Numerical analysis
of partial differential equations
arising in finance

J. Frédéric Bonnans
CHAPTER 1

Finite differences methods

Contents
Section 1 presents some partial differential equations (PDE) of parabolic type, arising in the evaluation of options in finance, and discusses their transformations under various changes of variables. Postponing any discussion on the well-posedness of the PDEs, sections 2 and 3 introduce explicit and implicit finite difference (FD) algorithms. Assuming, in the case of explicit schemes, certain limitations on the size of the time step, we show that the schemes are simply monotonic and that, if the data are uniformly bounded, the same holds for their solutions.

The previous analysis dealt with the relatively easy case of mono dimensional problems (space dimension equal to one) Algorithms for the multi dimensional case are discussed in sections 4. We present the “classical” finite difference scheme that needs the restrictive condition of diagonally dominant scaled diffusion matrices. Then we show in section 5 how to deal with larger classes of diffusion matrices using larger stencils, in the framework of the generalized finite differences (GFD) schemes.

Section 6 establishes error estimates in the uniform norm, by estimating the consistency error, assuming the solution of the PDE to be smooth enough. When the solution of the PDE is only Hölder continuous, we show how to combine this analysis with an approach of regularization by convolution. Energy estimates are established in section 7 both in the mono dimensional case and in the framework of generalized finite differences. Finally section 8 shows how to deal with the case of unbounded r.h.s. and final condition by a certain change of variable.

1. Informal discussion of the basic model, change of variables

1.1. First steps. In the standard Black-Scholes model [8], the value of European options are solutions of PDEs of the form

\[ v_t + xr(x,t)v_x + \frac{1}{2}x^2\sigma^2(x,t)v_{xx} - r(x,t)v = 0, \quad (x,t) \in \mathbb{R}_+ \times [0,T], \]

\[ v(x,T) = g(x), \quad x \in \mathbb{R}_+. \]

Here the data are the actualization coefficient \( r \geq 0 \), the volatility \( \sigma > 0 \), and the payoff function \( g \). We assume that \( r \) and \( \sigma \) are bounded and continuous functions over \( \mathbb{R}_+ \times [0,T] \). The payoff \( g \) is often unbounded, and sometimes discontinuous. Examples of payoff are the call and put

\[ g_{\text{call},K} := (x-K)_+; \quad g_{\text{put},K} := (K-x)_+, \]

where \( K > 0 \) is called the strike. There are also digital options, an example of which is

\[ 1_K(x) = \begin{cases} 
1 & \text{if } x \geq K, \\
0 & \text{otherwise.} 
\end{cases} \]

Financial interpretation. We briefly recall, in an informal way, the financial interpretation of the above equation. Given a financial asset \( S(t) \), called the underlying, with dynamics

\[ dS(t) = \mu(S(t),t)dt + \sigma(S(t),t)dW(t), \]

where \( W(t) \) is a standard Brownian motion, \( \mu(S,t) \) is the drift (whose value has no incidence in (1.1)), and \( r = r(S,t) \) is the (varying) interest rate. The price \( V(x,t) \) at time \( t \in [0,T] \) and when \( S = x \) of the option \( g(S(T)) \), whenever it exists, is defined as the amount of money that, partly invested (without transaction costs) in the asset \( S(t) \) and in a risky asset (with interest rate \( r(S,t) \)) in a dynamic way, allows to obtain the amount
1. FIRST STEPS IN FINITE DIFFERENCES METHODS

Invariance of identity. If \( g(x) = x \) then the solution of \( (1.1) \) is \( v(x, t) = x \). This is not surprising in view of the financial interpretation: the investment consist in buying one underlying asset and waiting until time \( T \).

Constant final value. In the case when \( g(x) = c \) (constant), if \( r(x, t) = r(t) \) is a function of time only, the solution is \( v(x, t) = ce^{-\int_0^t r(s)ds} \), and in particular, \( v(x, t) = ce^{-r(T-t)} \) if \( r \) is constant. Again, the financial interpretation is clear: the investment consist in buying \( ce^{-r(T-t)} \) units of the non risky asset.

Portfolios. By combining finitely many call and put options (it suffices of call options) for arbitrary strikes we obtain an arbitrary payoff function in the class of piecewise constant (with finitely many pieces) functions. Similarly, combining finitely many digital options, we obtain arbitrary payoff function in the class of piecewise affine (with finitely many pieces) functions. We can in turn use functions in these two classes in order to approximate more general payoff functions.

In the numerical analysis that will follow, we may of course expect better approximations in the case of a Lipschitz payoff than in the case of a discontinuous one. In any case, let us note that, in this setting, the payoff function typically have linear growth: \( |g(x)| = O(1 + |x|) \), with possible discontinuities.

1.2. Change of variables. Assuming \( (1.1) \) to have a solution \( v \) in the classical sense, i.e., with continuous first (second) derivatives in time (space), we observe the consequence of some changes of variables.

Actualization. Let \( V(x, t) := \beta(t)v(x, t) \), where \( \beta(t) \) plays the role of an artificial actualization coefficient, of the form \( \beta(t) = e^{-\int_0^t \alpha(s)ds} \). Then

\[
\begin{align*}
V_t(x, t) &= \beta(t)(v_t(x, t) - \alpha(t)v(x, t)), \\
V_z(x, t) &= \beta(t)v_z(x, t), \\
V_{zx}(x, t) &= \beta(t)v_{zx}(x, t).
\end{align*}
\]

Multiplying the first row in \( (1.1) \) by \( \beta(t) \) and substituting the partial derivatives of \( v \), we obtain

\[
V_t + xr(x, t)V_z + \frac{1}{2}x^2\sigma^2(x, t)V_{xx} + (\alpha(t) - r(x, t))V = 0,
\]

with terminal condition \( V(x, t) = \beta(T)g(x) \). We see that the result is to shift the coefficient of the zero order term, leaving the other terms unchanged. In particular, if \( r \) depends only on time, taking \( \alpha(t) = r(t) \), we obtain a similar PDE, but without the zero order term.

Logarithmic change of space variable. Consider the change of space variable \( z = \log x \). The function \( V(z, t) := v(e^z, t) \) satisfies, skipping arguments

\[
V_{t} = v_{t}; \quad V_z = e^z v_x = x v_z; \quad V_{zz} = e^{2z} v_{xx} + e^z v_x = x^2 v_{xx} + x v_x.
\]

Substituting \( v \) and its partial derivatives in \( (1.1) \), we obtain

\[
V_t(z, t) + (r(e^z, t) - \frac{1}{2}\sigma^2(e^z, t))V_z(z, t) + \frac{1}{2}\sigma^2(e^z, t)V_{zz}(z, t) - r(e^z, t)V(z, t) = 0,
\]

with final condition \( V(z, T) = g(e^z) \). The formulation \( (1.8) \) has the important advantage of having bounded coefficients. On the other hand, in the case e.g. of a European call,
the final condition is \((e^z - K)_+\): the linear growth in the original setting now becomes an exponential growth.

**Characteristic-like transformations.** Consider the ordinary differential equation 
\[
\dot{z}(t) = c(t),
\]
with terminal condition \(z(T) = z\), whose solution is
\[
\zeta(t) = z - \int_t^T c(s)\,ds.
\]
The transformation \((z, t) \mapsto (\zeta(z, t), t)\) is bijective, and we perform the change of variables
\[
\tilde{V}(z, t) = V(\zeta(z, t), t),
\]
where \(V\) is solution of (1.8). Using
\[
\left\{ \begin{array}{l}
\tilde{V}_t(z, t) = \tilde{V}_t(\zeta(z, t), t) + c(t)\tilde{V}_z(\zeta(z, t), t), \\
\tilde{V}_z(z, t) = V_2(\zeta(z, t), t), \\
\tilde{V}_{zz}(z, t) = V_{zz}(\zeta(z, t), t),
\end{array} \right.
\]
we obtain, writing \(\zeta\) for \(\zeta(z, t)\):
\[
\tilde{V}_t(z, t) + (r(\zeta, t) - c(t) - \frac{1}{2}\sigma^2(\zeta, t))\tilde{V}_z(z, t) + \frac{1}{2}\sigma^2(\zeta, t)\tilde{V}_{zz}(z, t) - r(\zeta, t)\tilde{V}(z, t) = 0.
\]
Setting
\[
\tilde{r}(z, t) := r(\zeta(z, t), t); \quad \tilde{\sigma}(z, t) := \sigma(\zeta(z, t), t);
\]
we can rewrite the PDE (1.11) in the form (coming back to the notation \(z\) for the space variable)
\[
\tilde{V}_t(z, t) + (\tilde{r}(z, t) - c(t) - \frac{1}{2}\tilde{\sigma}^2(z, t))\tilde{V}_z(z, t) + \frac{1}{2}\tilde{\sigma}^2(z, t)\tilde{V}_{zz}(z, t) - \tilde{r}(z, t)\tilde{V}(z, t) = 0,
\]
with \(z \in \mathbb{R}\), and the same final condition \(\tilde{V}(z, t) = g(e^z)\).

**Remark 1.1.** If \(r\) is constant, then we can combine the space and time transformation in order to cancel the term without derivatives. If \(\sigma\) is also constant then taking \(c = r - \frac{1}{2}\sigma^2\), we reduce the PDE to \(\tilde{V}_t + \frac{1}{2}\tilde{\sigma}^2\tilde{V}_{zz} = 0\). By an linear change of time variables we further reduce to the (backward) heat equation, after a change of notation, and with given final condition denoted by \(\tilde{g}\):
\[
v_t(x, t) + v_{xx}(x, t) = 0, \quad (x, t) \in \mathbb{R} \times [0, T], \quad v(x, T) = \tilde{g}(x).
\]

**Remark 2.1.** The characteristic-like transformation is of interest even if the coefficients are not constant, since it may allow to reduce the size of the coefficient of the first order space derivative. This may be useful for the convergence of the numerical schemes.

**Convolution of functions.** We recall that, given two measurable functions \(f_1\) and \(f_2\) over \(\mathbb{R}^n\), their convolution denoted by \(f_1 \ast f_2\) is defined (whenever the integral below makes sense) by
\[
f = f_1 \ast f_2(x) := \int_{\mathbb{R}^n} f_1(x - s)f_2(s)\,ds.
\]
This is the case, for instance, if \(f_1\) and \(f_2\) belong to \(L^1(\mathbb{R}^n)\), and then we have that \(f\) also belongs to \(L^1(\mathbb{R}^n)\), since after an obvious change of variables
\[
\int |f(x)|\,dx \leq \int \int |f_1(x - s)||f_2(s)|\,ds\,dx \leq \int |f_1(x)|\,dx \int |f_2(s)|\,ds
\]
so that
\[(1.17)\quad \|f_1 \ast f_2\|_1 \leq \|f_1\|_1 \|f_2\|_1.\]

Another case of interest (considered in section 6) is when \(f_1 \in L^\infty(\mathbb{R}^n)\) and \(f_2 \in L^1(\mathbb{R}^n)\).

We easily check then that
\[(1.18)\quad \|f_1 \ast f_2\|_\infty \leq \|f_1\|_\infty \|f_2\|_1.\]

In our present application we need to consider the more general case when \(f_1\) and \(f_2\) are locally integrable (i.e., their restriction to any bounded measurable set is integrable), and satisfy
\[(1.19)\quad \left|f_1(x-s)\right| \leq a_1(x)a_2(s); \quad a_2(s)f_2(s) \text{ is integrable.}\]

**Constant parameters.** We have shown that for constant \(r\) and \(\sigma\) we can reduce (1.1) to the heat equation (1.14). The solution to the latter is known to have a convolution product form:
\[(1.20)\quad v(x,t) = (G(\cdot,T-t) \ast \tilde{g}) \left( x \right) = \int_\mathbb{R} G(x-y,T-t)\tilde{g}(y)dy,\]

where \(G(x,t)\) is the heat kernel, or fundamental solution of the heat equation in forward time, when the initial condition is a Dirac measure at \(x = 0\):
\[(1.21)\quad G(x,t) = \frac{1}{(4\pi t)^{1/2}} e^{-|x|^2/(4t)}.\]

We have therefore
\[(1.22)\quad v(x,t) = \frac{1}{(4\pi(T-t))^{1/2}} \int_{\mathbb{R}^n} e^{-|x-y|^2/(4(T-t))} \tilde{g}(y)dy.\]

**Remark 1.3.** The convolution is obviously well-defined if \(\tilde{g}\) has exponential growth (which covers the case of European calls) since (1.19) then applies.

**Remark 1.4.** We recall that a (scalar) Gaussian variable with expectation \(\mu\) and standard deviation \(\sigma\) has density
\[(1.23)\quad \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}.\]

We see that \(G(x,t)\) is a Gaussian variable with expectation \(\mu = 0\) and standard deviation \(\sigma = \sqrt{2t}\). So \(v(x,t)\) is the expectation of \(\tilde{g}\) with the measure of probability whose density is \(G(x,T-t)\). It follows that we may obtain an estimate of \(v(x,t)\) by the Monte-Carlo method:
\[(1.24)\quad v(x,t) \approx \frac{1}{N} \sum_{i=1}^{N} \tilde{g}(y_i),\]

where the \(y_i\) are identically independent random variables with density \(G(x,T-t)\). About Monte-Carlo methods we refer to [15].

**Remark 1.5.** There are various cases where the solution of the heat equation (1.14) can be computed explicitly. In particular, when for the European call, the solution is given by the Black and Scholes formula [8], whose proof is precisely based on the reduction to the heat equation. See also e.g. Lamberton and Lapeyre [39].
2. Schemes for diffusions: one space variable

2.1. Orientation. Instead of (1.8) we will deal with the slightly more general format\(^1\) corresponds to the

\[
\begin{align*}
V_t + b(x,t)V_x + \frac{1}{2} a(x,t) V_{xx} - r(x,t) V + f(x,t) &= 0, \quad (x,t) \in \mathbb{R} \times [0,T], \\
V(x,T) &= g(x), \quad x \in \mathbb{R}.
\end{align*}
\]

We will call \(f\) the source term. In financial models it may represent some revenue (dividends). We assume for the moment that the functions \(a, b\) and \(r\) are continuous and bounded, and that \(a\) is nonnegative. We will be more specific on \(f\) and \(g\) later. We start by the simple case of diffusions, in which \(b\) and \(r\) vanish, and discuss explicit, and then implicit schemes. We emphasize the property of simple monotonicity of these schemes. Finally, we consider the more general \(\theta\) scheme for which simple monotonicity most often does not hold, but general energy estimates may be provided.

We then turn our attention to equations with a first-order term, starting with the case of transport equations, and then considering the general case.

We end with the discussion of consistency, and provide a method for studying the convergence in the case when the coefficient \(a\) is constant.

2.2. Explicit schemes. We start with the case of pure diffusion, in which \(b\) and \(r\) are identically zero, so that (1.1) reduces to

\[
\begin{align*}
V_t + \frac{1}{2} a(x,t) V_{xx} + f(x,t) &= 0, \quad (x,t) \in \mathbb{R} \times [0,T], \\
V(x,T) &= g(x), \quad x \in \mathbb{R}.
\end{align*}
\]

Let the time step be \(h_0 = T/N\), where \(N\) is a positive integer, and the space step be \(h_1 > 0\). By \(v_j^k\), where \(j \in \mathbb{Z}\) and \(k = 0\) to \(N\), we denote an approximation of \(V(jh_1,kh_0)\).

The standard explicit FD scheme is

\[
\begin{align*}
&\left\{ \begin{array}{l}
\frac{v_j^k - v_j^{k-1}}{h_0} + \frac{1}{2} v_j^k \frac{v_{j+1}^k v_{j-1}^k - 2v_j^k}{h_1^2} + f_j^k = 0, \quad j \in \mathbb{Z}, \quad k = 1 : N, \\
v_j^N = g_j, \quad j \in \mathbb{Z},
\end{array} \right.
\end{align*}
\]

where

\[
\begin{align*}
a_j^k := a(jh_1,kh_0), \quad f_j^k := f(jh_1,kh_0), \quad g_j := g(jh_1).
\end{align*}
\]

The ordered form (w.r.t. to the components of \(v^k\)) of the first relation in (1.27) is

\[
\begin{align*}
v_j^{k-1} = \left(1 - \frac{h_0}{h_1^2} a_j^k\right) v_j^k + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k (v_{j-1}^k + v_{j+1}^k) + h_0 f_j^k, \quad j \in \mathbb{Z}, \quad k = 1 : N.
\end{align*}
\]

The r.h.s. of (1.29) is, apart from the source term \(h_0 f_j^k\), a linear combination of the values at the previous step. The coefficients have a sum equal to 1. They are nonnegative if the following monotonicity condition holds:

\[
\frac{h_0}{h_1^2} \|a\|_\infty \leq 1.
\]

\(^1\)We adopt here the notation "\(x\)" for the space variable for a PDE with bounded coefficients, instead of "\(z\)" in the previous section.
1. FIRST STEPS IN FINITE DIFFERENCES METHODS

Figure 1. Explicit scheme: maximal time step, $t = 0$ and 0.95.

In that case, $v^{k-1}$ is a nondecreasing function of $v^k$ and $f^k$: we say that the scheme is simply monotonic.

Example 1.6. Let $f(x,t) = 0$ and $a(x,t) = 1$, so that we recover the heat equation, and take $h_0$ equal either to the maximal value $h_1^2$, or to the half of it. Then the scheme reduces to

\[
\begin{align*}
    v_j^{k-1} &= \frac{1}{2} (v_{j+1}^k + v_{j-1}^k) & \text{if } h_0 = h_1^2, \\
    v_j^{k-1} &= \frac{1}{2} v_j^k + \frac{1}{4} (v_{j+1}^k + v_{j-1}^k) & \text{if } h_0 = \frac{1}{2} h_1^2.
\end{align*}
\]

Observe that, if the final condition is such that $v^N_j = (-1)^j$, then in the first case, $v^{k-1} = -v^k$: no damping occurs, while in the case of an oscillating final condition, the solution of the heat equation rapidly vanishes. More generally, the maximal time step does not damp rapidly enough high frequencies.

Numerical illustrations. We illustrate the numerical properties when $a(x,t) = 1$ and $T = 1$, for a digital option $g(x) = 1_{R_+}(x)$. We solve on a spatial domain $\Omega = [-1,1]$, with boundary conditions

\[
\begin{align*}
    V(-1,t) &= 0; & V(1,t) &= 1, & t \in [0,T].
\end{align*}
\]

Such a domain is obviously too small, but is quite convenient in order to illustrate the possible instabilities of the scheme. We take 39 space steps, so that $h_1 = 2/39$, and display the values computed for $t = 0.95$ and $t = 0$. On figure 1 we have taken the maximal time step $h_0 = h_1^2$, and observe significant instabilities. When $h_0 = \frac{1}{2} h_1^2$, these instabilities disappear as can be seen in figure 2. What happens if the monotonicity condition (1.30) does not hold? When $h_0 = 1.05 h_1^2$ we observe significant instabilities for $t = 0.95$ in figure 3 and a numerical explosion of the solution computed for $t = 0$ in figure 4.

