider the general optimal control problem introduced at at the beginning of this subsection. Assume that $w(\boldsymbol{x}, t)$ solves the HJB equation* (3.8)*, with final condition $w(\boldsymbol{x}, T) = g(T)$. Assume also that $\boldsymbol{y}$ derived from w in the way discussed above fulfils the state constraint. Then $w(\boldsymbol{x}, t) = u(\boldsymbol{x}, t)$, where u is defined is the solution of the control problem.*

Proof. Assume that w is a solution of the HJB equation with final condition $w(\boldsymbol{x}, T) = g(\boldsymbol{x})$. Let $\boldsymbol{\alpha}_0(s)$ be an arbitrary control, and let $\boldsymbol{y}_0(s)$ be the solution of the controlled ODE with start $\boldsymbol{y}_0(t) = \boldsymbol{x}$ and control $\boldsymbol{\alpha}_0(s)$. Let us investigate the function $s \mapsto w(\boldsymbol{y}_0(s), s)$. We find

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}s} w(\boldsymbol{y}(s), s) &= \partial_s w(\boldsymbol{y}(s), s) + \nabla w(\boldsymbol{y}(s), s) \cdot \partial_s \boldsymbol{y}(s) \\ &= \underbrace{\partial_s w(\boldsymbol{y}(s), s) + \nabla w(\boldsymbol{y}(s), s) \cdot F(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)) + h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s))}_{(*)} - h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s)).\end{aligned}$$

Now for an arbitrary control, the expression $(*)$ is smaller or equal to zero. This is because, by the HJB equation that w satisfies, when we take the maximum over all possible values of $\boldsymbol{\alpha}(s)$ we get $(*) = 0$, so any other $\boldsymbol{\alpha}$ will give less. We can then integrate the last equation from t to T and get

$$w(\boldsymbol{y}(T), T) - w(\boldsymbol{y}(t), t) \leqslant -\int_t^T h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s))\, \mathrm{d}s,$$

and using $w(\boldsymbol{y}, T) = g(\boldsymbol{y})$ as well as $\boldsymbol{y}(t) = \boldsymbol{x}$, we find

$$g(\boldsymbol{y}(T)) + \int_t^T h(\boldsymbol{y}(s), \boldsymbol{\alpha}(s))\,\mathrm{d}s \leqslant w(\boldsymbol{x}, t).$$

So any arbitrary control can make the value function at most as large as $w(\boldsymbol{x}, t)$, and maximizing over the possible controls we find $u(\boldsymbol{x}, t) \leqslant w(\boldsymbol{x}, t)$. What remains to show is that $w(\boldsymbol{x}, t)$ is indeed a value function; so far we only know it to be the solution of a HJB equation. To see that w is indeed a value function, notice that when we choose the feedback control that we get from w (see the discussion before the theorem), then $(*)$ above is identically zero (independent of whatever $\boldsymbol{y}(s)$ happens to be). Thus, again by integrating, we see that w is indeed a value function. $\square$

3.4. **Stochastic optimal control.** We will now perturb the equation for the state $\boldsymbol{y}_t$ by noise, leading to the stochastic differential equation

$$\mathrm{d}\boldsymbol{y}_s = f(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)\,\mathrm{d}s + \sigma(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)\,\mathrm{d}\mathcal{W}_s, \tag{3.11}$$

where $\mathcal{W}_s$ is $\mathbb{R}^n$-valued Brownian motion. The control problem is to maximize the *expectation of the various utility functions*, giving the optimal value function

$$u(x,t) = \max_{\boldsymbol{\alpha}} \mathbb{E}_{\boldsymbol{y}(t)=\boldsymbol{x}}\Big(\int_t^T h(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)\,\mathrm{d}s + g(\boldsymbol{y}_T)\Big). \tag{3.12}$$

g and h are utility functions. The space of allowed controls is now such that for some $A \subset \mathbb{R}^m$, we need to have $\boldsymbol{\alpha}_s \in A$ for all s, and an additional condition is that $\boldsymbol{\alpha}_s$ is *adapted* to the Brownian motion; that is, $\boldsymbol{\alpha}_s$ depends only on the values $\{\mathcal{W}_r : r \leqslant s\}$. This condition is very natural, as it means that the controller cannot know the future of the (random) evolution modelled by the Brownian motion. Usually (and also in this lecture) it is enough to let the control $\boldsymbol{\alpha}_s$ depend only on s and $\boldsymbol{y}_s$, i.e. to consider a *feedback control.*
To find the HJB equation in this case, we proceed as in the deterministic case and work backwards from the final time T. Clearly,

$$u(\boldsymbol{x}, T) = \mathbb{E}_{\boldsymbol{y}_T=\boldsymbol{x}}(0 + g(\boldsymbol{y}_T)) = g(\boldsymbol{x}).$$