2.3. Undiscounted Markovian schemes. We will see later various other algorithms that, like the one above, may be put in the following general framework. We
introduce the *Markovian scheme*

\begin{equation}
\begin{aligned}
\begin{cases}
  v^{k-1}_j = \sum_{\ell} \left( \alpha_{j,\ell}^k v^k_\ell + h_0 \hat{\alpha}_{j,\ell}^k f^k_\ell \right), & j \in \mathbb{Z}, \quad k = 1 : N, \\
  v^N_j = g_j, & j \in \mathbb{Z},
\end{cases}
\end{aligned}
\tag{1.33}
\end{equation}

where the coefficients $\alpha_{j,\ell}^k$ and $\hat{\alpha}_{j,\ell}^k$ satisfy

\begin{equation}
\begin{aligned}
\alpha_{j,\ell}^k \geq 0; \quad \hat{\alpha}_{j,\ell}^k \geq 0; \quad \sum_{\ell} \alpha_{j,\ell}^k = \sum_{\ell} \hat{\alpha}_{j,\ell}^k = 1; \quad \text{for all } j \in \mathbb{Z}, k = 0, \ldots, N - 1.
\end{aligned}
\tag{1.34}
\end{equation}

In the case of the scheme \(1.29\), we have that

\begin{equation}
\begin{aligned}
\alpha_{j,\ell}^k = 0 \text{ if } |j - \ell| > 1 \text{ and } \hat{\alpha}_{j,\ell}^k = \delta_{j,\ell},
\end{aligned}
\tag{1.35}
\end{equation}
1. FIRST STEPS IN FINITE DIFFERENCES METHODS

Figure 4. Explicit scheme: 1.05 times maximal time step, $t = 0$.

and so, the above sums have finitely many nonzero terms. However in the study of implicit algorithms we will need the general case. We note that condition (1.34) is an infinite dimensional extensions of the notion of stochastic matrix.

**Remark 1.7.** When the coefficients of the equation of the PDE (1.26) are constant, the resulting abstract form (1.33) typically satisfies $\alpha_{j,\ell}^k = \hat{\alpha}_{j-\ell}$ for some nonnegative coefficients $\hat{\alpha}_j$ of finite sum (as is the case for the explicit scheme (1.27) if (1.30) holds). Define the discrete convolution for function of $\mathbb{Z}$ as

(1.36) \[ (\alpha * \beta)_i := \sum_{k \in \mathbb{Z}} \alpha_{i-k} \beta_k, \quad i \in \mathbb{Z}. \]

When $f = 0$, we may then write (1.33) in the form of a convolution product:

(1.37) \[ v^{k-1} = \hat{\alpha} * v^k = \hat{\alpha} * \cdots * v^N \quad \text{(convolution $N - k + 1$ times),} \quad k = 1 : N. \]

Although in applications the solution is typically unbounded, we start with estimates in the space $\ell^\infty$ of bounded “sequences” $x = (x_k), k \in \mathbb{Z}$ endowed with the supremum norm

(1.38) \[ \|x\|_\infty := \sup_k |x_k|. \]

**Lemma 1.8.** Let (1.34) holds. Then the Markovian scheme (1.33) is simply monotonic. It $g$ and $f$ are bounded, it has a uniformly bounded solution, and we have

\[
\begin{align*}
(\text{i}) & \quad \sup_j v_{j}^{k-1} \leq \sup_j v_{j}^{k} + h_0 \sup_j f_{j}^{k} \leq \sup g + (N - k + 1)h_0 \sup f, \\
(\text{ii}) & \quad \inf_j v_{j}^{k-1} \geq \inf_j v_{j}^{k} + h_0 \inf_j f_{j}^{k} \geq \inf g + (N - k + 1)h_0 \inf f, \\
(\text{iii}) & \quad \|v^{k-1}\|_\infty \leq \|v^k\|_\infty + h_0 \|f^k\|_\infty \leq \|g\|_\infty + T\|f\|_\infty.
\end{align*}
\]

\footnote{A stochastic matrix is a square matrix whose nonnegative coefficients sum to 1 on each row. If columns also sum to 1, we say that the matrix is bistochastic. The matrix of probability transitions of a finite Markov chain is a stochastic matrix.}
Proof. In view of (1.34), we have that
\begin{align}
v_j^{k-1} &\leq \sup_{\ell} v^{k}_{\ell} + h_0 \sup_{\ell} f^{k}_{\ell}.
\end{align}

Take the supremum of the l.h.s. over \(j\), we obtain the first inequality in (i). Summing these inequalities from \(k\) to \(N - 1\), we obtain the second inequality in (i). Relation (ii) is of the same nature, and (iii) follows from (i)-(ii).

### 2.4. Implicit schemes

The standard implicit scheme for the pure diffusion equation (1.26) is
\begin{align}
\begin{cases}
\frac{v_j^{k+1} - v_j^k}{h_0} + \frac{1}{2} a_j^k v_{j+1}^{k+1} + v_{j-1}^k - 2v_j^k + f_j^k = 0, & j \in \mathbb{Z}, \; k = 0, \ldots, N - 1, \\
v_j^{0} = g_j, & j \in \mathbb{Z}.
\end{cases}
\end{align}

The first relation in (1.41) may be written in the fixed-point form
\begin{align}
v_j^k = \left(1 + \frac{h_0}{h_1^2} a_j^k \right)^{-1} \left(\frac{h_0}{h_1} a_j^k (v_{j-1}^k + v_{j+1}^k) + v_j^{k+1} + h_0 f_j^k\right), & j \in \mathbb{Z}, \; k = 0 : N - 1.
\end{align}

We set
\begin{align}
\gamma_j^k := \frac{h_0}{h_1} a_j^k \left(1 + \frac{h_0}{h_1} a_j^k \right)^{-1}, & \gamma^k := \sup_j \gamma_j^k.
\end{align}

Since \(s \mapsto s/(1 + s)\) is increasing with image in \((-1, 1)\), we have that
\begin{align}
\gamma^k \leq \gamma(a, h) := \frac{h_0}{h_1^2} \|a\|_{\infty} \left(1 + \frac{h_0}{h_1^2} \|a\|_{\infty}\right)^{-1} < 1.
\end{align}

Let us denote by \(T\) the fixed point operator from \(\ell^\infty\) into itself, corresponding to the mapping in the r.h.s. of (1.42), with argument \(v^k\) (here \(v^{k+1}\) is already computed and hence is “fixed”), i.e.
\begin{align}
(Tw)_j := \left(1 + \frac{h_0}{h_1^2} a_j^k \right)^{-1} \left(\frac{h_0}{h_1} a_j^k (w_{j-1} + w_{j+1}) + v_j^{k+1} + h_0 f_j^k\right), & j \in \mathbb{Z}.
\end{align}

**Definition 1.9.** We denote by \(\mathcal{B}(X)\) the Banach space of bounded functions over a set \(X\), endowed with the uniform norm \(\|f\|_{\infty} := \sup_{x \in X} |f(x)|\). If \(X\) is a countable set we denote this space as \(\ell^\infty(X)\), or \(\ell^\infty\) if this is not confusing, and the norm as \(\|\cdot\|_{\infty}\).

The above norm should not be confused with the \(L^\infty\) norm over measured spaces. There will be no ambiguity about that in these notes.

**Lemma 1.10.** Let \(f\) and \(g\) belong to \(\mathcal{B} (\mathbb{R}^n)\). Then: (i) the scheme (1.41) has a unique solution \(v(f, g)\) in \(\ell^\infty\), (ii) the mapping \((f, g) \mapsto v(f, g)\) is simply monotonic, and (iii) we have that
\begin{align}
\|v^k\|_{\infty} \leq \|g\|_{\infty} + T \|f\|_{\infty}.
\end{align}

**Proof.** (i) By (1.44), we have that
\begin{align}
\|(Tw')_j - (Tw)_j\| \leq \frac{1}{2} \gamma_j^k (|w_{j+1}' - w_j| + |w_{j-1}' - w_j|) \leq \gamma(a, h) \|w' - w\|_{\infty},
\end{align}
proving that $T$ has contraction factor $\gamma(a, h)$, and so has a unique fixed point, as was to be proved.

(ii) Since the scheme is linear, its simple monotonicity follows from the fact that the solution is nonnegative if $(f, g)$ is. We prove the latter by backward induction over $k$. This reduces to the proof of nonnegativity of $v^k$, knowing that $v^{k+1}$ and $f$ are nonnegative. In this case consider the fixed-point iteration starting from the initial guess $v_{0}^{0} = 0$. Then

$$v_{1}^{k} = \left( 1 + \frac{h_0}{h_1^2} a_{j}^k \right)^{-1} \left( v_{1}^{k+1} + h_0 f_{j}^k \right) \geq 0 = v_{0}^{k}.$$ 

Since the operator $T$ is obviously nondecreasing it follows that the sequence $(v_{1}^{k})$ is also nondecreasing, and therefore its limit $v_{1}^{k}$ is nonnegative as was to be checked.

(iii) Let us denote by $v_{(f, g)}$ the solution of the scheme. Since the latter is simply monotonic, we have that

$$v\left(-\|f\|_{\infty}, -\|g\|_{\infty}\right) \leq v((f, g) \leq v((\|f\|_{\infty}, \|g\|_{\infty}),$$

where we identify a constant with the constant function having the same value. So it suffices to check that the result holds when $(f, g)$ are constant. But then we easily see that the solution of the scheme is $v_{1}^{k} = g + h_0 (N - k)f$. The conclusion follows. □

2.4.1. Practical use of the fixed point iteration. Denoting as in the above proof $v_{1}^{k,i} := T_{1}^{0}$ the fixed-point sequence initialized by zero, since we have a contraction factor $\gamma = \gamma(a, h)$, since $\|v_{1}^{k+1} + h_0 f_{j}^k\|_{\infty} \leq M := \|g\|_{\infty} + T \|f\|_{\infty}$, we have that

$$\|v_{1}^{k,i} - v_{1}^{k}\|_{\infty} \leq \gamma_{i} M.$$

Actually it is wiser to initialize the sequence $v_{1}^{k,i}$ with $v_{1}^{k+1}$. Assuming that we have enough regularity so that $\|v_{1}^{k+1} - v_{1}^{k}\|_{\infty} \leq c_1 h_0$, we will then obtain

$$\|v_{1}^{k,i} - v_{1}^{k}\|_{\infty} \leq c_1 \gamma_{i} h_0.$$ 

Since we have a first order approximation of the time derivative in the scheme, we can expect errors of at least $O(h_0^2)$ and so we may stop the algorithm when $\gamma_{i} h_0 \leq c_2 h_0^2$, i.e., when

$$i \log(\gamma) \leq \log h_0 + O(1).$$

If we choose $h_0$ of the order of $h_1^2$, for instance, such that $h_0 \|a\|_{\infty}/h_1^2 \leq 1$, then $\gamma \leq 1/2$ and we may stop when

$$i \geq \frac{1}{\log 2} \log \left( \frac{1}{h_0} \right) + O(1),$$

which in practice will be a small number since we cannot take $h_0$ too small.

However the motivation for implicit schemes is precisely to take much larger time steps. If we take $h_0$ such that $h_0 \|a\|_{\infty} >> h_1^2$, then

$$\gamma = \frac{1}{1 + \frac{1}{h_0^2 \|a\|_{\infty}}} \approx 1 - \frac{h_1}{h_0 \|a\|_{\infty}}$$

will be close to 1 and the convergence will be very slow. So, using the fixed point operator for solving the implicit scheme is not a good idea. It is more efficient to solve the tridiagonal linear system with an appropriate linear solver.
2.4.2. Link with Markovian schemes. The fixed-point form (1.42) is a particular case of the more general implicit scheme

\[(1.55) \quad v_j^k = \left(1 + \sum_{\ell} \bar{\alpha}_{j,\ell}^k \right)^{-1} \left( \sum_{\ell} \bar{\alpha}_{j,\ell}^k v_{\ell}^k + v_{j+1}^k + h_0 f_j^k \right),\] 

with \(\bar{\alpha}_{j,\ell}^k \geq 0, \sum_{\ell} \alpha_{j,\ell}^k < \infty\).

**Lemma 1.11.** The general implicit scheme (1.55) belongs to the Markovian class (1.33), with here \(\alpha_{j,\ell}^k = \alpha_{j,\ell}^k\), for all \(j \text{ and } \ell \in \mathbb{Z}\).

**Proof.** Set \(\bar{v}_j^k := v_{j+1}^k + h_0 f_j^k\). We must prove that \(v_j^k = \sum_{\ell} \alpha_{j,\ell}^k \bar{v}_{\ell}^k\), for some nonnegative coefficients \(\alpha_{j,\ell}^k\) with unit sum. Since \(\bar{v}\) is the sum of its positive and negative parts, it suffices to discuss the case when \(\bar{v}_j^k \geq 0\) for all \(j \in \mathbb{Z}\). Let again \(v^{k,i} := T^i 0\) denote the fixed-point sequence initialized by zero. We prove by induction that, for \(i \geq 1\):

\[(1.56) \quad v_j^{k,i} = \sum_{\ell} \alpha_{j,\ell}^{k,i} \bar{v}_{\ell}^k, \quad \text{for some } \alpha_{j,\ell}^{k,i} \geq 0 \text{ such that } A_i := \sup_{j} \sum_{\ell} \alpha_{j,\ell}^{k,i} \leq 1.\]

For \(i = 1\) this holds with \(\alpha_{j,\ell}^{k,1} := \left(1 + \sum_{\ell} \bar{\alpha}_{j,\ell}^k \right)^{-1} \delta_{j,\ell}\) and so \(A_1 \leq 1\), so that for \(i > 1\):

\[(1.57) \quad v_j^{k,i+1} = \left(1 + \sum_{\ell} \bar{\alpha}_{j,\ell}^k \right)^{-1} \left( \sum_{q} \bar{\alpha}_{j,q}^k \sum_{\ell} \alpha_{j,\ell}^{k,i} \bar{v}_{\ell}^k + \bar{v}_j^k \right),\]

This implies (1.56) since

\[(1.58) \quad \alpha_{j,\ell}^{k,i+1} = \left(1 + \sum_{\ell} \bar{\alpha}_{j,\ell}^k \right)^{-1} \left( \delta_{j,\ell} + \sum_{q} \bar{\alpha}_{j,q}^k \alpha_{j,\ell}^{k,i} \right),\]

and

\[(1.59) \quad \sum_{\ell} \alpha_{j,\ell}^{k,i+1} \leq \left(1 + \sum_{\ell} \bar{\alpha}_{j,\ell}^k \right)^{-1} \left(1 + \sup_{j} \sum_{q} \bar{\alpha}_{j,q}^k A_i \right) \leq 1.\]

We may see the sequence \(\alpha_{j,\ell}^{k,i+1}\) as initialized for \(i = 0\) with zero values. We have that \(\alpha_{j,\ell}^{k,1} \geq \alpha_{j,\ell}^{k,0}\) and since the mapping in (1.58) that with \(\alpha_{j,\ell}^{k,i}\) associates \(\alpha_{j,\ell}^{k,i+1}\) is nondecreasing, it follows that \(i \mapsto \alpha_{j,\ell}^{k,i}\) is nondecreasing, and is less than one in view of (1.56), and so, converges to some \(\alpha_{j,\ell}^k\). By the monotone convergence theorem (in the space \(\ell^1(\mathbb{Z})\) of summable sequences over \(\mathbb{Z}\)), \(\sum_{\ell} \alpha_{j,\ell}^k = \limsup A_i \leq 1\), and also \(v_j^{k,i} \rightarrow \sum_{\ell} \alpha_{j,\ell}^k \bar{v}_{\ell}^k\), for all \(j \in \mathbb{Z}\). At the same time, by (1.50), \(v^{k,i} \rightarrow v^k\) uniformly, and so \(v_j^k = \sum_{\ell} \alpha_{j,\ell}^k \bar{v}_{\ell}^k\). When \(\bar{v}_\ell^k = 1\) for all \(\ell\), we easily check that \(v_j^{k,-1} = 1\), and so, \(\sum_{\ell} \alpha_{j,\ell}^k = 1\). The result follows. \(\Box\)

**Remark 1.12.** If the diffusion coefficient \(a\) is unbounded, then the operator \(T\) defined by (1.45) is still well-defined from \(L^2_\gamma\) into itself, and is simply monotonic. However, the well-posedness of the scheme does not follow any more from contraction arguments since now \(\gamma(a, h) = 1\), and therefore \(T\) is no more contracting.

We observe the behavior of the implicit scheme with the same data as for the illustration of the explicit scheme. We see on figure 5 how smooth is the obtained solution, even for a time step much larger that the maximal time step of the explicit scheme. Of course,
1. FIRST STEPS IN FINITE DIFFERENCES METHODS

Figure 5. Implicit scheme: 5 times maximal explicit time step, \( t = 0 \) and 0.95.

for a very large time step the numerical errors will be important, and hence, having a smooth curve is not a guarantee of precision!

2.5. Diffusions with actualization. When the interest rate \( r \) is nonzero, we can formulate finite difference schemes similar to the ones stated above for the case when \( r = 0 \). They enter in an abstract framework which we may call discounted Markovian. We detail the case of explicit schemes, and consider here equations of the form

\[
v_t + \frac{1}{2}a(x,t)v_{xx} - r(x,t)v + f(x,t) = 0, \quad (x,t) \in \mathbb{R} \times [0,T],
\]

\[
v(x,t) = g(x), \quad x \in \mathbb{R}.
\]

It is convenient to write a variant of the scheme (1.27) in which the contribution of the actualization term is implicit rather than explicit, i.e.,

\[
\begin{align*}
& v_j^k - v_j^{k-1} - \frac{1}{2}a_j^k v_{jj+1}^k + v_{jj-1}^k - 2v_j^k - r_j^{k-1}v_j^{k-1} + f_j^k = 0, \quad j \in \mathbb{Z}, \ k = 1 : N, \\
& v_N^1 = g_j, \quad j \in \mathbb{Z}.
\end{align*}
\]

where \( r_j^k := r(jh_1, kh_0) \). The first relation in (1.61) may be written in the ordered form

\[
\begin{align*}
& v_j^{k-1} = (1 + h_0 r_j^{k-1})^{-1} \left[ \left( 1 - \frac{h_0}{h_1^2} a_j^k \right) v_j^k + \frac{h_0}{2} h_1^2 a_j^k \left( v_{jj-1}^k + v_{jj+1}^k \right) \right], \\
& j \in \mathbb{Z}, \ k = 1 : N,
\end{align*}
\]

where the bracket on the r.h.s. is the expression of the explicit scheme when \( r = 0 \).