Assume now that we have found $u(\boldsymbol{x}, t+\delta t)$ for some small δt. Then, by the dynamic programming principle (which applies also to this case, as one can easily see),

$$\begin{aligned} u(\boldsymbol{x},t) &= \max_{\boldsymbol{\alpha}} \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^{t+\delta t} h(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)\,\mathrm{d}s + u(\boldsymbol{y}_{t+\delta t}, t+\delta t)\Big) \\ &\approx \max_{\boldsymbol{\alpha}} \big(h(\boldsymbol{x}, \boldsymbol{\alpha})\delta t + \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\big(u(\boldsymbol{y}_{t+\delta t}, t+\delta t)\big)\big)\,. \end{aligned}$$

The approximate identity in the last line is justified by the fact that $s \mapsto \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(h(\boldsymbol{y}_s, \boldsymbol{\alpha}_s))$ is continuous, and as δt is very small, the integral is approximately given by the initial value of the integrand times the length of the integration interval. We reformulate this to read

$$0 = \max_{\boldsymbol{\alpha}} \Big(h(\boldsymbol{x}, \boldsymbol{\alpha})\,\delta t + \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\big(u(\boldsymbol{y}_{t+\delta t}, t+\delta t) - u(\boldsymbol{y}_t, t)\big)\Big). \tag{3.13}$$

Using the same trick that we have applied many times in the first few weeks, we find

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\big(u(\boldsymbol{y}_{t+\delta t}, t+\delta t) - u(\boldsymbol{y}_t, t)\big) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^{t+\delta t} \mathrm{d}u(\boldsymbol{y}_s, s)\Big) \\ &= \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^{t+\delta t} \big(\partial_t u(\boldsymbol{y}_s, s) + \nabla u(\boldsymbol{y}_s, s)\cdot f(\boldsymbol{y}_s, \boldsymbol{\alpha}_s) + \frac{1}{2}\sigma(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)^2 \Delta u(\boldsymbol{y}_s, s)\big)\,\mathrm{d}s\Big) \\ &\approx \Big(\partial_t u(\boldsymbol{x}, s) + \nabla u(\boldsymbol{x}, t)\cdot f(\boldsymbol{x}, \boldsymbol{\alpha}_t) + \frac{1}{2}\sigma(\boldsymbol{x}, \boldsymbol{\alpha}_t)^2 \Delta u(\boldsymbol{x}, t)\Big)\delta t. \end{aligned}$$

The equality between first and second line above follows from the Itô formula and the fact that the expectation of a stochastic integral is zero (it is here that we

need $\boldsymbol{\alpha}_s$ to be adapted!). The approximate identity between the second and third line follows in the same way as the one we have just discussed. We now use this in (3.13), and after letting $\delta t \to 0$ we get the following
Theorem: *$u(\boldsymbol{x}, t)$ from (3.12) is the solution of the Hamilton-Jacobi-Bellman equation*

$$\partial_t u(\boldsymbol{x},t) + \max_{\boldsymbol{\alpha}\in A} \Big(f(\boldsymbol{x},\boldsymbol{\alpha}) \cdot \nabla u(\boldsymbol{x},t) + h(\boldsymbol{x},\boldsymbol{\alpha}) + \frac{1}{2}\sigma^2(\boldsymbol{x},\boldsymbol{\alpha}) \nabla u(\boldsymbol{x},t) \Big) = 0, \tag{3.14}$$

with final condition $u(\boldsymbol{x},T) = g(\boldsymbol{x})$.
Note that if σ does not depend on $\boldsymbol{\alpha}$, then (3.14) becomes

$$\partial_t u + H(\nabla u, \boldsymbol{x}) + \frac{1}{2}\sigma^2 \Delta u = 0,$$

with $H(\boldsymbol{p},\boldsymbol{x}) = \max_{\boldsymbol{\alpha}}(f(\boldsymbol{x},\boldsymbol{\alpha}) \cdot \boldsymbol{p} + h(\boldsymbol{x},\boldsymbol{\alpha}))$ the same as in the deterministic case! As in the deterministic case, the derivation above was not fully rigorous, but once we have the result, we can give a rigorous proof.

Proof of the Theorem. We first show that if v solves the HJB equation (3.14), then for any adapted (not prescient) control $\boldsymbol{\alpha}_s$ we have

$$v(\boldsymbol{x},t) \geqslant \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}} \Big(\int_t^T h(\boldsymbol{y}_s, \boldsymbol{\alpha}_s)\, \mathrm{d}s + g(\boldsymbol{y}_T) \Big).$$

The proof is similar to the deterministic case: we consider the path $\boldsymbol{y}_s$ resulting from the stochastic differential equation controlled by our chosen control $\boldsymbol{\alpha}_s$, and plug this into the solution v of the HJB equation. The Itô formula then gives

$$\begin{aligned} \mathrm{d}v(\boldsymbol{y}_s,s) &= \partial_s v(\boldsymbol{y}_s,s)\,\mathrm{d}s + \nabla_{\boldsymbol{y}} v(\boldsymbol{y}_s,s) \cdot \mathrm{d}\boldsymbol{y}_s + \frac{1}{2}\Delta v(\boldsymbol{y}_s,s)(\mathrm{d}\boldsymbol{y}_s)^2 = \\ &= \partial_s v(\boldsymbol{y}_s,s)\mathrm{d}s + \nabla_{\boldsymbol{y}} v(\boldsymbol{y}_s,s) \cdot \Big(f(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)\,\mathrm{d}s + \sigma(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)\mathrm{d}\mathcal{W}_s \Big) + \frac{1}{2}\sigma(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)^2 \Delta v(\boldsymbol{y}_s,s)\,\mathrm{d}s. \end{aligned}$$

Since $\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(v(\boldsymbol{y}_T,T)) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(g(\boldsymbol{y}_T))$ by the final condition of the HJB equation, and since $\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(v(\boldsymbol{y}_t,t)) = v(\boldsymbol{x},t)$, we have

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(g(\boldsymbol{y}_T)) - v(\boldsymbol{x},t) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(v(\boldsymbol{y}_T,T)) - \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}(v(\boldsymbol{y}_t,t)) = \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^T \mathrm{d}v(\boldsymbol{y}_s,s) \Big) = \\ &= \mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^T \big(\partial_s v(\boldsymbol{y}_s,s) + \nabla_{\boldsymbol{y}} v(\boldsymbol{y}_s,s) \cdot f(\boldsymbol{y}_s,\boldsymbol{\alpha}_s) + \frac{1}{2}\sigma(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)^2 \Delta v(\boldsymbol{y}_s,s) + h(\boldsymbol{y}_s,\boldsymbol{\alpha}_s) - h(\boldsymbol{y}_s,\boldsymbol{\alpha}_s) \big)\,\mathrm{d}s \Big) \\ &\leqslant -\mathbb{E}_{\boldsymbol{y}_t=\boldsymbol{x}}\Big(\int_t^T h(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)\,\mathrm{d}s \Big). \end{aligned}$$

The last inequality follows, as in the deterministic case, from the fact that v solves the HJB equation, i.e. the maximum over all controls of all the terms under the integral except the last one is zero. So we find that

$$v(\boldsymbol{x},t) \geqslant \boldsymbol{E}_{\boldsymbol{y}_t=\boldsymbol{x}} \Big(\int_t^T h(\boldsymbol{y}_s,\boldsymbol{\alpha}_s)\,\mathrm{d}s + g(\boldsymbol{y}_T) \Big),$$

and maximizing over all controls gives that v is at least as large as the optimal value function. Now again, we can see that (for each path of the Brownian motion), the feedback control obtained from the HJB equation leads to the value function $v(\boldsymbol{x},t)$, so that $v(\boldsymbol{x},t)$ is not only a solution to the HJB equation, but indeed also a value function. Therefore it must be the optimal value function. □

3.5. **Application: Optimal portfolio selection and consumption.** This is a problem considered by Robert Merton (J. Econ. Theory 3, (1971) 373-413). Here is the setup:
b_s is a riskless asset with $\mathrm{d}b_s = rb_s\,\mathrm{d}s$, so $b_s = b_0\,\mathrm{e}^{rs}$.
p_s is a risky asset solving $\mathrm{d}p_s = \mu p_s \mathrm{d}s + \sigma p_s \mathrm{d}\mathcal{W}_s$.
x is our wealth at the starting time t.
The control parameters are $\alpha_1(s)$, the fraction of wealth in the risky asset p_s at time s; clearly, we need $0 \leqslant \alpha_1 \leqslant 1$.
$\alpha_2(s)$ is our rate of consumption at time s. We want $\alpha_2 \geqslant 0$.
The equation for the total wealth controlled by α_1 and α_2 is then

$$\mathrm{d}y_s = (1-\alpha_1(s))y_s r\mathrm{d}s + \alpha_1(s)y_s(\mu\,\mathrm{d}s + \sigma\,\mathrm{d}\mathcal{W}_s) - \alpha_2(s)\,\mathrm{d}s. \tag{3.15}$$