Lemma 1.13. Assume that \( r \geq 0 \). If the monotonicity relation (1.30) holds, then

\[
\|v\|_\infty \leq (\|g\|_\infty + T\|f\|_\infty).
\]

The result follows from corollary 1.15 which involves a general framework introduced below.
2. SCHEMES FOR DIFFUSIONS: ONE SPACE VARIABLE

2.5.1. **Discounted Markovian schemes.** Assuming the monotonicity relation (1.30) to hold, we see that the scheme (1.62) is a particular case of the discounted Markovian setting (extension of (1.33)) below:

\[
\begin{aligned}
  v_j^{k-1} &= \beta_j^k \left( \sum_\ell \alpha_{j,\ell}^k v_\ell^k + h_0 \sum_\ell \hat{\alpha}_{j,\ell}^k f_\ell^k \right), \\
  v_N^k &= g_j,
\end{aligned}
\]

where the coefficients \(\alpha_{j,\ell}^k\) and \(\hat{\alpha}_{j,\ell}^k\) still satisfy (1.34), \(\beta_j^k\) is nonnegative and

\[
\bar{\beta} := \sup \left\{ \beta_j^k, \quad k = 1, \ldots, N, \ j \in \mathbb{Z} \right\} < \infty.
\]

(1.65) (Note that this holds for the scheme (1.62), with \(\beta_j^k := (1 + h_0 v_j^{k-1})^{-1}\).) Set

**Lemma 1.14.** We have that, \(\bar{\beta}\) being defined by (1.65):

\[
\begin{aligned}
  (i) \quad (\sup_j v_j^{k-1})_+ &\leq \bar{\beta}^{N-k+1}(\sup_j g_j)_+ + h_0 \sum_{\ell=k}^{N} \bar{\beta}^{\ell-k+1}(\sup_j f_\ell^j)_+ , \\
  (ii) \quad (\inf_j v_j^{k-1})_- &\geq \bar{\beta}^{N-k+1}(\inf_j g_j)_- + h_0 \sum_{\ell=0}^{k} \bar{\beta}^{\ell-k+1}(\inf_j f_\ell^j)_-, \\
\end{aligned}
\]

and also

\[
\| v^{k-1} \|_{\infty} \leq \bar{\beta}^{N-k+1} \| g \|_{\infty} + h_0 \sum_{\ell=k}^{N} \bar{\beta}^{\ell-k+1} \| f_\ell \|_{\infty},
\]

(1.67)

**Proof.** It follows from (1.64) that

\[
(v_j^{k-1})_+ \leq \bar{\beta} \left( \sup_i (v_i^k)_+ + h_0 \sup_i (f_i^k)_+ \right).
\]

Taking the supremum over \(j\), we deduce that

\[
\sup_i (v_i^{k-1})_+ \leq \bar{\beta} \left( \sup_i (v_i^k)_+ + h_0 \sup_i (f_i^k)_+ \right).
\]

from which (1.66)(i) easily follows. Inequality (1.66)(ii) is obtained in the same way, and these two relations imply (1.67).

**Corollary 1.15.** If \(\bar{\beta} \leq 1 - c_2 h_0, \) with \(c_2 \geq 0,\) then for \(k = 0\) to \(N:\)

\[
\| v^k \|_{\infty} \leq e^{-c_2(T-t_k)} \| g \|_{\infty} + (T-t_k) \| f \|_{\infty}.
\]

**Proof.** Set \(k' = N - k + 1.\) Then \(k' h_0 = T - t_{k-1}.\) Since \(\log(1+x) \leq x,\) we have that

\[
\bar{\beta}^{k'} = e^{k' \log(1-c_2 h_0)} \leq e^{-c_2 k' h_0} = e^{-c_2(T-t_{k-1})}.
\]

We obtain then with lemma 1.14 the desired estimate for \(k - 1,\) when \(k = 1\) to \(N.\) The result follows.

2.6. **Markov chain interpretation.** It is interesting to observe that, while the PDE to be solved can be interpreted as expectations of integral plus terminal values associated with certain stochastic process, the finite difference schemes that we have presented can also be interpreted as a computation of expectations of integral plus terminal terms, associated with a Markov chains systems.
Let us associate with the Markov setting \([1.33]\) the following “Markov chain” model with countable state space \(Z\) and time set \(\{0, \ldots, N\}\). The Markov process \(X^k, k = 0\) to \(N\), is such that if \(X^{k-1} = j\), then the probability that \(X^k = \ell\) is \(\alpha_{j,\ell}\), i.e.,

\[
\mathbb{P}[X^k = \ell | X^{k-1} = j] = \alpha_{j,\ell}^k.
\]

Let \(\mathcal{F}_k\) be the \(\sigma\)-field generated by \(X^1, \ldots, X^k\). We denote by \(\mathbb{E}_k(\cdot) := \mathbb{E}[\cdot | \mathcal{F}_k]\) the conditional expectations associated with this filtration; we have that \(\mathbb{E}_{k-1} = \mathbb{E}_{k-1} \circ \mathbb{E}_k\).

Set \(\hat{X}_{IE}\) conditional expectations associated with this filtration; we have that \(\hat{Z}\) with countable state space \(\{0, \ldots, N\}\). Let \(\hat{F}_{j} := \sum_\ell \alpha_{j,\ell}^k f_{j,\ell}^k\); With a trajectory \(X = X^0, X^k, \ldots, X^N\), we associate the reward function starting at time \(k - 1\), for \(k = 1\) to \(N\):

\[
Y_{X}^{k-1} := h_0 \hat{f}_{X}^k + \cdots + h_0 \hat{f}_{X}^N + g(X^N) = h_0 \hat{f}_{X}^k + Y_{X}^{k+1}.
\]

The corresponding expectation function is

\[
V_{j}^{k-1} := \mathbb{E}[Y_{X}^{k-1} | X^{k-1} = j],
\]

or equivalently

\[
V_{j}^{k-1} := h_0 \hat{f}_{j}^k + \mathbb{E}[h_0 \hat{f}_{X}^{k+1} + \cdots + h_0 \hat{f}_{X}^N + g(X^N) | X^{k-1} = j].
\]

We obviously have

\[
V_{j}^{k-1} := h_0 \hat{f}_{j}^k + \mathbb{E}[Y_{X}^{k} | X^{k-1} = j], \quad k = 1, \ldots, N, \ j \in \mathbb{Z},
\]

and the final condition

\[
V_{j}^{N} = g(x_j), \quad j \in \mathbb{Z}.
\]

**Lemma 1.16.** The functions \(V_{j}^{k}\) coincide with the solution \(v_{j}^{k}\) of the Markovian scheme \((1.33)\).

**Proof.** By the Markov property, since \(Y_{X}^{k}\) depends only on \(X^k, \ldots, X^N\) we have that \(\mathbb{E}_k Y_{X}^{k} = \mathbb{E}[Y_{X}^{k} | X^k]\), and so,

\[
\mathbb{E}_{k-1} Y_{X}^{k} = \mathbb{E}_{k-1} \mathbb{E}_k Y_{X}^{k} = \mathbb{E}_{k-1} \mathbb{E}[Y_{X}^{k} | X^k] = \mathbb{E}[\mathbb{E}[Y_{X}^{k} | X^k] | X^{k-1}]
\]

and since \(\mathbb{P}[X^k = \ell | X^{k-1} = j] = \alpha_{j,\ell}^k\):

\[
\mathbb{E}[Y_{X}^{k} | X^{k-1} = j] = \mathbb{E}[\mathbb{E}[Y_{X}^{k} | X^k] | X^{k-1} = j] = \sum_\ell \alpha_{j,\ell}^k \mathbb{E}[Y_{X}^{k} | X^k = \ell]
\]

so that

\[
V_{j}^{k-1} = h_0 \hat{f}_{j}^k + \mathbb{E}[Y_{X}^{k} | X^{k-1} = j] = h_0 \hat{f}_{j}^k + \sum_\ell \alpha_{j,\ell}^k V_{j}^{k}.
\]

This is nothing that the recurrence relation of the Markovian scheme, and since the final condition is identical, the conclusion follows. \(\square\)

**Remark 1.17.** In the setting of example \([1.6]\), when we take the maximal time step, we see that the Markov chain corresponds to a variation of \(\pm h_1\), which is a standard discretization for the Brownian motion. It is interesting to note that oscillations were present in numerical results, in relation with poor damping properties; this suggests to discretize in a different way (e.g. corresponding to half of the maximal time step) the Brownian motion.
Problems with actualization. When the coefficient $r$ is non zero, and more generally when the FD scheme enters in the format (1.64), we consider the same Markov chain model associated with the following discounted expression of the cost:

\begin{equation}
V_j^k := \beta_j^k \left( h_0 f_j^k + \mathbb{E} \left[ \beta_{X_{k+1}}^k \left( h_0 f_{X_{k+1}}^k + \beta_{X_{k+2}}^k \left( h_0 f_{X_{k+2}}^k + \cdots \right) + h_0 \beta_{X_{N-1}}^k \left( f_{X_{N-1}}^N + \beta_{X_N}^N g(X^N) \right) \right) \right] \right) | X^{k-1} = j \right). \tag{1.81}
\end{equation}

Again, the inner conditional expectation appears to be the value of $V_{X_k}^{k+1}$, and we see that we recover the solution of (1.64).

Remark 1.18. It follows from the previous results that we can evaluate the solution of the simply monotonic explicit FD schemes by Monte-Carlo methods, in which we simulate the simply monotonic operators is simply monotonic, if the following condition holds,

\begin{equation}
\theta \in [0,1], \ 
\tag{1.82}
\begin{cases}
\frac{v_{j+1}^{k} - v_{j}^{k}}{h_0} + \frac{1}{2} \theta a_j^k v_{j+1}^{k} + v_{j-1}^{k} - 2v_j^k + \frac{1}{2} (1-\theta) a_j^k v_{j+1}^{k+1} + v_{j-1}^{k+1} - 2v_j^{k+1} + f_j^k = 0, \\
\text{for } j \in \mathbb{Z}, \ k = 0, \ldots, N-1 \\
v_j^N = g_j, \ j \in \mathbb{Z}.
\end{cases}
\end{equation}

The scheme is explicit if $\theta = 0$, and implicit otherwise; for $\theta = 1$ it reduces to the standard implicit scheme (1.41). It can be seen as the combination of an explicit step followed by an implicit step, both of length $h_0$. Indeed, it is equivalent to the decomposed form below (to see this, just add the two equations):

\begin{equation}
\begin{cases}
\frac{v_{j+1}^{k+1/2} - v_{j}^{k+1/2}}{h_0} + \frac{1}{2} (1-\theta) a_j^k v_{j+1}^{k+1/2} + v_{j-1}^{k+1/2} - 2v_j^{k+1/2} + f_j^k = 0, \\
\frac{v_{j+1/2}^{k} - v_{j}^{k}}{h_0} + \frac{1}{2} \theta a_j^k v_{j+1/2}^{k} + v_{j-1}^{k} - 2v_j^k = 0, \quad j \in \mathbb{Z}, \ k = 0, \ldots, N-1, \\
v_j^N = g_j, \ j \in \mathbb{Z}.
\end{cases}
\end{equation}

We know that the second (implicit) step is simply monotonic. Since a combination of simply monotonic operators is simply monotonic, if the following condition holds,

\begin{equation}
(1-\theta) h_0 \frac{h_0}{h_1^2} \|a\|_{\infty} \leq 1, \tag{1.84}
\end{equation}

then by the results of section 2.2 the explicit step is simply monotonic, and hence so is the $\theta$ scheme.

Remark 1.19. In general the $\theta$ scheme is not simply monotonic (except of course when $\theta = 1$) when $h_0$ is not of the order of $h_1^2$, and different from 1, see [13].

3. Schemes for the general equation, one space variable

3.1. Transport equations. In order to emphasize the qualitative aspects of equations involving first-order terms we start by analyzing the situation (seldom encountered...
in financial problems) when in the basic model the diffusion coefficient \( a \) is equal to 0. We obtain the hyperbolic first-order equation

\[
\begin{align*}
& V_t + b(x,t)V_x + f(x,t) = 0, \quad (x,t) \in \mathbb{R} \times [0,T], \\
& V(x,T) = g(x), \quad x \in \mathbb{R}.
\end{align*}
\]  

(1.85)

The simplest situation is when \( b \) is constant and \( f \) is equal to 0, i.e.,

\[
\begin{align*}
& V_t + bV_x = 0, \quad (x,t) \in \mathbb{R} \times [0,T], \\
& V(x,T) = g(x), \quad x \in \mathbb{R}.
\end{align*}
\]  

(1.86)

In that case the solution is known to be

\[
V(x,t) = g(x + b(T-t)).
\]  

(1.87)

We obtain a classical solution if \( g \) is of class \( C^1 \), and a generalized solution otherwise. The solution at time \( t \) is a translation of amount \( bt \) of the solution at time 0; we may call \( b \) the speed of propagation and interpret \( V(x,t) \) as a signal propagating at speed \( b \). The lines \( x + b(T - t) = x_T \) (constant) of the \((x,t)\) plane are called characteristics; the values are invariant along the characteristics. The latter can be parameterized in the form \( (x(t) = c - b(T-t) \), or equivalently \( x(T) = x_T \), \( \dot{x}(t) = b \).

**Remark 1.20.** More generally, when \( b \) depends on \((x,t)\), we may define the *characteristic curves* in the \((x,t)\) plane as the solutions of the ODE

\[
x(T) = x_T; \quad \dot{x}(t) = b(x(t), t).
\]  

(1.88)

Assuming, say, \( b \) to be Lipschitz and bounded, we have (by the Cauchy-Lipschitz theorem) that the characteristic curves are well-defined. In addition, if \( V \) has a solution of class \( C^1 \), then along a characteristic we have that

\[
\frac{d}{dt} V(x(t), t) = -f(x(t), t),
\]  

(1.89)

so that the computation of say \( V(x_0, 0) \) reduces to the integration of a family of ODEs.

**Remark 1.21.** When the speed \( b \) depend also on \( V \), the characteristic curves may cross, and in this case the solution of the transport equation has to be understood using some elaborated concepts of solutions, see John [33]. Fortunately, this situation does not occur in our setting.

This theory suggests that a numerical scheme should (approximately) follow the values at previous stages of integration related to the characteristic curve. In particular, when \( b(x,t) \geq 0 \), it is sensible to use larger values of \( x \) (since we integrate backwards). So a sensible choice is to approximate the first-order derivative by a right (left) first-order difference of the propagation speed \( b \) is nonnegative (nonpositive). Setting \( b^k_j := b(x^j, t_k) \) and \( f^k_j := f(x^j, t_k) \), we may write an *explicit discretization upwind* scheme\(^3\) as follows:

\[
\begin{align*}
& \frac{v^k_j - v^k_{j-1}}{h_0} + (b^k_j)_+ \frac{v^k_{j+1} - v^k_j}{h_1} + (b^k_j)_- \frac{v^k_j - v^k_{j-1}}{h_1} + f^k_j = 0, \quad j \in \mathbb{Z}, \quad k = 1 : N,
\end{align*}
\]  

(1.90)

\(^3\) In the setting of a forward equation, i.e., for a given initial condition at time \( t = 0 \), the upwind scheme consists of course in the opposite rule, for instance, take the left FD if \( b \geq 0 \), with the following interpretation: look at what is coming to you.
or in ordered form
\[
\begin{cases}
  v_j^{k-1} = \left(1 - \frac{h_0}{h_1} |b_j^k| \right) v_j^k + \frac{h_0}{h_1} \left( (b_j^k)_+ v_{j+1}^k - (b_j^k)_- v_{j-1}^k \right) + h_0 f_j^k, & j \in \mathbb{Z}, \; k = 1 : N, \\
  v_N^0 = g_j, & j \in \mathbb{Z}.
\end{cases}
\]

We see that this scheme is simply monotonic whenever
\[
\frac{h_0}{h_1} \|b\|_\infty \leq 1.
\]

This is the famous CFL (Courant-Friedrichs-Lewy) condition \cite{19, 20}. Its physical interpretation is that the speed of propagation $h_1/h_0$ of the numerical scheme must not be less than the physical one. Observe that in the case of the diffusion with explicit scheme, we had the much more restrictive condition (1.30) for simple monotonicity, that implies $h_0 = O(h_1)^2$.

**Exercice 1.22.** Assume that $b(x, t) = 1$ and $f(x, t) = 0$ for all $(x, t)$, when $h_0 = h_1$.
(i) Check that the scheme reduces to the “shift” operator $v_j^{k-1} = v_{j+1}^k$.
(ii) Give an interpretation in term of numerical integration along the characteristic curves.

**Exercice 1.23.** Assume again that $b(x, t) = 1$, $f(x, t) = 0$ for all $(x, t)$, and $h_0 = \frac{1}{2}h_1$.
Check that the scheme reduces to $v_j^{k-1} = \frac{1}{2}(v_j^k + v_{j+1}^k)$ and display the numerical results.

**Exercice 1.24.** Show that the explicit scheme with centered finite difference in space:
\[
\begin{cases}
  v_j^k - v_{j}^{k-1} + b_j^k v_{j+1}^k - v_{j-1}^k + f_j^k = 0, & j \in \mathbb{Z}, \; k = 1 : N, \\
  v_N^0 = g_j, & j \in \mathbb{Z},
\end{cases}
\]
leads to a non simply monotonic scheme. Confirm the expected bad numerical behavior on a simple example.
1. FIRST STEPS IN FINITE DIFFERENCES METHODS

Figure 7. Transport equation: $CFL = \frac{1}{2}$, $h_0 = 1/80, t = 0$.

![Figure 7](image1)

Figure 8. Transport equation: $CFL = 1.02$, $t = 0.95$: growing instability.

![Figure 8](image2)

Example 1.25. Consider the case when $b = 1$ and $CFL = \frac{1}{2}$, over the domain $\Omega = [0, 20]$, with $g(x) := 1_{x \geq 10}(x)$, $T = 10$, $h_1 = 1/20$, $h_0 = 1/40$. The exact solution is equal to $1_{x \geq t}(x)$. We display on figure 6 the value at time $t = 0$, for $x \in [0, 2]$. Observe the important “numerical diffusion”, showing that (for that given $CFL$) the space and time steps should be much smaller. For $h = 1/80$ the result is displayed on figure 7. It is clearly more accurate, but still far from the exact solution. Finally on figure 8 we display a case when the $CFL$ is slightly higher than 1. After only 0.05 time units we see an important instability that is starting; for $t = 0$ we obtain values of order $\pm 10^5$.

3.2. Explicit schemes. We next present two explicit schemes for discretizing the general equation (1.25), combining the ideas of section 2.2 with the above material. We assume for the sake of simplicity that $r = 0$. 
3. SCHEMES FOR THE GENERAL EQUATION, ONE SPACE VARIABLE

3.2.1. Upwind scheme. The scheme is as follows

\[
\begin{cases}
\frac{v_j^k - v_j^{k-1}}{h_0} + (b_j^k) + \frac{v_j^{k+1} - v_j^k}{h_1} + \frac{v_j^k - v_j^{k-1}}{h_1} \\
+ \frac{1}{2} a_j v_j^{k+1} + v_j^{k-1} - 2 v_j^k + \frac{1}{h_1^2} f_j^k = 0, \\
\end{cases}
\]

or in ordered form

\[
\begin{cases}
v_{j}^{k-1} = \left(1 - \frac{h_0}{h_1} |b_j^k| - \frac{h_0}{h_1^2} a_j^k\right) v_j^k + \left(\frac{h_0}{h_1} (b_j^k) + \frac{1}{2} h_0 a_j^k \right) v_{j+1}^k \\
+ \left(-\frac{h_0}{h_1} (b_j^k) - \frac{1}{2} h_0 a_j^k \right) v_{j-1}^k + h_0 f_j^k, \\
v_j^N = g_j, \\
\end{cases}
\]

and we see that this scheme is simply monotonic if

\[
h_0 \left(1 - \frac{1}{h_1} \|b\|_\infty + \frac{1}{h_1^2} \|a\|_\infty \right) \leq 1.
\]

When \(a \neq 0\) and \(h_1\) is small enough we recover a condition of the type \(h_0 = O(h_1^2)\).

**Lemma 1.26.** Provided that (1.96) holds, the scheme is well-defined, simply monotonic, and its unique solution satisfies

\[
\|v\|_\infty \leq \|g\|_\infty + T \|f\|_\infty.
\]

**Proof.** As in the case of diffusion equations, we can prove that the above scheme enters in the Markovian framework. We conclude with lemma 1.8(iii). \( \square \)

3.2.2. Explicit centered schemes. It may happen that the simple monotonicity of the second order term compensates the non simple monotonicity of the first order one. Indeed, let us use the centered first order finite difference. The scheme is then

\[
\begin{cases}
\frac{v_j^k - v_j^{k-1}}{h_0} + \frac{h_0}{h_1} (b_j^k) v_j^k + \frac{1}{2} a_j v_j^{k+1} + v_j^{k-1} - 2 v_j^k + \frac{1}{h_1^2} f_j^k = 0, \\
\end{cases}
\]

The ordered form is

\[
\begin{cases}
v_{j}^{k-1} = \left(1 - \frac{h_0}{h_1} |b_j^k| - \frac{h_0}{h_1^2} a_j^k\right) v_j^k + \left(\frac{h_0}{h_1} (b_j^k) + \frac{1}{2} h_0 a_j^k \right) v_{j+1}^k \\
+ \left(-\frac{h_0}{h_1} (b_j^k) - \frac{1}{2} h_0 a_j^k \right) v_{j-1}^k + h_0 f_j^k, \\
v_j^N = g_j, \\
\end{cases}
\]

and we deduce the following result:

**Lemma 1.27.** The above scheme is simply monotonic whenever

\[
\begin{cases}
(i) \frac{h_0}{h_1^2} \|a\|_\infty \leq 1, \\
(ii) h_1 |b(x,t)| \leq a(x,t), \quad \text{for all } (x,t) \in \mathbb{R} \times [0,T],
\end{cases}
\]
and in that case $\|v\|_\infty \leq \|g\|_\infty + T\|f\|_\infty$.

**Remark 1.28.** We say that the diffusion coefficient is nondegenerate if

$$a := \inf_{x,t} a(x,t) > 0.$$  

In that case (1.100)(ii) holds whenever $h_1 \leq a/\|b\|_\infty$.

**Exercice 1.29.** Analyze a variant in which the discretization of the first order term is a convex combination of the centered and implicit scheme, with the constraint that the resulting scheme is simply monotonic. Extend the study to the case of implicit scheme.

### 3.3. Implicit schemes.

#### 3.3.1. Standard implicit schemes.

The implicit upwind scheme is as follows

$$
\begin{cases}
 v_{j}^{k+1} - {v_j}^k = \frac{h_0}{h_1} \left( {v_j}^{k+1} - {v_j}^k \right) + (b_j)^+ + (b_j)^- - \frac{v_j^k - v_{j-1}^k}{h_1} \\
 + \frac{1}{2} a_j v_{j+1}^{k+1} + v_{j-1}^{k+1} - 2v_j^k + f_j^k = 0, \quad j \in \mathbb{Z}, \quad k = 0 : N - 1,
\end{cases}
$$

or equivalently

$$
\begin{cases}
 \left( 1 + \frac{h_0}{h_1} |b_j| + \frac{h_0}{h_1} a_j \right) v_j^k = \left( \frac{h_0}{h_1} (b_j)^+ + \frac{1}{2} \frac{h_0}{h_1} a_j \right) v_{j+1}^k + \\
 - \frac{h_0}{h_1} (b_j)^- + \frac{1}{2} \frac{h_0}{h_1} a_j \right) v_{j-1}^k + v_j^{k+1} + h_0 f_j^k, \quad j \in \mathbb{Z}, \quad k = 0 : N - 1,
\end{cases}
$$

Multiplying both sides by $\left( 1 + \frac{h_0}{h_1} |b_j| + \frac{h_0}{h_1} a_j \right)^{-1}$ (similarly to what has been done in section 2.4) we put this equation in a fixed-point form, the fixed-point operator being both contracting and simply monotonic in $\ell^\infty$. We obtain the following:

**Lemma 1.30.** Assume that $a \geq 0$ and that $a$, $b$ and $f$ are bounded. Then the implicit upwind scheme (1.102) is well-defined and simply monotonic.

#### 3.3.2. Semi implicit schemes.

With in view the possibility of having the time step of the order of the space step, we may consider an explicit (implicit) discretization of the first (second) order term. Since we are now used to finite difference schemes we may write directly the fixed-point form

$$
\begin{cases}
 \left( 1 + \frac{h_0}{h_1} a_j \right) v_j^k = \frac{1}{2} h_0 a_j \left( v_{j+1}^k + v_{j-1}^k \right) + \left( 1 - \frac{h_0}{h_1} |b_j| \right) v_j^{k+1} \\
 + \frac{h_0}{h_1} (b_j)^+ v_{j+1}^k + \frac{h_0}{h_1} (b_j)^- v_{j-1}^k + h_0 f_j^k, \quad j \in \mathbb{Z}, \quad k = 0 : N - 1.
\end{cases}
$$

This algorithm can be written in the following form: for $k = N - 1$ to 0, first compute $v_{k+1/2}$ solution of

$$
v_{k+1/2} = \left( 1 - \frac{h_0}{h_1} |b_j| \right) v_j^{k+1} + \frac{h_0}{h_1} (b_j)^+ v_{j+1}^{k+1} - \frac{h_0}{h_1} (b_j)^- v_{j-1}^{k+1} + h_0 f_j^k, \quad j \in \mathbb{Z},
$$
and then obtain $v^k$ solution of
\begin{equation}
(1.106) \quad \left(1 + \frac{h_0}{h_1} a_j^k\right) v_j^k = \frac{1}{2} \frac{h_0}{h_1} a_j^k \left(v_{j+1}^k + v_{j-1}^k\right) + v^{k+1/2} \quad j \in \mathbb{Z}.
\end{equation}

This a splitting method, in the sense that we have decomposed in two steps, each corresponding to one term of the dynamics.

**Remark 1.31.** Since the abstract setting in (1.33) applies also here, the conclusion of lemma 1.26 holds, provided of course that the CFL condition (1.92) is satisfied.

3.3.3. Implicit centered schemes. The implicit centered scheme is (compare to (1.98))
\begin{equation}
(1.107) \quad \begin{cases}
\frac{v_{j+1}^k - v_j^k}{h_0} + \frac{b_j^k}{2h_1} v_{j+1}^k - v_j^k + \frac{1}{2} a_j^k v_{j+1}^k + v_{j-1}^k - 2 v_j^k + f_j^k = 0, \quad j \in \mathbb{Z}, \ k = 1 : N, \\
v_j^N = g_j, \quad j \in \mathbb{Z}.
\end{cases}
\end{equation}

and the corresponding fixed-point form is
\begin{equation}
(1.108) \quad \begin{cases}
v_j^{k-1} = \left(1 + \frac{h_0}{h_1} a_j^k\right)^{-1} \left(\frac{1}{2} \left(\frac{h_0}{h_1} b_j^k + \frac{h_0}{h_1} a_j^k\right) v_{j+1}^k \right. \\
\quad \left. + \frac{1}{2} \left(-\frac{h_0}{h_1} b_j^k + \frac{h_0}{h_1} a_j^k\right) v_{j-1}^k + v_j^k + h_0 f_j^k\right), \quad j \in \mathbb{Z}, \ k = 1 : N, \\
v_j^N = g_j, \quad j \in \mathbb{Z}.
\end{cases}
\end{equation}

When (1.100)(ii) holds, the coefficients on the r.h.s. are nonnegative. It easily follows that the algorithm is well-defined and that its solution satisfies $\|v\|_\infty \leq \|g\|_\infty + T\|f\|_\infty$.

3.4. Some practical aspects. In practice we have to bound the domain of computation. We may assume that the domain has length $L$, so that there are $N_x = L/h_1$ space steps. One this is done, given that the domain includes space steps, we have that:

- The explicit algorithms need, since $h_0 = O(h_1^2)$, an order of $T/h_1^2$ time steps. Since $O(1)$ operations are performed at each point of the grid, a total of $O(TLh_1^{-3})$ operations is needed.
- The implicit algorithms involves tridiagonal linear systems of size $N_x$, which can be solved by Gaussian elimination in $O(N_x)$ operations. So an order of $TL/(h_0h_1)$ operations is needed. If $h_0 = O(h_1)$, the gain is of order $1/h_1$.

We next briefly discuss how to bound the domain. If $b$ and $\sigma$ are constant, then the stochastic process (remember that this is after the logarithmic transformation) say $z$, is such that $z(T) = z(0) + bT + \sigma W(T)$, i.e., has a Gaussian law with mean $\bar{z}_T := z(0) + bT$ and standard variation $\sigma_T := \sigma T^{1/2}$. We know that the probability that a Gaussian variable is say at distance of more than 5 times the standard variation from its expected value is very small. So, if we want to get significant results for $z(0) \in [\zeta_1, \zeta_2]$, we may bound the domain to
\begin{equation}
(1.109) \quad [\zeta_1 - |b|T - 5\sigma\sqrt{T}, \zeta_2 + |b|T + 5\sigma\sqrt{T}].
\end{equation}

**Example 1.32.** We have to compute an European call option with maturity one year, for a value of the underlying that is close to the strike equal to 1, so that $z(0) = 0$, and volatility of value (typically) 0.3. If the main contribution in the above interval is due to
the volatility, the computation domain will be $[-1.5, 1.5]$, that corresponds to $[0.22, 4.48]$ in the original space.

Remark 1.33. A more precise argument is as follows, if the computation domain is $[c, d]$, for a European call we should choose $d$ so that the following amount is be small enough, denoting by $p$ the density of $S(T)$:

$$
A := \int_d^\infty (e^s - K)p(s)\,ds = \frac{1}{\sigma_T(2\pi)^{1/2}}\int_d^\infty (e^s - K)\exp(-(s - \bar{z}_T)^2/(2\sigma_T^2))\,ds
$$

that we can evaluate, using the known asymptotics for the normal distribution.

Remark 1.34. We need boundary conditions when solving the discretization on a bounded domain; these conditions are often deduced from the structure of the problem. For instance, in the case of a European call option, when solving in the domain $[z_1, z_2]$, with $z_1 < K < z_2$, we usually take $z_1$ small enough and $z_2$ large enough, so that the expected value when $z = z_1$ is close to 0 (the final value of the underlying will be less that $K$ with high probability), and when $z = z_2$ is close to $z_2 - K$ (the final value of the underlying will be more that $K$ with high probability).

4. Multi dimensional problems

4.1. Orientation. In this section we will study multidimensional problems $(x \in \mathbb{R}^n)$ of the following form:

$$
(1.111) \quad \begin{cases}
V_t + \sum_{i=1}^n b_i(x,t)V_{x_i} + \frac{1}{2} \sum_{i,t=1}^n a_{it}(x,t)V_{x_ix_t} + f(x,t) = 0, & (x,t) \in \mathbb{R}^n \times [0,T], \\
V(x,T) = g(x), & x \in \mathbb{R}^n.
\end{cases}
$$

We have removed the term $"-r(x,t)V"$, present in the applications, but whose discretization makes no difficulty. We always assume that $a, b$ are bounded and that the diffusion matrix $a$ is symmetric, positive semidefinite.[4]

$$
(1.112) \quad a(x,t) \succeq 0, \quad \text{for all } (x,t) \in \mathbb{R}^n \times [0,T].
$$

Example 1.35. A standard two dimensional basket option problem (introducing a r.h.s. $f(x,t)$), is as follows:

$$
(1.113) \quad \begin{cases}
V_t + r(xV_x + yV_y) + \frac{1}{2}x^2\sigma_1^2V_{xx} + \frac{1}{2}y^2\sigma_2^2V_{yy} + \rho xy\sigma_1\sigma_2V_{xy} - rV + f = 0, & (x,y,t) \in \mathbb{R}^2 \times [0,T], \\
V(x,y,T) = g(x,y), & (x,y) \in \mathbb{R}^2.
\end{cases}
$$

Here $(x,y) \in \mathbb{R}^2$ is the notation for the space variables, and we skipped arguments $(x,y,t)$ of $r, f, \sigma$ and of the correlation $\rho \in [-1,1]$. After a logarithmic change of variable in space, we get the general format, with diffusion matrix is

$$
(1.114) \quad a = \begin{pmatrix}
\sigma_1^2 & \rho \sigma_1\sigma_2 \\
\rho \sigma_1\sigma_2 & \sigma_2^2
\end{pmatrix}.
$$

[4] We say that a symmetric matrix $A$ of size $n$ is positive semidefinite (resp. positive definite), and write $A \succeq 0$ (resp. $A \succ 0$) if $x^\top Ax \geq 0$, for all $x \in \mathbb{R}^n$ (resp. $x^\top Ax > 0$, for all nonzero $x \in \mathbb{R}^n$). If $B$ is another symmetric matrix of same size, we say that $A \succeq B$ if $A - B \succeq 0$, and $A \succ B$ if $A - B \succ 0$. 

The original final condition, in the case of a European call, would be $g_{EC}(x, y) := (\alpha x + \beta y - K)_+$, for some nonnegative parameters $\alpha$, $\beta$, and $K$. After the logarithmic transformation, the final condition is of the form $g_{EC}(x) := \left( \sum_{i=1}^{2} \alpha_i \exp(x_i) - K \right)_+$, for some nonnegative parameters $\alpha_1$, $\alpha_2$ and $K$.

**Remark 1.36.** In the previous example, when computing on a bounded domain, we can take Dirichlet boundary conditions as follows. If the boundary corresponds to $x$ or $y$ large we may expect that the final value of $\alpha x + \beta y$ will be greater than $K$ with high probability, and therefore choose $\alpha x + \beta y - K$ as boundary condition. If say the boundary corresponds to the minimum value of say $x$, then we may neglect the contribution of the $x$ variable. Then the evaluation on this part of the boundary reduces to the one dimensional European call problem. More generally, computing the boundary conditions in higher dimensions requires the computation of smaller dimensional problems option problems. An alternative is to choose poorer estimates for the boundary value, and compensating by taking a larger domain.

The non diagonal coefficient of the diffusion matrix may be interpreted as covariances between underlyings. In real-world examples these covariances are always nonzero. However, for pedagogical reasons, and since we can sometimes reduce to this case, we will start with the study of the case of diagonal diffusions.

### 4.2. Diagonal diffusions.

When the diffusion matrix $a$ is diagonal, we may write (1.111) in the form

$$
\begin{align*}
V_t + \sum_{i=1}^{n} b_i(x, t)V_{x_i} + \frac{1}{2} \sum_{i=1}^{n} a_{ii}(x, t)V_{x_i x_i} + f(x, t) &= 0, \quad (x, t) \in \mathbb{R}^n \times [0, T], \\
V(x, T) &= g(x), \quad x \in \mathbb{R}^n,
\end{align*}
$$

and in view of (1.112), we have that

$$
a_{ii}(x, t) \geq 0, \quad \text{for all } (x, t) \in \mathbb{R}^n \times [0, T] \text{ and } i = 1, \ldots, n.
$$

**Remark 1.37.** If the diffusion matrix $a(x, t)$ is constant, choosing as a basis of the state space an orthonormal set of eigenvectors of $a$, we can recover the case of a diagonal diffusion. Since diagonal diffusions may be dealt with efficiently by the finite difference numerical schemes, it is advisable to reduce to this case by a transformation in the state space whenever possible. Note that, however, it may be then more difficult to set boundary conditions.