We impose the state constraint $y_s \geqslant 0$. The most elegant way to do this is to define $\tau(x) = \inf\{s \geqslant t : y_s = 0\}$; we then have the optimal value function (with discounting) given by

$$u(x.t) = \max_{\alpha_1,\alpha_2} \mathbb{E}_{y_t=x}\Big(\int_t^{\min(T,\tau(x))} \mathrm{e}^{-\rho s}\,h(\alpha_2(s))\,\mathrm{d}s\Big).$$

We want the utility function h to be monotone increasing and concave, as in the deterministic case. Our eventual choice will be $h(\alpha) = \alpha^\gamma$ with $0 < \gamma < 1$.
The derivation of the HJB equation with discounting is entirely parallel to the general case that we just treated. The result is

$$\partial_t u + \max_{\alpha_1,\alpha_2}\Big(\mathrm{e}^{-\rho t}\,h(\alpha_s) + (xr + \alpha_1(\mu-r)x - \alpha_2)\partial_x u + \frac{1}{2}x^2\sigma^2\alpha_1^2\partial_x^2 u\Big) = 0. \tag{3.16}$$

We can find the optimal α_1 by simple differentiation: the determining equation is

$$x(\mu-r)\partial_x u + \sigma^2 x^2 \partial_x^2 u\alpha_1 = 0,$$

giving for the optimal α_1:

$$\alpha_1^* = -\frac{(\mu-r)\partial_x u}{\sigma^2 x\partial x^2 u}. \tag{3.17}$$

Have we actually found a maximum? Only if $\partial_x^2 u > 0$, otherwise it would be a minimum! We need to keep the in mind and check at the end that it holds. Also, we have so far ignored the constraint $0 \leqslant \alpha_1 \leqslant 1$. We will have to come back to this later, too.
The optimal α_2 is now determined by the equation

$$\mathrm{e}^{-\rho t}\,h'(\alpha_2) = \partial_x u(x,t), \tag{3.18}$$

where h' is the derivative of h. It is intuitively clear that $\partial_x u > 0$, as greater initial wealth will give greater optimal value (try to find a mathematical argument for this!). Also, $h' > 0$ by the assumption that h is monotone increasing, $h'' < 0$ by concavity. So the optimal $\alpha_2^* > 0$ is nonnegative.
We now specialize to the case $h(\alpha) = \alpha^\gamma$ to make further progress. In the same way as on last week's problem sheet, we see that u must be of the form $u(x,t) = g(t)x^\gamma$. Since the optimal value is certainly nonnegative, we will have $g(t) \geqslant 0$. Thus $\partial_x u = \gamma g(t) x^{\gamma-1}$, and $\partial_x^2 u = \gamma(\gamma-1)g(t)x^{\gamma-2}$. Note that this means $\partial_x^2 u < 0$, which was one of the conditions that we had to remember checking.
Now (3.17) becomes

$$\alpha_1^* = \frac{\mu - r}{\sigma^2(1-\gamma)},$$

and (3.18) reads

$$\alpha_2^* = (\,\mathrm{e}^{\rho t}\, g(t))^{1/(\gamma-1)} x.$$

We can see that $0 \leqslant \alpha_1^* \leqslant 1$ if

$$0 \leqslant \mu - r \leqslant \sigma^2(1-\gamma). \tag{3.19}$$

We will assume for the moment that this extra condition holds. Putting $u(x,t) = g(t)x^\gamma$ back into the equation, we find that g needs to satisfy

$$\partial_t g(t) + \nu\gamma g(t) + (1-\gamma) g(t) \Big(\,\mathrm{e}^{\rho t}\, g(t)\Big)^{1/(\gamma-1)} = 0,$$

with final condition $g(T) = 0$, and $\nu = r + \frac{(\mu-r)^2}{2\sigma^2(1-\gamma)}$. This is of the same form as the equation that we have seen in the deterministic optimal consumption problem, and by following the steps given in Problem 1 on Sheet 5, we find that the solution is

$$g(t) = \,\mathrm{e}^{-\rho t}\Big(\frac{1-\gamma}{\rho - \nu\gamma}\Big(1 - \,\mathrm{e}^{-\frac{(\rho-\nu\gamma)(T-t)}{1-\gamma}}\,\Big)\Big)^{1-\gamma}.$$