We can obtain finite difference algorithms by applying to derivatives w.r.t. $x_i$ the finite difference operators of the previous section. We denote by $h_i > 0$ the space step for the $i$th space component, $e_i$ the $i$th element of the natural basis of $\mathbb{R}^n$, and set for $j \in \mathbb{Z}^n$ and $k \in \mathbb{N}$:

$$
x_j = \sum_i h_j e_i; \quad t_k = h_0 k; \quad b_{j,i}^k = b_i(x_j, t_k); \quad a_{j,ii}^k = a_{ii}(x_j, t_k),
$$
and so on. The explicit upwind algorithm is

\[
\begin{align*}
\frac{v_j^k - v_j^{k-1}}{h_0} + \sum_{i=1}^{n} \left( (b_{j,i})_+ \frac{v_{j+i}^k - v_j^k}{h_i} + (b_{j,i})_- \frac{v_j^k - v_{j-i}^k}{h_i} \right) \\
+ \frac{1}{2} \sum_{i=1}^{n} a_{j,ii} v_{j+i}^k + v_{j-i}^k - 2v_j^k h_i^2 + f_j^k = 0, \quad j \in \mathbb{Z}^n, \quad k = 1 : N,
\end{align*}
\]

(1.118)

The associated ordered form is

\[
\begin{align*}
\frac{v_j^k - v_j^{k-1}}{h_0} + \sum_{i=1}^{n} \left( (b_{j,i})_+ \frac{v_{j+i}^k - v_j^k}{h_i} + (b_{j,i})_- \frac{v_j^k - v_{j-i}^k}{h_i} \right) \\
+ \frac{1}{2} \sum_{i=1}^{n} a_{j,ii} v_{j+i}^k + v_{j-i}^k - 2v_j^k h_i^2 + f_j^k = 0, \quad j \in \mathbb{Z}^n.
\end{align*}
\]

(1.119)

A sufficient condition for simple monotonicity is therefore

\[
h_0 \left( \sum_{i=1}^{n} \frac{1}{h_i} \|b_i\|_\infty + \sum_{i=1}^{n} \frac{1}{h_i^2} \|a_{ii}\|_\infty \right) \leq 1.
\]

(1.120)

**Example 1.38.** Let \( n = 2, a \) be the identity, \( b = 0 \) (so that the PDE reduces to the heat equation), \( f = 0 \), and \( h_1 = h_2 \). We choose the maximum time step \( h_0 = \frac{1}{8} h_1^2 \). Then the explicit scheme reduces to

\[
v_j^{k-1} = \frac{1}{4} \left( v_{j+1}^k + v_{j-1}^k + v_{j+2}^k + v_{j-2}^k \right),
\]

(1.121)

with the same lack of damping as observed for one dimensional problems (example 1.6). If \( h_0 = \frac{1}{4} h_1^2 \), then the scheme reduces to

\[
v_j^{k-1} = \frac{1}{8} v_j^k + \frac{1}{8} \left( v_{j+1}^k + v_{j-1}^k + v_{j+2}^k + v_{j-2}^k \right).
\]

(1.122)

Other forms of FD algorithms are easily derived; for instance, an implicit upwind algorithm, written in the fixed point form, is (compare to (1.103)):

\[
\begin{align*}
\left( 1 + \sum_{i=1}^{n} \frac{h_0}{h_i} |b_{j,i}| + \sum_{i=1}^{n} \frac{h_0}{h_i^2} a_{j,ii} \right) v_j^k \\
= \sum_{i=1}^{n} \left( \frac{h_0}{h_i} (b_{j,i})_+ + \frac{1}{2} h_0 a_{j,ii} \right) v_{j+i}^k \\
+ \sum_{i=1}^{n} \left( -\frac{h_0}{h_i} (b_{j,i})_- + \frac{1}{2} h_0 a_{j,ii} \right) v_{j-i}^k + v_j^{k+1} + h_0 f_j^k, \quad j \in \mathbb{Z}^n, \quad k = 0 : N - 1,
\end{align*}
\]

(1.123)

\[v_j^N = g_j, \quad j \in \mathbb{Z}^n.
\]

Again this enters in the Markovian schemes setting, and so, by lemma 1.8, we have that:
Figure 9. Approximation of $D^2_{ij}$: case when $a_{ij} > 0$

**Lemma 1.39.** The above explicit and implicit algorithms are (assuming that (1.120) holds in the case of the explicit algorithm) well-posed, simply monotonic, and such that

$$\|v\|_{\infty} \leq \|g\|_{\infty} + T\|f\|_{\infty}. \quad (1.124)$$

**Remark 1.40.** The linear system to be solved for the implicit algorithm is quite expensive, even in dimension 2. Therefore it may be useful to consider splitting algorithms, in the spirit of e.g. [44].

**Exercise 1.41.** Give the expression of the explicit scheme when $n = 2$, $a$ is the identity, $b_1 = b_2 = 1$, $f = 0$, $h_1 = h_2$, and $h_0$ is the maximal time size.

Hint: note that the coefficient of $v_j^k$ in the ordered form is zero, and that $h_0 = \frac{1}{2h_1^2}/(1+h_1)$.

### 4. Diagonally dominant scaled diffusions

In this section we discuss how to approximate cross derivatives by finite differences methods; we will see that the resulting formulas are effective in the case of **diagonally dominant scaled diffusions matrices**. Since the resulting algorithm is simple to implement, one should try to reduce (by changes of variables) to the case of diagonally dominant diffusions matrices whenever possible.

When the diffusion matrix $a$ is not diagonal, in order to approximate cross derivatives, we have to introduce in the scheme shifts involving more than one space variable. Here we adopt the simple point of view of approximation of each term $v_{x_i x_\ell}$ separately. We use the notation $S_{\pm i}$, $i = 1$ to $n$, for the shift of $\pm 1$ in direction $i \in \{1, \ldots, n\}$ ($S_0$ for the identity) and $S_{\pm i, \pm \ell}$ for the combined shifts of $\pm 1$ in direction $i$ and $\ell$, with the usual abuse of notation by writing $S_{\pm i}w_j$ instead of $(S_{\pm i}w)_j$:

$$S_{\pm i}w_j := w_{j \pm e_i}; \quad S_{\pm i, \pm \ell} := w_{j \pm e_i \pm e_\ell}. \quad (1.125)$$

We can approximate the crossed derivative of $w = (w_j)$ at $x_j$, by the expression in the **upper right corner**:

$$\frac{w_{j + e_i + e_\ell} - w_{j - e_i - e_\ell} - w_{j + e_i} - w_{j + e_\ell}}{h_i h_\ell} = \frac{S_{i, \ell} + S_0 - S_i - S_\ell}{h_i h_\ell} w_j. \quad (1.126)$$
Indeed, if $\phi$ is smooth $\mathbb{R}^2 \to \mathbb{R}$, then by a Taylor expansion, we have that

\begin{align}
(1.127) \quad \frac{\phi(x + h_i e_i + h_\ell e_\ell) + \phi(x) - \phi(x + h_i e_i) - \phi(x + h_\ell e_\ell)}{h_i h_\ell} = D^2_{x_i x_\ell} \phi(x) + r(h),
\end{align}

with $r(h) = o\left(\frac{h_i^2 + h_\ell^2}{h_i h_\ell}\right)$, which is close to 0 if $h_i$ and $h_\ell$ are of the same order of magnitude.

We might as well consider shifts in the opposite directions (lower left corner):

\begin{align}
(1.128) \quad \frac{w_{j-e_i-e_\ell} + w_j - w_{j-e_i} - w_{j-e_\ell}}{h_i h_\ell} = \frac{S_{-i,-\ell} + S_0 - S_{-i} - S_{-\ell}}{h_i h_\ell} w_j.
\end{align}

Even better, we can take the centered formula obtained by taking the average of the two previous ones, whose operator, represented in figure 9, is

\begin{align}
(1.129) \quad \Delta^2_{i\ell} := \frac{S_{i,\ell} + S_{-i,-\ell} + 2S_0 - S_i - S_\ell - S_{-i} - S_{-\ell}}{2h_i h_\ell}.
\end{align}

Now we could as well introduce the alternative centered formula for which the two space variables vary in opposite directions, in the upper left and lower right corners, i.e.,

\begin{align}
(1.130) \quad \tilde{\Delta}^2_{i\ell} := \frac{-S_{i,-\ell} - S_{-i,\ell} - 2S_0 + S_i + S_\ell + S_{-i} + S_{-\ell}}{2h_i h_\ell}.
\end{align}

However, in order to have a simply monotonic algorithm, we have no choice. Indeed, when using say $\Delta^2_{i\ell}$, we see that the coefficient in the ordered form of an explicit algorithm of $v^{k+1}_{j+e_i+e_\ell}$ and $v^{k+1}_{j-e_i-e_\ell}$ will be $a^{k+1}_{j,i\ell}$. So the coefficient will be nonnegative iff $a^{k+1}_{j,i\ell}$ is so. Otherwise the use of $\tilde{\Delta}^2_{i\ell}$ allows to get a nonnegative coefficient for $v^{k+1}_{j+e_i-e_\ell}$ and $v^{k+1}_{j-e_i+e_\ell}$.

This leads to the following explicit algorithm, where we use the symmetry of $a(x,t)$:

\begin{align}
(1.131) \quad \begin{cases}
\quad v^k_j - v^{k-1}_j h_0 + \sum_{i=1}^{n} \left( (b^{k}_{j,i} v^{k}_{j+e_i} - v^{k}_{j}) + (b^{k}_{j,i}) v^{k}_{j-e_i} - v^{k}_{j-e_i} \right) \\
\quad + \frac{1}{2} \sum_{i=1}^{n} a^{k}_{j,i} v^{k}_{j+e_i} + v^{k}_{j-e_i} - 2v^{k}_j \\
\quad + \sum_{i<\ell, a^{k}_{j,i\ell} \geq 0} a^{k}_{j,i\ell} \Delta^2_{i\ell} v^{k}_j + \sum_{i<\ell, a^{k}_{j,i\ell} < 0} a^{k}_{j,i\ell} \Delta^2_{i\ell} v^{k}_j + f^k_j = 0, \quad j \in \mathbb{Z}^n, \quad k = 1 : N, \\
\quad v^N_j = g_j, \quad j \in \mathbb{Z}^n.
\end{cases}
\end{align}
The ordered form is
\[
\begin{aligned}
v_j^{k-1} &= \left(1 - \sum_{i=1}^{n} \frac{h_0}{h_i} |b_{j,i}^k| - \sum_{i=1}^{n} \frac{h_0}{h_i^2} a_{j,ii}^k + \frac{1}{2} \sum_{i \neq \ell} \frac{h_0}{h_i h_\ell} |a_{j,i\ell}^k|\right) v_j^k \\
&\quad + \sum_{i=1}^{n} \left(\frac{h_0}{h_i} (b_{j,i}^k)_+ + \frac{1}{2} \frac{h_0}{h_i^2} a_{j,ii}^k - \frac{1}{2} \sum_{\ell \neq \ell} \frac{h_0}{h_i h_\ell} |a_{j,i\ell}^k|\right) v_{j+e_i}^k \\
&\quad + \frac{1}{2} \sum_{i < \ell} \frac{h_0}{h_i h_\ell} \left((a_{j,ii}^k)_+ + (a_{j,\ell i}^k)_- - (a_{j,\ell i}^k)_- - (a_{j,\ell i}^k)_- + v_{j+e_i-e_\ell}^k - v_{j-e_i+e_\ell}^k\right) \\
&\quad + h_0 f_j^k, \quad j \in \mathbb{Z}^n, \quad k = 1 : N, \\
v_j^N &= g_j, \quad j \in \mathbb{Z}^n.
\end{aligned}
\]

The simple monotonicity condition involves now not only the coefficient of \(v_j^k\), but those of \(v_{j \pm e_i}\):
\[
\begin{aligned}
(1.133) \quad &h_0 \sum_{i=1}^{n} \left(\frac{1}{h_i} |b_{j,i}^k| + \frac{1}{h_i^2} a_{j,ii}^k - \frac{1}{2} \sum_{\ell \neq \ell} \frac{1}{h_i h_\ell} |a_{j,i\ell}^k|\right) \leq 1, \\
(1.134) \quad &\left(\frac{1}{h_i} (b_{j,i}^k)_+ + \frac{1}{2} \frac{1}{h_i^2} a_{j,ii}^k - \frac{1}{2} \sum_{\ell \neq \ell} \frac{1}{h_i h_\ell} |a_{j,i\ell}^k|\right) \geq 0, \quad i = 1, \ldots, n, \\
(1.135) \quad &\left(- \frac{1}{h_i} (b_{j,i}^k)_- - \frac{1}{2} \frac{1}{h_i^2} a_{j,ii}^k - \frac{1}{2} \sum_{\ell \neq \ell} \frac{1}{h_i h_\ell} |a_{j,i\ell}^k|\right) \geq 0, \quad i = 1, \ldots, n.
\end{aligned}
\]

A sufficient condition for the two last relations is that
\[
(1.136) \quad \frac{1}{h_i^2} a_{j,ii}^k \geq \sum_{\ell \neq \ell} \frac{1}{h_i h_\ell} |a_{j,i\ell}^k|.
\]

This is equivalent to the property of diagonally dominance for the scaled diffusion matrix \(a^h\) defined below:
\[
(1.137) \quad a_{i\ell}^h(x,t) := \frac{1}{h_i h_\ell} a_{i\ell}(x,t).
\]

Note that this relation is also necessary when the stepsize are small enough. We have obtained the following result, see Lions and Mercier \[42\].

**Proposition 1.42.** Let \(a^h(x,t)\) be diagonally dominant for all \((x,t) \in \mathbb{R}^n \times [0,T]\), and assume that
\[
(1.138) \quad \sum_{i=1}^{n} \left(\frac{1}{h_i} \|b_i\|_\infty + \sup_{x,t} \left(\frac{1}{h_i^2} a_{ii}(x,t) - \frac{1}{2} \sum_{\ell \neq \ell} \frac{1}{h_i h_\ell} |a_{i\ell}(x,t)|\right)\right) \leq \frac{1}{h_0},
\]
Then the above scheme is simply monotonic, and its solution satisfies
\[
\|v\|_\infty \leq \|g\|_\infty + T\|f\|_\infty.
\]

**Remark 1.43.** An obvious sufficient condition for (1.138) is
\[
\sum_{i=1}^{n} \left( \frac{1}{h_i} \|b_i\|_\infty + \frac{1}{h_i^2} \|a_{ii}\|_\infty \right) \leq \frac{1}{h_0}.
\]

On the other hand, since \(a^h\) is diagonally dominant, we have that
\[
\frac{1}{2} \sum_{i=1}^{n} \frac{1}{h_i^2} a_{ii}(x,t) \leq \sum_{i=1}^{n} \left( \frac{1}{h_i^2} a_{ii}(x,t) - \frac{1}{2} \sum_{\ell \neq i} \frac{1}{h_i h_\ell} |a_{i\ell}(x,t)| \right).
\]

Therefore, a necessary condition for (1.138) is
\[
\sum_{i=1}^{n} \left( \frac{1}{h_i} \|b_i\|_\infty + \frac{1}{2} \sup_{x,t} \left( \frac{1}{h_i^2} a_{ii}(x,t) \right) \right) \leq \frac{1}{h_0}.
\]

This implies that \(h_0 = O(\min_i h_i^2)\) whenever \(a \neq 0\).

**Example 1.44.** Consider the case when \(n = 2\), \(a(x,t) = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}\), \(b = 0\), \(f = 0\), \(h_1 = h_2\). The expression of the explicit scheme in ordered form (1.132) is, setting \(\chi := h_0/h_1^2\):
\[
v_j^{k-1} = (1 - \chi) v_j^k + \frac{1}{2} \chi (v_{j+e_1+e_2}^k + v_{j-e_1-e_2}^k).
\]
The maximal time step is therefore \(h_0 = h_1^2\), and then the scheme reduces to
\[
v_j^{k-1} = \frac{1}{2} (v_{j+e_1+e_2}^k + v_{j-e_1-e_2}^k).
\]
When taking the half of the maximal time step we obtain
\[
v_j^{k-1} = \frac{1}{2} v_{j+e_1+e_2}^k + \frac{1}{3} (v_{j+e_1+e_2}^k + v_{j-e_1-e_2}^k).
\]

**Example 1.45.** Consider the same problem, but now with \(a(x,t) = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}\), and \(b_1 = b_2 = 1\). The expression of the scheme (1.132) is, setting \(\chi_1 := h_0/h_1\) and \(\chi_2 := h_0/h_1^2\):
\[
v_j^{k-1} = (1 - 2\chi_1 - 3\chi_2) v_j^k + (\chi_1 + \frac{1}{2} \chi_2)(v_{j+e_1+e_2}^k + v_{j+e_2}^k) + \frac{1}{2} \chi_2 (v_{j-e_1+e_2}^k + v_{j-e_1-e_2}^k).
\]

**Remark 1.46.** When checking the correctness of the expression of the scheme, we may use the fact that the sum of weights of either \(\chi_1\) or \(\chi_2\) has to be zero.

## 5. Generalized finite differences

### 5.1. Basic tools

We consider again the problem of discretizing the PDE (1.111). In financial applications the diffusion matrix, while being semidefinite positive, is in general, even after a change of variables, not diagonally dominant. We explain in this section how to deal with this case, following the approach of [14]. We first introduce the *stencil*, defined as the set of integer spatial displacements allowed for points entering in the scheme. In
the case of a diagonal diffusion, the scheme of section 4.3 most often uses the diagonal stencil

\( \Xi_D := (\pm e_1, \ldots, \pm e_n) \).

In the diagonally dominant case, the stencil was the set of integers with components 0, 1, or -1, and with at most two nonzero components, that we call diagonally dominant stencil:

\( \Xi_{DD} = (\pm e_1, \ldots, \pm e_n) \cup (\cup \{ \pm e_i \pm e_\ell, 1 \leq i < \ell \leq n \}) \).

We next consider, for \( p = 1, 2, \ldots \), stencils of the form:

\( \Xi_p = \{ q \in \mathbb{Z}^n ; |q_i| \leq p, 1 \leq i \leq n \} \).

Note that \( \Xi_{DD} \subset \Xi_1 \), with equality when \( n = 2 \). We discuss a simple example in order to motivate the approach to be developed later.

**Example 1.47.** Consider the case of the equation

\( v_t + \frac{1}{2} \sum_{i, \ell=1}^{n} a_{i\ell} v_{x_i x_\ell} + f(x, t) = 0, \)

where \( a \succeq 0 \) is a constant rank one diffusion matrix; then \( a = \Sigma \xi^T \), for some \( \xi \in \mathbb{R}^n \). Take for instance

\( a = \begin{pmatrix} 4 & 2 \\ 2 & 1 \end{pmatrix} \).

This means that diffusion occurs along the direction of the vector \( \xi \). If \( h_1 = h_2 \) and \( a \) is as above, it seems therefore sensible to include the points \( \pm (2 1) \) in the stencil. Indeed, we have the explicit scheme

\( \frac{v_j^k - v_j^{k-1}}{h_0} + \frac{1}{2} \frac{v_j^{k+2e_1+e_2} + v_j^{k-2e_1-e_2} - 2v_j^k}{h_1^2} + f_j^k = 0, \quad j \in \mathbb{Z}^2, \quad k = 1 : N. \)

This is a consistent approximation since, if \( \phi \) is a smooth function over \( \mathbb{R}^2 \):

\( \varphi(x + 2h_1e_1 + h_1e_2) + \varphi(x - 2h_1e_1 - h_1e_2) - 2\varphi(x) = h_1^2 D^2 \varphi(x)(2e_1, e_2)^2 + o(h_1^2) = h_1^2 \sum_{i, \ell} a_{i\ell} D^2 x_i x_\ell \varphi(x) + o(h_1^2). \)

In view of the ordered form below, this scheme is also simply monotonic when \( h_0 \leq h_1^2 \):

\( \begin{cases} \frac{v_j^{k-1}}{h_0} = \left( 1 - \frac{h_0}{h_1^2} \right) v_j^k + \frac{h_0}{h_1^2} \left( v_j^{k+2e_1+e_2} + v_j^{k-2e_1-e_2} \right) + h_0 f_j^k, \quad j \in \mathbb{Z}^2, \quad k = 1 : N, \\
v_j^N = g_j, \quad j \in \mathbb{Z}^2. \end{cases} \)

We now come back to the general case. With the scaled diffusion matrix \( a^h \) defined in \( \Xi_{DD} \), we associated the values at the grid points defined by:

\( a_{j,i\ell}^h := \frac{a_{i\ell}(x_j, t_k)}{h_i h_\ell}. \)
DEFINITION 1.48. Given positive time and space steps \((h_0, \ldots, h_n)\), and a stencil (finite subset of \(\mathbb{Z}^n\)) \(\Xi\), we call \textit{stencil decomposition} w.r.t. \(\Xi\), or \(\Xi\) decomposition of the diffusion matrix \(a_j^{h,k}\), a collection of \textit{nonnegative} coefficients \(\eta_{j,\xi}^{h,k}\), \(\xi \in \Xi\), such that
\[
a_j^{h,k} = \sum_{\xi \in \Xi} \eta_{j,\xi}^{h,k} \xi \xi^\top.
\]

REMARK 1.49. The stencil decomposition is reminiscent of the one using eigenvalues, with the following important difference that the set of vectors is now constrained to belong to the stencil. Since it is determined by finitely many equalities and inequalities, the set of decompositions for a given stencil is a (possibly empty) polyhedron.

EXAMPLE 1.50. Consider the “heat equation” \(V_t + \frac{1}{2} \Delta V = 0\), corresponding to the case when \(a(x, t)\) is the identity \(I_d\). When \(n = 2\) and \(h_1 = h_2\), the \(\Xi_1\) decomposition corresponding to the usual discretization is
\[
I_d = h_1^{-2}(e_1 e_1^\top + e_2 e_2^\top),
\]
where the \(e_i\) are the elements of the natural basis. Yet another \(\Xi_1\) decomposition is
\[
I_d = \frac{1}{2} h_1^{-2}((e_1 + e_2)(e_1 + e_2)^\top + (e_1 - e_2)(e_1 - e_2)^\top).
\]

REMARK 1.51. When several stencil decompositions are possible we prefer those with the smallest norms of the elements \(\xi\) for which the coefficients are nonzero, so that the neighbouring points involved in the scheme are as close as possible.

With a given stencil decomposition, we associate the following finite difference operator:
\[
A_{\eta} v_j^k := \sum_{\xi \in \Xi} \eta_{j,\xi}^k \left( v_{j+\xi}^k + v_{j-\xi}^k - 2v_j^k \right).
\]
If \(\phi : \mathbb{R}^n \to \mathbb{R}\) is smooth, we remind that by a second order Taylor expansion:
\[
\phi(x_{j+\xi}) + \phi(x_{j-\xi}) - 2\phi(x_j) = \phi''(x_j)(x_{\xi}, x_{\xi}) + o(|x_{\xi}|^2) = \sum_{i,l=1}^n h_i h_{\xi_i} \xi_l \frac{\partial^2 \phi(x)}{\partial x_i \partial x_l} + o((\max_i h_i)^2).
\]
The finite difference operator in \((1.159)\) can be interpreted as a discretization of the operator of the PDE, since by \((1.156)\) \(a_{j,i,\xi} \eta_{j,\xi} = \sum_{\xi \in \Xi} \eta_{j,\xi}^k \xi \xi^\top\), and so:
\[
\sum_{\xi \in \Xi} \sum_{i,l=1}^n h_i h_{\xi_i} \xi_l \frac{\partial^2 \phi(x)}{\partial x_i \partial x_l} = \sum_{i,l=1}^n a_{j,i,\xi} h_i h_{\xi_l} \frac{\partial^2 \phi(x)}{\partial x_i \partial x_l} = \sum_{i,l=1}^n a_{j,i,\xi} \frac{\partial^2 \phi(x)}{\partial x_i \partial x_l}.
\]
Accordingly, we obtain (in the case of a pure diffusion \((1.150)\) for simplicity) the \textit{explicit scheme} (skipping the final condition)
\[
\frac{v_j^k - v_j^{k-1}}{h_0} + \frac{1}{2} \sum_{\xi \in \Xi} \eta_{j,\xi}^k \left( v_{j+\xi}^k + v_{j-\xi}^k - 2v_j^k \right) + f_j^k = 0, \quad j \in \mathbb{Z}^n, \quad k = 1 : N,
\]
or in \textit{ordered form}
\[
v_j^{k-1} = \left( 1 - h_0 \sum_{\xi \in \Xi} \eta_{j,\xi}^k \right) v_j^k + \frac{1}{2} h_0 \sum_{\xi \in \Xi} \eta_{j,\xi}^k \left( v_{j+\xi}^k + v_{j-\xi}^k \right) + h_0 f_j^k, \quad j \in \mathbb{Z}^n, \quad k = 1 : N.
\]
We deduce the following:

**Lemma 1.52.** A sufficient condition for the explicit scheme (1.162) to be simple monotone is

\[ h_0 \sum_{\xi \in \Xi} \eta_{j,\xi}^k \leq 1. \]  

If it holds, then the solution of the scheme satisfies

\[ \|v\|_{\infty} \leq \|g\|_{\infty} + T\|f\|_{\infty}. \]

We next check that (1.164) implies, as in the case \( n = 1 \), that the time step should be of the order of square of the space step:

**Lemma 1.53.** The following estimate for the coefficients of the decomposition holds:

\[ \sum_{\xi \in \Xi} \eta_{j,\xi}^k \leq \text{trace}(a_{j,\xi}^h) = \sum_{i=1}^n \frac{1}{h_i^2} a_{ii}(x_j, t_k) \leq \frac{\sup(\text{trace}(a))}{\min_i h_i^2}, \]

and therefore the simple monotonicity condition (1.164) is satisfied whenever

\[ \frac{h_0}{\min_i h_i^2} \sup(\text{trace}(a)) \leq 1. \]

**Proof.** Computing the trace on both sides of (1.156), and using \( \text{trace}(\xi \xi^\top) = |\xi|^2 \geq 1 \), we obtain (1.166), from which (1.167) follows. \( \square \)

The *implicit scheme* in ordered form may be written as (compare to (1.163)):

\[ \left(1 + h_0 \sum_{\xi \in \Xi} \eta_{j,\xi}^k\right) v_j^k = \frac{1}{2} h_0 \sum_{\xi \in \Xi} \eta_{j,\xi}^k \left(v_{j+\xi}^k + v_{j-\xi}^k\right) + v_{j+1}^k + \sum_{j, k, j \in \mathbb{Z}^n, k = 0 : N - 1} h_0 f_{j, k}. \]

By arguments similar to those of the diagonal matrices we deduce that

**Lemma 1.54.** The implicit scheme (1.168) is well-posed and simply monotonic, and its solution satisfies

\[ \|v\|_{\infty} \leq \|g\|_{\infty} + T\|f\|_{\infty}. \]

**Exercice 1.55.** In the case of a *diagonal diffusion*, check that a decomposition compatible with the diagonal stencil is

\[ a_{j,ii}^h = \sum_{i=1}^n a_{j,ii}^h e_i e_i^\top, \]

and that one recovers then the schemes of section 4.2.

**Exercice 1.56.** In the case of a *diagonally dominant scaled diffusion matrix*, denote the nonnegative dominance coefficients by

\[ \beta_{j,ii}^h := a_{j,ii}^h - \sum_{\ell \neq i} a_{j,\ell\ell}^h, \quad i = 1, \ldots, n. \]

Check that (i) a decomposition compatible with the diagonally dominant stencil (1.148) is

\[ \sum_{i=1}^n \beta_{j,ii}^h e_i e_i^\top + \sum_{i < \ell, a_{j,\ell\ell}^h \geq 0} a_{j,\ell\ell}^h (e_i + e_\ell)(e_i + e_\ell)^\top - \sum_{i < \ell, a_{j,\ell\ell}^h < 0} a_{j,\ell\ell}^h (e_i - e_\ell)(e_i - e_\ell)^\top, \]
(ii) the associated finite difference operator is

\[
\sum_{i=1}^{n} \gamma_{j,i}^{h,k} (v_{j+e_i} + v_{j-e_i} - 2v_j) + \sum_{i<\ell, a_{j,i\ell}^{h,k} \geq 0} a_{j,i\ell}^{h,k} (v_{j+e_i+e_\ell} + v_{j-e_i-e_\ell} - 2v_j)
\]

\[
- \sum_{i<\ell, a_{j,i\ell}^{h,k} < 0} a_{j,i\ell}^{h,k} (v_{j+e_i-e_\ell} + v_{j-e_i+e_\ell} - 2v_j),
\]

(iii) one recovers then the schemes of section 4.3.

5.2. Characterization of compatibility with a given stencil.

5.2.1. General relations.

We have obtained an effective way of generalizing the standard finite differences, provided the decomposition exists and can be computed quickly. This is an important difference w.r.t. the usual situation for finite differences algorithms where the coefficients of the scheme are given. We concentrate first on the existence of a decomposition. For that we first review briefly some results of the theory of linear programming and convex polyhedra [45].

**Convex cones and duality.** Let \( E \) be an Euclidean space. We say that \( C \subset E \) is a cone if \( \alpha c \in C \) for all \( \alpha \geq 0 \) and \( c \in C \). Let \( C \) be a non empty closed convex cone in \( E \). The **(positive) dual cone** \( C^+ \) is defined by

\[
C^+ := \{ y \in E; \ y \cdot z \geq 0, \ \text{for all} \ z \in C \}.
\]

This is obviously a convex cone, and the following holds, see [45] Thm 14.1:

**Lemma 1.57.** Let \( C \) be a non empty closed convex cone in \( E \). Then the bidual cone \( C^{++} := (C^+)^+ \) is equal to \( C \). In other words, duality is an involutive mapping on the set of non empty closed convex cones.

Let \( G \) be a finite subset of \( E \). We denote by \( \text{cone}(G) \) the set of linear combinations of elements of \( G \) with nonnegative coefficients; this is the smallest convex cone containing \( G \). If \( \text{cone}(G) = C \), we say that that \( C \) is **finitely generated** and that \( G \) is a **generator** of \( C \). In that case, we easily see that

\[
C^+ = \{ y \in E; \ y \cdot z \geq 0, \ \text{for all} \ z \in G \},
\]

and it is known that \( C^+ \) also has a finite generator say \( G' \). Since \( C^{++} \) is polar to \( C^+ \), it follows that

\[
C = \{ z \in E; \ y \cdot z \geq 0, \ \text{for all} \ y \in G' \}.
\]

Therefore checking if \( z \in C \) is easy once a generator of \( C^+ \) has been computed (at least if this generator is not too large).

**Application to the compatibility problem.** Let \( a, b \) be \( n \times p \) matrices. The **Frobenius norm** and related scalar product are defined by

\[
\|a\|_F := \left( \sum_{i,\ell} a_{i\ell}^2 \right)^{1/2}; \quad \langle a, b \rangle_F := \sum_{i,\ell} a_{i\ell}b_{i\ell} = \text{trace}(ab^T).
\]

We will apply the previous cone duality results, in the Euclidean space \( S_n^+ \) of symmetric matrices of size \( n \), endowed with the Frobenius norm.
We say that the (symmetric, positive semidefinite) matrix $a$ is **compatible** with the stencil $\Xi$ if a decomposition exists, i.e., if $a$ belongs to the convex cone

$$C(\Xi) := \text{cone}(\{\xi \xi^T; \xi \in \Xi\}).$$

Observe that the set of generators of $\Xi_p$ type stencils is invariant under (i) permutation of variables in $\mathbb{R}^n$, (ii) change of sign of some components of a variable in $\mathbb{R}^n$. Let us denote by $\mathcal{M}_p$ the set of transformations (invertible linear mappings) from $\mathbb{R}^n$ into itself that leaves $\Xi_p$ invariant:

$$M\xi \in \Xi_p, \text{ for all } M \in \mathcal{M}_p.$$ 

This set is a **group** containing all permutations as well as the operators of sign changing for some components. By the definition of $C(\Xi)$, we have that

$$MEM^T \in C(\Xi_p), \text{ for all } E \in C(\Xi_p) \text{ and } M \in \mathcal{M}_p.$$ 

The set of transformations $S^n \to S^n$, $E \mapsto MEM^T$, for $M \in \mathcal{M}_p$, is a group that leaves $C(\Xi_p)$ invariant.

For $M \in \mathcal{M}_p$, define $\hat{M} \in L(S^n)$ by $\hat{M}(E) := MEM^T$, and let

$$\hat{\mathcal{M}}_p := \{\hat{M}; \ M \in \mathcal{M}_p\}.$$ 

**Lemma 1.58.** The dual cone $C(\Xi_p)^+$ is invariant under the set of dual transformations $S^n \to S^n$, $\Lambda \mapsto M^T \Lambda M$, for $M \in \mathcal{M}_p$.

**Proof.** Let $\Lambda \in C(\Xi_p)^+$ and $M \in \mathcal{M}_p$. For any $E \in C(\Xi_p)$, we have that

$$0 \leq \langle \Lambda, MEM^T \rangle_F = \text{trace}(\Lambda MEM^T) = \langle \Lambda E, M \rangle_F = \langle EM^T \Lambda, M^T \rangle_F = \text{trace}(EM^T \Lambda M) = \langle E, M^T \Lambda M \rangle_F$$

from which the result follows. \(\square\)

**Remark 1.59.** This property will allow to express in a compact form the inequalities characterizing the inclusion in $C(\Xi_p)$, since it suffices essentially to state one inequality for each equivalent class of the quotient $G^+/\mathcal{M}^*$, where $G^+$ is a generator of $C(\Xi_p)^+$, and $\mathcal{M}^*$ is the above set of dual transformations.

**Example 1.60.** By exercise [1.56](#), the cone of diagonally dominant matrices coincides with $C(\Xi_{DD})$, where $\Xi_{DD}$ was defined in (1.148). This stencil is also invariant under permutations and sign changes, and so the conclusion of the above lemma holds in this case. One of the linear inequalities expressing diagonal dominance is $\sum_{j=2}^n a_{1j} \leq a_{11}$. In the space of symmetric matrices, this is equivalent to $\langle \Lambda, a \rangle_F \leq 0$ with

$$\Lambda = \begin{pmatrix} -1 & \frac{1}{2} & \cdots & \frac{1}{2} \\ \frac{1}{2} & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{2} & 0 & \cdots & 0 \end{pmatrix}.$$ 

All other inequalities defining diagonal dominance are image of this one by permutations and sign changes, and so there is only one equivalent class in this case.
5.2.2. Some explicit characterizations of compatibility. Computations of a generator of $C(\Xi)^+$ lead to complex expressions that can hardly be done by hand, except in rather trivial situations. We reproduce some of the results of [14], based on the quickhull algorithm [4], whose implementation uses floating-point computation, that computes extreme points of polytopes (from which the computation of a generator of a convex cone is easily done). In dimension 2 we already know that $C(\Xi_1)$ is the set of diagonally dominant matrices. The set $C(\Xi_2)$ is characterized by 8 constraints and 2 equivalence classes:

$$a \in C(\Xi_2) \iff \begin{cases} 2a_{ii} \geq |a_{ij}|, \\ 2a_{ii} + a_{jj} \geq 3|a_{ij}|, \end{cases} \quad \text{for } 1 \leq i \neq j \leq 2.$$ 

Characterizations were obtained also for $C(\Xi_3)$ to $C(\Xi_7)$; the latter is characterized by 72 constraints and 18 equivalence classes. For $n = 3$ the sets $C(\Xi_1)$ and $C(\Xi_2)$ have been computed. The expression of $C(\Xi_1)$ is

$$\{ a_{ii} \geq |a_{ij}|, \\ a_{ii} + a_{jj} \geq (-1)^p a_{ik} + (-1)^q a_{jk} + 2(-1)^{p+q+1}a_{ij} \} \quad \text{for } i \neq j \neq k \text{ and } p, q \in \{1, 2\}.$$ 

When the dimension of the state space increases, the complexity of the compatibility relations increases very rapidly, so that checking these relations may be more expensive than the resolution of the linear program allowing to compute the decomposition whenever it exists.

**Remark 1.61** (Existence of a decomposition for non degenerate operators). For a (uniformly) nondegenerate operator compatibility with the stencil $\Xi_p$ (for any point of the grid) holds for large enough $p$. See Krylov [35], in which smoothness of the coefficients is discussed, and also Kuo and Trudinger [37].

Note that the generalized finite difference methods enter in the more general framework of Markov chain approximations, presented in Kushner and Dupuis [38].

5.2.3. Best approximation decomposition. When a matrix $a$ is not compatible with a given stencil $\Xi$, one possibility is to compute instead the decomposition of the compatible matrix that is the closest to $a$ in a certain sense. When choosing the Frobenius norm (1.176), this amounts to solve the following quadratic optimization problem:

(1.183) \[ \min_{\eta \geq 0} \frac{1}{2} \left\| \sum_{\xi \in \Xi} \eta \xi \xi^T - a \right\|_F^2; \]

This is a convex problem, whose solution is not necessarily unique. However, for any solution $\eta$, the matrix $\sum_{\xi \in \Xi} \eta \xi \xi^T$ is nothing but the projection denoted $P_{\Xi}a$ of $a$ onto $C(\Xi)$, for the Frobenius norm. We can then use $P_{\Xi}a$ instead of $a$ for the design of the numerical algorithms.

For two dimensional problems the maximum relative errors (denoted $\varepsilon_{p_{max}}$) as functions of $p$ have been computed in [9]; their results are displayed in table [1]. An algorithm involving only the closest neighbour can make up to 17% of relative error on diffusion matrices, and hence, will perform poorly in general. A relative precision of 1% needs to take $p = 5$. We denote by $p_\varepsilon$ the minimal value of $p$ that allows to obtain a relative error of $\varepsilon$.