So, the optimal value function is $g(t)x^\gamma$ with the above $g(t)$. The optimal control α_1^* is constant, i.e. it depends neither on time nor on the current wealth. This means that our investment decision is not influenced by our current wealth, and also not by the time we still have to consume. Instead, it is fully determined by the difference $\mu - r$ of the expected return of the asset and the bond rate, divided by a factor $\sigma^2(1-\gamma)$. This factor σ^2 is easy to interpret: large uncertainty σ makes it more unattractive to invest in the risky asset. The factor $(1-\gamma)$ is less obvious. It means that if we can consume larger amounts of wealth with relatively little penalty (i.e. γ close to 1), then we should invest less into the risky asset. Now, all of this is for $\mu > r$. In the case $\mu < r$, our extra condition (3.19) does not hold; it is not difficult to see that in this case, $\alpha_1^* = 0$ is the optimal allowed control. So, if the expected rate of return for the risky asset is less than the bond rate, it is not worth investing into it at all. On the other hand, if $\mu - r > \sigma^2(1-\gamma)$, then the return of the risky asset is so much better that the bond rate, that we will put all our wealth into it, and $\alpha_1^* = 1$ in that case.

Unlike α_1^*, the optimal α_2^* in (3.18) does depend on time. Now that we know $g(t)$, we put it into (3.18) (this is part of the feedback!) and find

$$\alpha_2^*(x,t) = \frac{1-\gamma}{\rho-\nu\gamma}\Big(1-\mathrm{e}^{-\frac{(\rho-\nu\gamma)(T-t)}{1-\gamma}}\Big)x.$$

At given time s, our wealth will also be known to be y_s. So at that time, we replace x with y_s in the above equation. This is the second part of the feedback. This means that the optimally controlled asset solves the SDE

$$\mathrm{d}y_s = (r+(\mu-r)\alpha_1^*)y_s r\mathrm{d}s - \frac{1-\gamma}{\rho-\nu\gamma}\Big(1-\mathrm{e}^{-\frac{(\rho-\nu\gamma)(T-s)}{1-\gamma}}\Big)y_s + \alpha_1^*\sigma y_s\,\mathrm{d}\mathcal{W}_s.$$

The fact that $\alpha_2^*(y_s,s)$ is proportional to y_s guarantees that the whole right hand side of the SDE is proportional to y_s. This means that $y_s \geqslant 0$ (since, should it ever hit zero (from above, obviously), its time derivative will be zero, and it will be stuck there). Thus luckily the state constraint is automatically fulfilled, and indeed we do not need $\tau(x)$ in the end. The optimal consumption rate is easy as a function of wealth (proportional to it), but rather difficult as a function of time. It seems strange that it goes to zero as $t \to T$. One would have thought it should go to infinity then, as there is nothing to lose by consuming it all in the last instant. I don't fully understand this intuitively. One explanation is that we are optimizing the *expected* utility, and so the optimal consumption is made so that in the last instant, on average there won't be much left to consume. But this is not fully clear to me.

4. Numerical solutions of PDE

There are very few PDE that can be solved analytically. Therefore, it is important to understand the techniques for solving PDE on a computer, but also the difficulties and pitfalls that can arise when we try to do so. Let us start with a simpler setting.

4.1. **Ordinary differential equations.**

Consider the ordinary differential equation

$$\partial_t y(t) = F(y(t), t), \qquad y(t_0) = y_0. \tag{4.1}$$

The simplest way to solve this equation on a computer is the *forward Euler scheme:*
Step 1: Discretize the t-axis with step size h, giving grid points $t_0, t_0+h, t_0+2h, \ldots$.
Step 2: Use Taylor-expansion to find that a solution of (4.1) fulfils

$$y(t+h) = y(t) + h\partial_t y(t) + \mathcal{O}(h^2) = y(t) + hF(y(t), t) + \mathcal{O}(h^2),$$

where the notation $\mathcal{O}(h^2)$ means the following: it is possible to find a function that, when written in place of the symbol $\mathcal{O}(h^2)$, ensures that the equality sign is a true statement. This function may depend on $t, y(t), h$ or whatever other parameters there are, but it must vanish at least as quickly as a constant times h^2 (or whatever expression we write into brackets after the $\mathcal{O}$), when $h \to 0$.
Step 3: Recall that the initial value of the ODE is y_0. Define

$$y_1 = y_0 + hF(y_0, t_0),\; y_2 = y_1 + F(y_1, t_2),\; \ldots,\; y_{n+1} = y_n + F(y_n, t_n).$$

The values y_j should approximate the values $y(t_j)$ of the true solution; after all, they solve the difference equation from Step 2 that is an approximation to the ODE.