**Remark 1.62.** In the case when all $h_i$, $i > 0$ are equal (to which we can reduce by a scaling of the space variables), we see that the resulting scheme can be interpreted as a
consistent and simply monotonic scheme for the perturbed problem (written in the case of a pure diffusion (1.150))

\[
v_t + \frac{1}{2} \sum_{i,j} \hat{a}_{ij}(x,t)v_{ij} + f(x,t) = 0,
\]

where \( \hat{a}(x,t) := P_\Xi(a(x,y)) \). This leads to two questions:

a) for a given stencil, what will be the maximal relative distance between a (positive semidefinite) matrix \( a \) and its projection? We just discussed this question when \( n = 2 \).

b) Can we estimate the distance between the solutions of (1.184) and the one of the original diffusion? We note that, in the context of viscosity solutions, Jakobsen and Karlsen \[32\] obtain estimates of the form

\[
\|v - v'\|_\infty \leq C\|a - \hat{a}\|_\infty^{1/2}.
\]

5.3. A fast 2D decomposition algorithm. We study here the best approximation problem when \( n = 2 \). It was shown in \[9\] that, using some arithmetic properties in the theory of rational approximation, we are able to compute a decomposition of a semidefinite positive matrix in \( \Xi_p \), with a complexity of only \( O(p) \) operations. With a symmetric matrix \( a \in S^2_+ \) we associate its view:

\[
\text{View}(a) := \left( \frac{a_{11} - a_{22}}{2}, \frac{2a_{12}}{a_{11} + a_{22}} \right).
\]

We easily check that see that the set of views of positive semidefinite matrices is the unit ball of \( \mathbb{R}^2 \) for the Euclidean norm:

\[
\text{View}(S^2_+) = \bar{B}(0, 1)
\]

The view of the identity is the zero vector, and the view of \( \eta \eta^\top \), where \( \eta := (1 \ 0)^\top \), is \( (1 \ 0) \). The lemma below eases the computation of the view of any rank one symmetric nonnegative matrix, and is illustrated in figure [13]

**Lemma 1.63.** Let \( \eta_\theta := (\cos \theta, \sin \theta) \). Then the view of \( \eta \eta^\top \) is \( \eta_{2\theta} \).

**Proof.** Since \( \eta \eta^\top = \begin{pmatrix} \cos^2 \theta & \cos \theta \sin \theta \\ \cos \theta \sin \theta & \sin^2 \theta \end{pmatrix} \), the associated view is

\[
(\cos^2 \theta - \sin^2 \theta, 2 \cos \theta \sin \theta) = (\cos 2\theta, \sin 2\theta),
\]

as was to be proved. \( \square \)

**Lemma 1.64.** The view of diagonally dominant matrices is the unit ball in the \( \ell^1 \) norm.
Proof. This follows immediately from the following (easy to check) characterization of diagonally dominant two dimensional symmetric matrices:

\[ |a_{11} - a_{22}| + 2|a_{12}| \leq a_{11} + a_{22}. \]

Diagonally dominant matrix have the well-known decomposition (particular case of (1.171) when \( n = 2 \))

\[
a = (a_{11} - |a_{12}|) \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1 0) + (a_{22} - |a_{12}|) \begin{pmatrix} 0 \\ 1 \end{pmatrix} (0 1)
+ \max(a_{12}, 0) \begin{pmatrix} 1 \\ 1 \end{pmatrix} (1 1) + \max(-a_{12}, 0) \begin{pmatrix} -1 \\ 1 \end{pmatrix} (-1 1).
\]

Let us call “inner region” of the positive semidefinite cone, the set of diagonally dominant matrices. There are four outer regions corresponding to the violation of one of the four constraints \( \pm a_{12} \leq a_{ii} \), for \( i = 1, 2 \). Define the outer region \( I \) is the set of positive semidefinite and non diagonally dominant matrices such that \( a_{22} < a_{12} < a_{11} \). It is easy to reduce any diffusion matrix to this case by permutation of variables and change of sign of one state variable. Therefore in the sequel we will discuss essentially the fast decomposition of such matrices. Note that for positive semidefinite and diagonally dominant matrices in region \( I \) an alternative decomposition, involving the identity matrix is

\[
a = (a_{11} - a_{22}) \begin{pmatrix} 1 \\ 0 \end{pmatrix} (1 0) + (a_{22} - a_{12}) \begin{pmatrix} 1 \\ 1 \end{pmatrix} (0 1) + a_{12} \begin{pmatrix} 1 \\ 1 \end{pmatrix} (1 1).
\]

**Figure 10. Family relations in regular grid**

**The Stern-Brocot tree.** Since the algorithm should use points in the stencil as close to 0 as possible, it suffices to take such \( \xi \) with relatively prime components.
5. GENERALIZED FINITE DIFFERENCES

For two dimensional problems on which we focus now, such points have a specific structure. For reason of symmetries, we have displayed in figure [10] one eighth of the neighbouring points, namely the points $\xi$ in $\mathbb{N}_+^2$, such that $\xi_2 \leq \xi_1$. Those with an irreducible associated (symbolic) fraction $\xi_2/\xi_1$, that we will call irreducible points, are in red (boldface in black and white printing). These points are connected by segments that represent the arcs of a tree that we introduce now.

A very effective way for generating direction with irreducible components is to use the Stern-Brocot tree, see e.g. [28], displayed in figure [11]. In the sequel, when we write $q/p$ this should be understood as the pair $(p,q)$, so that $p = 0$ makes no problem.

The tree starts with two roots $0/1$ and $1/0$. At any stage of the construction, between two adjacent nodes $q/p$ and $q'/p'$, called the parents, insert the child node $(q+q')/(p+p')$. The two roots are adjacent, and hence, the first child is $1/1$. Then each child is made adjacent with each of his two parents, and we can repeat the process of generating children (in any order).

Figure [10] shows the links between parents and child for the first nodes of the Stern-Brocot tree. One finds the two parents of a child-node following the two segments starting from this point and going to the left. For convenience we reproduce a short proof of classical properties of the Stern-Brocot tree, see also [28 section 4.5].

**Figure 11. Stern-Brocot tree**

**Lemma 1.65.** Let $q/p$ and $q'/p'$ be adjacent nodes such that $q/p < q'/p'$, with child $q''/p''$, where $p'' = p + p'$, $q'' = q + q'$. Then

(i) $q/p < q''/p'' < q'/p'$,

(ii) every node of the Brocot tree is irreducible,

(iii) every irreducible fraction $b/a$ belongs to the Brocot tree.

Furthermore, if $q/p$ and $q'/p'$ are adjacent nodes of the tree such that $q/p < b/a < q'/p'$, then

(1.192) \[ a \geq p + p'; \quad b \geq q + q'. \]
Proof. (i) It is easily checked that \( q/p < (q + q')/(p + p') < q'/p' \). (This property explains why generation of sons may be made in any order.)

(ii) We prove by induction that, if \( q/p \) and \( q'/p' \) are adjacent nodes of the tree, then

\[
q'p - qp' = 1.
\]

The relation is obviously true for the root nodes 0/1 and 1/0. Assume that it is satisfied for adjacent nodes \( q/p \) and \( q'/p' \). It follows from (1.193) that \( q'(p + p') - p'(q + q') = 1 \) and \( p(q + q') - q(p + p') = 1 \), proving the induction. Combining (1.193) and Bézout’s theorem, we obtain (ii).

(iii) Let \( b/a \) be an irreducible fraction, with \( 0 < b/a < 1 \), and \( q/p, q'/p' \) be adjacent nodes of the tree such that \( q/p < b/a < q'/p' \). Then \( bp - aq \geq 1 \) and \( ap' - bp' \geq 1 \). Multiply the first (second) inequality by \( p' \) (by \( p \)) and add them; multiply the first (second) inequality by \( q' \) (by \( q \)) and add them; using (1.193), relation (1.192) follows. Since \( p'' \geq \max(p, p') + 1 \), this relation implies that there is a finite number of couple of adjacent nodes \( (q/p, q'/p') \) in the tree such that \( q/p < b/a < q'/p' \) holds. This is the case for the two root nodes. Assume now that \( b/a \) does not belong to the Stern-Brocot tree. If \( q/p < b/a < q'/p' \), setting \( q'' = q + q' \) and \( p'' = p + p' \), we see that either \( q/p < b/a < q''/p'' \), or \( q''/p'' < b/a < q'/p' \). In this way we generate an infinite sequence of adjacent nodes such that \( q/p < b/a < q'/p' \). The desired contradiction follows.

\[\text{Figure 12. Correspondence of directions}\]

**Decomposition of the scaled diffusion matrix.** In the sequel we will present a fast algorithm for computing the decomposition of diffusion matrices, when the stencil is \( \Xi_p \). Since we want that the decomposition involves small elements of the stencil when possible we will use only in the decomposition the set of directions with integer irreducible components.

As discussed before, it suffices to discuss the case when the matrix \( a^h \) is in outer region \( I \); i.e., when it is positive semidefinite and non diagonally dominant, and \( a_{22} < a_{12} < a_{11} \). On figure 12 this means that the view of \( a^h \) belongs to the quarter of ball in the upper right side, and is not in the triangle with vertices of coordinates \((0, 0)\), \((1, 0)\) and \((0, 1)\). The latter correspond to the identity matrix, and to degenerate diffusions with horizontal
and angle of $\pi/4$ diffusions. (The cone generated by these three points is a set of diagonally dominant matrices).

With every node $q/p$ of the Stern-Brocot tree, $q \leq p$, associate directions $\xi_{p,q} := (p/q)^{\top}$ and $X_{p,q} := \xi_{p,q}^\top \xi_{p,q}$. With two adjacent nodes is associated the plane $H(q/p, q'/p')$ generated by $X_{p,q}$ and $X_{p',q'}$, and two half spaces, the inner one (containing the identity matrix) and the outer one. If $p/q > p'/q'$, let $y := \text{View}(X_{p,q}) \wedge \text{View}(X_{p',q'})$ (where here $\wedge$ denote the vector product in $\mathbb{R}^2$). Then a matrix $a$ belongs to the outer region iff $y \cdot \text{View}(a) \geq 0$, see figure 12. Denote by $P_H(q/p, q'/p')$ the orthogonal projection onto this plane (since the mapping (??) onto $\mathbb{R}^3$ is isometric, projection w.r.t. the Frobenius norm is equivalent to the Euclidean projection in the image space $\mathbb{R}^3$).

Beginning the search of a decomposition, we are in the following situation: the matrix $a^h$ belongs to the outer half space of $H(0/1, 1/1)$. So, let us assume more generally that $a^h$ belongs to the outer half space of $H(q/p, q'/p')$, where $q/p$ and $q'/p'$ are adjacent nodes. In figure 12 we have drawn the views of the first nodes of the Stern-Brocot tree; the segments are the views of the segment between two neighbouring nodes of this tree.

We see, using lemma 1.63 that we have to use another element of stencil of the form $\hat{q}/\hat{p}$, with $\hat{q}$ and $\hat{p}$ nonnegative, such that $q/p < \hat{q}/\hat{p} < q'/p'$, and as small as possible. In view of (1.192), the optimal choice is to take the child $q''/p'' = (q + q')/(p + p')$. Then (see figure 12) there are two possibilities:

- The matrix $a^h$ belongs to both inner half spaces of $H(q/p, q''/p'')$ and $H(q'/p'', q'/p')$. Then $a^h$ belongs to the cone generated by $X_{p,q}$, $X_{p',q'}$ and $X_{p'',q''}$. Since these three matrices are linearly independent, the corresponding coefficients are unique nonnegative solution of the invertible (three dimensional) system

\[ \eta_{p,q} X_{p,q} + \eta_{p',q'} X_{p',q'} + \eta_{p'',q''} X_{p'',q''} = a^h. \]

- The matrix $a^h$ belongs to at least one outer half space. Since $X_{p'',q''}$ belongs to the boundary of the cone of positive semidefinite matrices, $a^h$ cannot belong to both outer half spaces (see figure 12). We are therefore lead to the situation at the beginning, setting either $q/p$ or $q'/p'$ to $q''/p''$.

If $p'' > p_{\max}$, we replace $a^h$ by its projection onto the cone generated by matrices of the form $X_{p,q_i}$, with either $q_i/p_i < q/p$ or $q'/p' < q_i/p_i$. Note that this projection belongs to the cone generated by $X_{p,q}$ and $X_{p',q'}$. As above, since these two matrices are linearly independent, the corresponding coefficients are unique nonnegative solution of the system

\[ \eta_{p,q} X_{p,q} + \eta_{p',q'} X_{p',q'} = P_H(q/p, q'/p') a^h. \]

This leads to an effective algorithm, that will stop either if the exact decomposition is obtained, or if either $p'' > p_{\max}$, or if the projection of $a^h$ onto $H(q/p, q'/p')$ is close enough to $a^h$. The precise algorithm is as follows: $\varepsilon$ is a threshold for the distance to the projection of $a^h$ onto the class of consistent matrices, and $p_{\max}$ is the size of stencil:

Algorithm DECOMP

**Initial Phase:** Data $\varepsilon \geq 0$, $p_{\max} > 0$. Set $k := 0$.
- If $a^h$ is diagonally dominant: set $\eta$ using (1.191) and stop.
- Reduction to region I, i.e. $a^h_{22} < a^h_{12} < a^h_{11}$.

Set $q_0/p_0 := 0/1$, $q_0'/p_0' := 1/1$.

**Repeat**
- Compute $a^r := P_H(q/p, q'/p') a^h$. 

• If \( \|a' - a^h\| \leq \varepsilon\|a^h\| \) or \( p + p' > p_{\text{max}} \): compute \( \eta \), decomposition of \( a' \) using (1.195) and stop.
• Set \( q''/p'' := (q + q')/(p + p') \).
• If \( a^h \) in inner half spaces of \( H(q/p,q''/p'') \) and \( H(q/p,q''/p'') \): compute \( \eta \) using (1.194) and stop.
• If \( a \) is in outer half space of \( H(q/p,q''/p'') \): \( q/p \) := \( q''/p'' \).
• \( k := k + 1 \).

END REPEAT

From the above discussion we have the following result.

**Theorem 1.66.** Algorithm **DECOMP** provides a decomposition of \( a^h \) with a relative error lower than \( \varepsilon \), and stops after at most \( p_{\text{max}} \) iterations. The cost of each iteration is \( O(1) \) operations, and hence, its total cost is no more than \( O(p_{\text{max}}) \).

![Figure 13. Correspondence of angles](image)

We plot in figure [14] the values of the numerical errors when the space steps vanish, for various values of \( p \); see details of the model in [9]. As expected, the higher \( p \), the better the precision.

### 6. Convergence analysis

#### 6.1. Basic analysis based on consistency and simple monotonicity

We analyze the error estimates in the case of the one dimensional space equation (1.25), assuming that (1.25) has a smooth solution \( V \) (i.e., with bounded derivatives of sufficiently large order). Denote by \( v_j^k \) the solution of the **explicit scheme**

\[
\begin{align*}
&v_j^k - v_j^{k-1} + (b_j^k) + \frac{v_{j+1}^k - v_j^k}{h_0} + (b_j^k) - \frac{v_j^k - v_{j-1}^k}{h_1} + \\
&\frac{1}{2} a_j^k v_{j+1}^k + v_j^{k-1} - 2v_j^k - \frac{r_j^k v_j^k + f_j^k = 0}{h_1} = 0, \quad j \in \mathbb{Z}, \quad k = 1 : N, \\
v_j^N = g(x_j), \quad j \in \mathbb{Z},
\end{align*}
\]

(1.196)
We easily check that the scheme is simply monotonic whenever the following condition holds:

\[(1.197) \quad h_0 \left( \|r\|_\infty + \|b\|_\infty \frac{1}{h_1} + \|a\|_\infty \frac{1}{h_1^2} \right) \leq 1.\]

Set \( \hat{v}^k_j := V(x_j, t_k) \). We may view \( \hat{v} \) as a solution of the discretized scheme with a perturbed r.h.s.:

\[(1.198) \quad \left\{ \begin{array}{l}
\frac{\hat{v}^k_j - \hat{v}^{k-1}_j}{h_0} + \frac{(b^k_j)_+ - \hat{v}^k_{j+1} - \hat{v}^{k-1}_j}{h_1} + \frac{(b^k_j)_- - \hat{v}^k_{j-1} - \hat{v}^{k-1}_j}{h_1} \\
+ \frac{1}{2} a^k_j \hat{v}^{k+1}_j + \hat{v}^{k-1}_j - 2 \hat{v}^k_j - r^k_j \hat{v}^k_j + f^k_j = e^k_j, \quad j \in \mathbb{Z}, \quad k = 1 : N,
\end{array} \right.\]

where the consistency error \( e^k_j \) of the scheme can be decomposed as a sum of contributions of the terms of time derivative, transport and diffusion:

\[(1.199) \quad e^k_j = e^k_{0,j} + e^k_{1,j} + e^k_{2,j},\]

with

\[(1.200) \quad e^k_{0,j} := \frac{\hat{v}^k_j - \hat{v}^{k-1}_j}{h_0} - V_t(x_j, t_k),\]

\[(1.201) \quad e^k_{1,j} := (b^k_j)_+ \frac{\hat{v}^k_{j+1} - \hat{v}^k_j}{h_1} + (b^k_j)_- \frac{\hat{v}^k_{j-1} - \hat{v}^{k-1}_j}{h_1} - b^k_j V_x(x_j, t_k),\]

\[(1.202) \quad e^k_{2,j} := \frac{1}{2} a^k_j \left( \frac{\hat{v}^k_{j+1} + \hat{v}^{k-1}_j - 2 \hat{v}^k_j}{h_1^2} - V_{xx}(x_j, t_k) \right).\]
Lemma 1.67. Assume that \( \inf r \geq 0 \) and that the monotonicity condition (1.197) holds. Then the following error estimate holds:

\[
(1.203) \quad \sup_{k,j} |V(x_j, t_k) - v_j^k| \leq T \sup_{k,j} |e^k_j|.
\]

Proof. Since \( \tilde{v}_j^k := V(x_j, t_k) - v_j^k \) is solution of the same scheme with zero final condition and source term \(-e^k_j\), this is an immediate consequence of lemma 1.14. \(\square\)

Assume now that the function \( V \) is smooth with bounded derivatives of any order. Using standard first-order Taylor expansions with integral remainders of \( v_t \) and \( v_x \), we obtain

\[
(1.204) \quad |e_{0,j}^k| = \left| \int_0^1 (V_t(x_j, t_k - h_0s) - V_t(x_j, t_k)) \, ds \right| \leq \|V_t\|_{\infty} h_0,
\]

and similarly

\[
(1.205) \quad \left| \frac{\tilde{v}_{j+1}^k - \tilde{v}_j^k}{h_1} - V_x(x_j, t_k) \right| \leq \|V_x\|_{\infty} h_1 \Rightarrow |e_{1,j}^k| \leq \|b\| \|V_x\|_{\infty} h_1.
\]

Next consider the following third order expansion (we skip the time argument and take \( \pm \) equal to either 1 or -1 everywhere)

\[
(1.206) \quad V(x \pm h_1) = V(x) \pm V_x(x)h_1 + \frac{1}{2}h_1^2V_{xx}(x) \pm \frac{1}{2}h_1^3 \int_0^1 (1-s)^2 V_{xxx}(x \pm sh_1) \, ds.
\]

Summing the above two equalities for \( \pm 1 \) and \( x = x_j \), we obtain

\[
(1.207) \quad \left| \frac{\tilde{v}_{j+1}^k + \tilde{v}_{j-1}^k - 2\tilde{v}_j^k}{h_1^2} - V_{xx}(x_j, t_k) \right| \leq \frac{h_1}{2} \int_0^1 (1-s)^2 |V_{xxx}(x + sh_1) - V_{xxx}(x - sh_1)| \, ds.
\]

Using \( |V_{xxx}(x + sh_1) - V_{xxx}(x - sh_1)| \leq 2\|V_{xxx}\|_{\infty} h_1 \), we obtain that an upper bound of the r.h.s. is \( \|V_{xxx}\|_{\infty} h_1^2 \), and therefore

\[
(1.208) \quad |e_{2,j}^k| \leq \frac{1}{2} \|a\| \|V_{xxx}\|_{\infty} h_1^2.
\]

Note that since this is an even order derivative and successive grid points are equidistant, we have one error order more than for the other error terms. Combining the previous estimates we get

\[
(1.209) \quad \|e\|_{\infty} \leq \|V_t\|_{\infty} h_0 + \|b\| \|V_x\|_{\infty} h_1 + \frac{1}{2} \|a\| \|V_{xxx}\|_{\infty} h_1^2.
\]

We deduce then from lemma 1.67 the following error estimate.