The second step above was a bit arbitrary. We could as well have used

$$y(t-h) = y(t) - h\partial_t y(t) + \mathcal{O}(h^2) = y(t) - hF(y(t), t) + \mathcal{O}(h^2),$$

which would have led to the scheme

$$y_0 = y_1 - hF(y_1, t_1), y_1 = y_2 - hF(y_2, t_2), \ldots$$

This is the *backwards Euler scheme*, or implicit Euler scheme. In each step we now still have to solve a (possibly difficult) equation to obtain the value of y_{j+1} as a function of the value y_j and t_j. It is not clear at this moment why anyone would like to do such a thing, but we will see later when we treat PDE that there are large benefits in doing so.
However we do it, we hope that the values $y_1, y_2, \ldots$ approximate $y(t_1), y(t_2), \ldots$, where $y(t)$ is the true solution. We now want to quantify how good that approximation is. For this, consider a finite time interval $[t_0, T]$, and put $t_j = t_0 + hj$, with N such that $t_N = T$. Let y solve (4.1), and let y_n be the approximating solution under some numerical scheme. We define

$$e_j = |y_j = y(t_j)| = \text{ the error we make at time } t_j,$$

and

$$E = \max_{0 \leqslant j \leqslant N} e_j = \text{ the worst error we make on the interval.}$$

Definition: A numerical scheme is called *convergent* if $\lim_{h\to 0} E(h) = 0$. It is called *convergent of order* p if $E(h) \leqslant Ch^p$ for some $C > 0$ and all h small enough.

How can we check convergence? A good indicator would be how similar our numerical scheme is to the true ODE, for small h. Assume our numerical scheme is given by

$$y_{n+1} = y_n + h\phi(y_n, t_n, h) \tag{4.2}$$

for some function ϕ that may depend on y_n, t_n but also on h. In the forward Euler scheme we had $\phi(y_n, t_n, h) = F(y_n, t_n)$, so in that case ϕ was independent of h. Now let again $y(t)$ be the true solution of the ODE.

Definition: The *truncation error* $\tau_n(h)$ at step n is defined by the equation

$$y(t_{t+1}) = y(t_n) + h\Big(\phi(y(t_n), t_n, h) + \tau_n(h)\Big).$$

Interpretation: the true solution does not fulfil the recursive scheme (4.2), but instead fulfils another recursive scheme. $\tau_n(h)$ measures how different the two schemes are at time t_n and for discretisation paramter h.

Definition: A numerical scheme is called *consistent* if

$$\lim_{h\to 0}\Big(\max_{0 \leqslant n \leqslant N} \tau_n(h)\Big) = 0.$$

It is called *consistent of order* p if

$$\max_{0 \leqslant n \leqslant N} \tau_n(h) \leqslant Ch^p$$

for some $C > 0$ and all sufficiently small $h > 0$.

So, consistency means that the two *equations* become similar as $h \to 0$, while convergence means that the two *solutions* become similar. You should check that explicit Euler is consistent of order 1.
What about the connections between consistency and convergence? For ODE, the connection is rather simple.
Theorem: For a numerical scheme given by (4.2) let us assume that

$$\Big|\phi(y, t, h) - \phi(\tilde{y}, t, h)\Big| \leqslant L|y - \tilde{y}|, \tag{4.3}$$

for some $L > 0$, all $t \leqslant T$ and all $h < h_0$ with some $h_0 > 0$. (This is called the *Lipschitz condition.*) Assume that the scheme is consistent of order p. Then it is also convergent of order p.

Proof. On the exercise sheet you are asked to prove the following fact (discrete Gronwall Lemma): For nonnegative numbers $z_1, z_2, \dots$ assume that there are $C, D > 0$ such that $z_{n+1} \leqslant Cz_n + D$ for all n. Then

$$z_n \leqslant D\frac{C^n - 1}{C - 1} + z_0 C^n \tag{4.4}$$

for all n. We use this fact in the following argument: Since

$$\begin{aligned} y_{n+1} &= y_n + \phi(y_n, t_n, h) \\ y(t_{n+1}) &= y(t_n) + h\phi(y(t_n), t_n, h) + h\tau_n(h), \end{aligned}$$

we find

$$\begin{aligned} e_{n+1} = |y_{n+1} - y(t_{n+1})| &= e_n + h\Big(\phi(y_n, t_n, h) - \phi(y(t_n), t_n, h)\Big) + h\tau_n(h) \\ &\leqslant e_n + hL\Big|y_n - y(t_n)\Big| + h\tau_n(h) = (*). \end{aligned}$$

The inequality above is due to our assumption (4.3). Now we have assumed consistency of the scheme of order p, thus there is an $M > 0$ such that $\tau_n(h) \leqslant Mh^p$ for all n and all small enough h. Also, $|y_n - y(t_n)| = e_n$. So $(*) \leqslant e_n(1+hL) + Mh^{p+1}$, and by (4.4) with $C = 1 + hL$ and $D = Mh^{p+1}$, we get (notice that $e_0 = 0$)

$$e_n \leqslant Mh^{p+1}\frac{(1 + hL)^n - 1}{1 + hL - 1} = Mh^p\frac{1}{L}\Big((1 + hL)^n - 1\Big).$$

Finally,

$$\begin{aligned} 0 \leqslant (1 + hL)^n - 1 &\leqslant \left(1 + hL + \frac{(hL)^2}{2} + \frac{(hL)^3}{3!} + \dots\right) - 1 \\ &= \big(\mathrm{e}^{hL}\big)^n - 1 = \mathrm{e}^{hLn} - 1 \leqslant \mathrm{e}^{(T-t_0)L} - 1, \end{aligned}$$

where in the last equality we used that h discretizes the interval $[t_0, T]$, so for $n < N = N(h)$, hn can never be larger that $[t_0, T]$. □

Before leave the ODE case and look at numerics for PDE, let us give an example that shows how implicit schemes can be useful even for ODE. Consider the simple equation

$$\partial_t y(t) = -\lambda y(t), \qquad \lambda > 0, y(0) = y_0.$$

The solution is of course $y(t) = y_0\,\mathrm{e}^{-\lambda t}$, and it decays to 0 rapidly and monotonously as $t \to \infty$. Can we say the same for our numerical schemes? Let's look at the Euler forward scheme:

$$y_{n+1} = y_n + h\lambda y_n = (1 - \lambda h)y_n, \qquad \Longrightarrow \qquad y_n = (1 - h\lambda)^n y_0.$$

If λ is large, we will need h to be small for this scheme to give something sensible: indeed, $y_n \to 0$ as $n \to \infty$ *only* if $h \leqslant 2/\lambda$; and, $n \mapsto y_n$ is monotone decreasing *only* if $h \leqslant 1/\lambda$. So, while the true ODE becomes more and more easy to understand when λ gets large, the numerical scheme becomes more difficult to implement in the sense that the time step needed for a sensible solution becomes smaller. After

all, h will always be finite in real world situations, and the smaller we have to take it, the longer we need to run a computer in order to find $y(10)$, say.
The situation is different for the Euler backward scheme. In that scheme, we find

$$y_n = y_{n+1} + h\lambda y_{n+1} \qquad \Longrightarrow \qquad y_n = \frac{1}{(1+h\lambda)^n} y_0,$$

which is monotonously decreasing to zero regardless of the size of h. We say that the forward Euler scheme is *conditionally stable* (i.e. for small enough h it is stable), while the backward scheme is *unconditionally stable*. A proper definition of what a stable scheme means will be given below when we treat PDE.

4.2. **Forward schemes for PDE.** We consider the general PDE

$$\begin{aligned} &\partial_t u(x,t) = (Fu)(x,t) \qquad (a < x < b, t > 0), \\ &u(x,t_0) = u_0(x) \qquad \text{(initial condition)}, \\ &u(a,t) = g_a(t), u(b,t) = g_b(t) \qquad \text{(boundary conditions)}. \end{aligned} \tag{4.5}$$

Above, F can be any linear or nonlinear operator, such as

$$\begin{aligned} &Fu = \frac{\sigma^2}{2}\partial_x^2 u \qquad \text{(heat equation)}, \\ &Fu = -\frac{1}{2}\sigma^2 x^2 \partial_x^2 u + b(x\partial_x u - u) \qquad \text{(Black-Scholes PDE)} \\ &Fu = -\max_{\alpha \in A}\left(f(x,\alpha)\partial_x u(x,t) + h(x,\alpha) + \frac{1}{2}\sigma^2(x,\alpha)\partial_x^2 u \right) \qquad \text{(HJB equation)}. \end{aligned}$$

We now want to solve these numerically, so we discretize time as $t_n = t_0 + hn$ and space as $x_j = h_x j$ with $n \in \mathbb{N}$ and $j \in \mathbb{Z}$. h_x can and usually will be different from h, so we do not discretize the space-time domain into squares but rather rectangles. For discretizing the spatial derivatives we use again Taylor expansion. One possibility is

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

This can be useful if there is a preferred direction of space, but for the heat equation and related equations, there is none, and so the symmetric finite difference

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

is usually better. Whether we best use central of ordinary finite differences, or yet another approximation, depends on the equation and is more an art than a science. We will not go into this here and only use central spatial differences. The second derivative is thus approximated by

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

Higher derivatives can be disrcetized similarly, but we will not need them here. Plugging the discretized derivatives into Fu gives

$$Fu(x_j,t_n) \approx \tilde{F}\Big(u(x_{j+1},t_n), u(x_j,t_n), u(x_{j-1},t_n), t_n, h_x\Big),$$

for some function $\tilde{F}$, in the case where we have only second derivatives. For higher derivatives, the function on the right hand side above will depend on more values of $u(x,t_n)$, but we will not need this here. As an example, for the heat equation we find

$$\tilde{F}\Big(u(x_{j+1},t_n), u(x_j,t_n), u(x_{j-1},t_n), t_n, h_x\Big) = \frac{\sigma^2}{2h_x^2}\Big(u(x_{j+1},t_n)+u(x_{j-1},t_n)-2u(x_j,t_n)\Big).$$

The approximation to the PDE then becomes

$$u(x_j,t_{n+1}) \approx u(x_j,t_n) + h\tilde{F}\Big(u(x_{j+1},t_n), u(x_j,t_n), u(x_{j-1},t_n), t_n, h_x\Big), \tag{4.6}$$

and the numerical scheme is

$$u_{n+1}^j = u_j^n + h\tilde{F}(u_{j+1}^n, u_j^n, u_{j-1}^n, t_n, h_x),$$

with $u_j^0 = u_0(x_j)$ as initial condition and $u_j^n = g_{a,b}(t_n)$ if $x_j = a, b$, respectively, as boundary conditions. u_j^n are again what we would like to be approximations to $u(x_j,t_n)$. In our case, it is easy to compute the point $u(x_j,t_n)$; it is a known function (namely, $\tilde{F}$) of the points $u(x_{j-1},t_n)$, $u(x_{j+1},t_n)$, $u(x_j,t_n)$ from the previous time discretisation point, and so we can recursively get all values.

Of course, the question is again how good this approximation actually is. As in the ODE case, we define the *truncation error* that measures how different the numerical scheme is from the actual PDE:

Definition: The *truncation error* at point (x_j,t_n) of the numerical scheme (4.6) is defined to be

$$T(x_j,t_n) = \frac{1}{h}\Big(u(x_j,t_{n+1})-u(x_j,t_n)\Big)-\tilde{F}\Big(u(x_{j+1},t_n), u(x_j,t_n), u(x_{j-1},t_n), t_n, h_x\Big),$$

i.e. the extent to which the true solution does not solve the approximate PDE (difference equtaion).

Definition: The scheme is *consistent* if $\lim T(x,t) = 0$ for all x and t as h_x and h go to zero. Here, if we want to be very precise, we need to define $T(x,t)$ as the limit of $t(x_j,t_n)$ for sequences (t_j) and (x_n) of grid points with $x_j \to x$ and $t_n \to t$ as both h and h_x go to zero.

As for ODE, consistency is the minimal requirement that we have for any numerical scheme. Let us check consistency for the forward scheme of the heat equation: we have seen that in this case,

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

so the truncation error is

$$\begin{aligned}T(x_j,t_n) &= \frac{1}{h}\Big(u(x_j,t_{n+1})-u(x_j,t_n)\Big)-\frac{\sigma^2}{2}\frac{1}{h_x^2}\Big(u(x_{j-1},t_n)+u(x_{j+1},t_n)-2u(x_j,t_n)\Big)\\ &\to \partial_t u(x,t)-\frac{\sigma^2}{2}\partial_x^2 u(x,t)=0\end{aligned}$$

as $h, h_x \to 0$, and $x_j \to x$, $t_n \to t$. So, the scheme is consistent. Is it convergent? There seems to be little hope unless $h \leqslant Ch_x^2$ for some $C > 0$ since otherwise the prefactor h/h_x^2 for getting from one time step to the next would diverge as $h \to 0$. But as we will see in the next section, even this condition does not ensure convergence.

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 $\boldsymbol{u}^n = (u_1^n, \ldots, u_{J-1}^n)$. Then (4.7) reads

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

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 $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{J-1}$ be the eigenvectors and $\lambda_1, \ldots, \lambda_{J-1}$ 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 $k = 1, \ldots J-1$. 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

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

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. □