Theorem 1.68. For the scheme (1.94), we have the following error estimate:

\[
(1.210) \quad \sup_{j,k} |V(x_j, t_k) - v_j^k| \leq T \left( \|V_t\|_{\infty} h_0 + \|b\| \|V_x\|_{\infty} h_1 + \frac{1}{2} \|a\| \|V_{xxx}\|_{\infty} h_1^2 \right).
\]

Exercise 1.69. Derive similar estimates for the implicit or semi implicit schemes, and for problems with space dimension greater than one.

As we will see, the solution of (1.25) is not always not so smooth, and this will lead us to introduce a method of regularization by convolution in order to obtain error estimates in the general case.
6.2. Smoothness of the solution of \((1.25)\). We recall the notion of Hölderian mappings.

**Definition 1.70.** Let \( E \subset \mathbb{R}^n \). We say that \( w : E \to \mathbb{R} \) is Hölder with constant \( C_w \geq 0 \) and exponent \( \mu_w \in (0, 1] \) (or with constants \( (C_w, \mu_w) \)), if the following holds:

\[
|w(x) - w(y)| \leq C_w |x - y|^{\mu_w}, \quad \text{for all } x, y \in E.
\]

Note that, if \( \mu_w = 1 \) we recover the definition of a Lipschitz function. More generally, for \( w : \mathbb{R} \times [0, T] \to \mathbb{R} \), we denote by \( \mu_{w,x}, \mu_{w,t} \in (0, 1] \) the Hölder exponents w.r.t. \( x \) and \( t \); these are (whenever they exist) the biggest constants in \((0, 1]\) such that, for some \( C_w > 0 \):

\[
|w(x', t') - w(x, t)| \leq C_w \left( |x' - x|^{\mu_{w,x}} + |t' - t|^{\mu_{w,t}} \right), \quad \text{for all } x \text{ and } t.
\]

Consider the space of bounded functions that are Lipschitz in space and Hölder with exponent \( 1/2 \) in time:

\[
\mathcal{V}_{1,1/2} := \{ w : \mathbb{R} \times [0, T] \to \mathbb{R} \text{ bounded; } \mu_{w,x} = 1, \mu_{w,t} = 1/2 \}.
\]

**Remark 1.71.** Usually, in financial applications, using the Feynman Kac-formula, which gives a representation as an expectation of a cost associated to some stochastic process, one can usually associate with a pair \((f, g)\), with \( f \in \mathcal{V}_{1,1/2}\) and \( g \) is Lipschitz and bounded, a solution \( V[f, g] \) of the PDE which is the unique solution in \( \mathcal{V}_{1,1/2} \) and such that, if \( V^1 = V[f^1, g^1] \) and \( V^2 = V[f^2, g^2] \), then for some \( c > 0 \) not depending on \((f^1, f^2, g^1, g^2)\):

\[
\|V^2 - V^1\|_{\infty} \leq c \left( \|f^2 - f^1\|_{\infty} + \|g^2 - g^1\|_{\infty} \right).
\]

See e.g. Fleming and Soner [26].

**Remark 1.72.** It may happen that the solution is more regular. In some cases including the one of the heat equation, if \( f \in \mathcal{V}_{1,1/2} \) and \( g \) is bounded and has Lipschitz second derivatives, the solution \( V \) is such that \( V_t, V_x, V_{xx} \) are functions in \( \mathcal{V}_{1,1/2} \); see Krylov [34, Thm. 9.1.2]. For financial applications, such hypotheses are however too strong.

**Exercise 1.73.** Consider the heat equation with final condition \( g(x) = x^2 \). Compute \( V(0, t) \) using e.g. \((1.22)\). Deduce that we cannot have in general more than the \( 1/2 \) Hölder exponent in time.

**Remark 1.74.** We will obtain in chapter \([2]\) additional regularity results, needing in particular the diffusion matrix to be uniform elliptic.

We see that we cannot in general apply theorem \([1.68]\) since the solution \( V \) of the PDE is not smooth enough. However we can regularize \( V \) by convolution and see if it is solution of a PDE close to the original one. This is the subject of the next section.

6.3. Regularization by convolution. We next present an approach that allows to obtain error estimates when the solution of the PDE is Hölder, but not necessarily differentiable, as is typically the case in applications.
Let \( \psi \) be a nonnegative \( C^\infty \) function over \( \mathbb{R}^n \) with support in \( B(0,1) \), and such that \( \int \psi(x)dx = 1 \). For \( \varepsilon > 0 \), call the family on functions \( \psi_\varepsilon(x) := \varepsilon^{-n}\psi(x/\varepsilon) \) a regularizing kernel. The function \( \psi_\varepsilon(x) \) has support in \( B(0,\varepsilon) \), and has a unit integral:

\[
\int_{\mathbb{R}^n} \psi_\varepsilon(x)dx = \int_{\mathbb{R}^n} \psi \left( \frac{x}{\varepsilon} \right) d \left( \frac{x}{\varepsilon} \right) = \int_{\mathbb{R}^n} \psi(y)dy = 1.
\]

Let \( w \) be a bounded, Hölder function over \( \mathbb{R}^n \), with Hölder constants \( C_w > 0, \mu_w \in (0,1] \). Its \( \varepsilon \) regularization (by the regularizing kernel \( \psi_\varepsilon \)) is defined as \( w_\varepsilon(x) := \psi_\varepsilon * w \), i.e.

\[
w_\varepsilon(x) = \int_{\mathbb{R}^n} w(y)\psi_\varepsilon(x-y)dy = \int_{\mathbb{R}^n} w(x-y)\psi_\varepsilon(y)dy = \int_{\mathbb{R}^n} w(x-\varepsilon z)\psi(z)dz,
\]

where we have made the change of variables \( z = y/\varepsilon \). Since

\[
w_\varepsilon(x) - w(x) = \int_{\mathbb{R}^n} (w(x-\varepsilon z) - w(x))\psi(z)dz,
\]

it follows that

\[
\| w_\varepsilon \|_\infty \leq \| w \|_\infty \quad \text{and} \quad \| w_\varepsilon - w \|_\infty \leq \varepsilon^{\mu_w} C_w.
\]

Call multiindex an element of \( \mathbb{N}^n \). With each multiindex \( \alpha \) is associated the differential operator

\[
D^\alpha := \frac{\partial^{\alpha_1}}{\partial x_1^{\alpha_1}} \cdots \frac{\partial^{\alpha_n}}{\partial x_n^{\alpha_n}}, \quad \text{of order } |\alpha| := \sum_{i=1}^{n} \alpha_i.
\]

In view of (1.217), \( w_\varepsilon \) is of class \( C^\infty \) and satisfies

\[
D^\alpha w_\varepsilon(x) = \int_{\mathbb{R}^n} w(x-y)D^\alpha \psi_\varepsilon(y)dy, \quad \text{for any multiindex } \alpha.
\]

Since \( D^\alpha \psi_\varepsilon(y) = \varepsilon^{-|\alpha|} D^\alpha \psi(y/\varepsilon) \), with the change of variable \( z = y/\varepsilon \), we get

\[
D^\alpha w_\varepsilon(x) = \varepsilon^{-|\alpha|} \int_{\mathbb{R}^n} w(y)D^\alpha \psi \left( \frac{y}{\varepsilon} \right) d \left( \frac{y}{\varepsilon} \right) = \varepsilon^{-|\alpha|} \int_{B(0,1)} w(x-\varepsilon z)D^\alpha \psi(z)dz.
\]

In addition, if \( \alpha \neq 0 \), \( \int_{B(0,1)} D^\alpha \psi(z)dz = 0 \), and therefore we also have

\[
D^\alpha w_\varepsilon(x) = \varepsilon^{-|\alpha|} \int_{B(0,1)} (w(x-\varepsilon z) - w(x))D^\alpha \psi(z)dz, \quad \text{if } \alpha \neq 0.
\]

We deduce that

\[
\|D^\alpha w_\varepsilon\|_\infty \leq \varepsilon^{\mu_w-|\alpha|} C_w \|D^\alpha \psi\|_1, \quad \text{if } \alpha \neq 0.
\]

**Remark 1.75.** We can refine this estimate for \( \alpha = \beta + \gamma \), with \( \beta \) and \( \gamma \) multiindexes, when \( D^\beta w \) exists and is itself bounded and Hölder. Indeed, since \( D^\beta w_\varepsilon = \psi_\varepsilon * D^\beta w \) and \( D^\alpha w_\varepsilon(x) = D^\gamma \left( D^\beta w_\varepsilon \right) \), applying (1.222) with \( (D^\beta w, \gamma) \) in lieu of \( (w, \alpha) \), we obtain:

\[
D^\alpha w_\varepsilon(x) = \int_{\mathbb{R}^n} D^\beta w(y)D^\gamma \psi_\varepsilon(x-y)dy = \varepsilon^{-|\gamma|} \int_{\mathbb{R}^n} D^\beta w(x-\varepsilon z)D^\gamma \psi(z)dz.
\]

Let us denote by \( (C_{w,\beta}, \mu_{w,\beta}) \) the Hölder constants of \( D^\beta w \). Applying now (1.224) with again \( D^\beta w(y) \) and \( \gamma \) in lieu of \( w(y) \) and \( \alpha \), we get:

\[
\|D^\alpha w_\varepsilon\|_\infty \leq \varepsilon^{\mu_{w,\beta}-|\gamma|} C_{w,\beta} \|D^\gamma \psi\|_1, \quad \text{if } \gamma \neq 0.
\]
Anisotropic spaces. We have seen that we need to consider functions of space and time that are Lipschitz in space and Hölder in time with exponent $\frac{1}{2}$. So consider $\psi : \mathbb{R}^{n+1} \to \mathbb{R}_+$ of class $C^\infty$ (elements of $\mathbb{R}^{n+1}$ denoted by $(x, t)$ with $x \in \mathbb{R}^n$) such that

$$
\int \psi(x)dx = 1; \quad \text{supp} \psi \subset \hat{B},
$$

where

$$
\hat{B} := \{(x, t) \in \mathbb{R}^{n+1}; \quad |x| + \sqrt{|t|} \leq 1/2\}.
$$

Note that $\hat{B} \subset B$ (the Euclidean unit ball). Redefine $\psi_\varepsilon$ and $w_\varepsilon$ as

$$
\psi_\varepsilon(x, t) = \varepsilon^{-n-2}\psi(x/\varepsilon, t/\varepsilon^2); \quad w_\varepsilon := \psi_\varepsilon * w.
$$

Note that $\psi_\varepsilon$ is nonnegative with integral 1, and satisfies

$$
\text{supp} \psi_\varepsilon \subset \{(x, t) \in \mathbb{R}^{n+1}; \quad |x| + \sqrt{|t|} \leq \varepsilon/2\} \subset \varepsilon B.
$$

A multiindex in this setting is of the form $\alpha = (\alpha_x, \alpha_t) \in \mathbb{N}^n \times \mathbb{N}$, with associated differential operator

$$
D^\alpha := \frac{\partial^{\alpha_x}}{\partial x_1^{\alpha_1} \ldots \partial x_n^{\alpha_n}} \left( \frac{\partial^{\alpha_t}}{\partial t^{\alpha_t}} \right).
$$

We assume that $w$ is Lipschitz in space and Hölder with exponent 1/2 in time:

$$
|w(x', t') - w(x, t)| \leq C_w \left( |x' - x| + |t' - t|^{1/2} \right), \quad \text{for all } (x', t') \text{ and } (x, t) \text{ in } \mathbb{R}^{n+1}.
$$

Let us set

$$
w_\varepsilon(x, t) := \int_{\mathbb{R}^{n+1}} w(x - y, t - s) \psi_\varepsilon(y, s)d(y, s).
$$

Setting $z := y/\varepsilon$ and $\tau := s/\varepsilon^2$, we obtain

$$
w_\varepsilon(x, t) = \int_{B(0, 1)} w(x - \varepsilon z, t - \varepsilon^2 \tau) \psi(z, \tau)d(z, \tau).
$$

It follows that

$$
\|w_\varepsilon\|_\infty \leq \|w\|_\infty; \quad \|w_\varepsilon - w\|_\infty \leq C_w \varepsilon.
$$

Similarly, for a multiindex $\alpha \neq 0$ we have that

$$
D^\alpha w_\varepsilon(x, t) = \varepsilon^{-|\alpha_x|-2|\alpha_t|} \int_{\mathbb{R}^{n+1}} w(x - y, t - s) D^\alpha \psi \left( \frac{y}{\varepsilon}, \frac{s}{\varepsilon^2} \right) d \left( \frac{y}{\varepsilon}, \frac{s}{\varepsilon^2} \right)
$$

$$
= \varepsilon^{-|\alpha_x|-2|\alpha_t|} \int_{B(0, 1)} w(x - \varepsilon z, t - \varepsilon^2 \tau) D^\alpha \psi(z, \tau)d(z, \tau),
$$

and again we may write if $\alpha \neq 0$:

$$
D^\alpha w_\varepsilon(x, t) = \varepsilon^{-|\alpha_x|-2|\alpha_t|} \int_{B(0, 1)} (w(x - \varepsilon z, t - \varepsilon^2 \tau) - w(x, t)) D^\alpha \psi(z, \tau)d(z, \tau).
$$

Therefore, denoting by $C_w$ the Hölder coefficient of $w$, we have that, using an adaptation to the present setting of (1.223):

$$
\|D^\alpha w_\varepsilon\|_\infty \leq \varepsilon^{1-|\alpha_x|-2|\alpha_t|} C_w \|D^\alpha \psi\|_1, \quad \text{when } \alpha \neq 0.
$$
6.4. Error estimates in the case of constant coefficients. We next apply the approach of regularization by convolution in the simple case of the PDE (1.25), with constant coefficients $a$, $b$, $r$.

**Theorem 1.76.** Let (1.214)–(1.215) hold, and $(a, b, r)$ be constant. Let $v$ be the solution of the explicit scheme (1.196), such that the monotonicity condition (1.197) holds. Then, for some $c > 0$ not depending on $h$ and $\varepsilon \in (0, 1]$:

$$
\sup_{j,k} |V(x_j, t_k) - v_j^k| \leq c(\varepsilon + \varepsilon^{-3}h_0 + \varepsilon^{-1}h_1 + \varepsilon^{-3}h_1^2).
$$

In particular, when $h_0 = O(h_1^3)$ (monotonicity condition), taking $\varepsilon = O(h_1^{1/2})$ we deduce that

$$
\sup_{j,k} |V(x_j, t_k) - v_j^k| = O(h_1^{1/2}).
$$

**Proof.** It suffices to prove the first statement. We first extend $f$ and $V$ for $t \in [-1, 0)$ by setting $f(x, t) := f(x, 0)$, $t \in [-1, 0) < 0$, and by taking $V$ as the solution of the PDE extended for $t \in [-1, 0)$.

Remember that $\psi_\varepsilon$ was defined in (1.229). We may assume that $\psi(x, t) = 0$ if $t \leq 0$. Then $f^\varepsilon := f * \psi_\varepsilon$ and $V^\varepsilon := V * \psi_\varepsilon$ are well-defined, for $(x, t) \in \mathbb{R} \times [0, T]$ and $\varepsilon \leq 1$. Computing the convolution of both sides of (1.25) with $\psi_\varepsilon$, we find that $V^\varepsilon$ is solution of the perturbed PDE

$$
\begin{align*}
V^\varepsilon_t + bV^\varepsilon_x + \frac{1}{2}aV^\varepsilon_{xx} - rV^\varepsilon + f^\varepsilon &= 0, \quad (x, t) \in \mathbb{R} \times [0, T], \\
V^\varepsilon(x, T) &= g^\varepsilon(x), \quad x \in \mathbb{R},
\end{align*}
$$

where $g^\varepsilon(x) := V^\varepsilon(x, T)$ for all $x \in \mathbb{R}^n$. Since $V$ in $V_{1,1/2}$, we easily get that

$$
\|f^\varepsilon - f\|_\infty + \|g^\varepsilon - g\|_\infty = O(\varepsilon).
$$

By (1.215), we have that

$$
\|V - V^\varepsilon\|_\infty = O(\varepsilon).
$$

Next, denote by $v^\varepsilon$ the solution of the scheme (1.196), with $(f^\varepsilon, g^\varepsilon)$ in lieu of $(f, g)$. By the $\ell^\infty$ estimates for this Markovian scheme, combined with (1.242), we know that

$$
\sup_{j,k} |v_j^k - v_j^{\varepsilon,k}| \leq \|g^\varepsilon - g\|_\infty + T |f^\varepsilon - f|_\infty = O(\varepsilon).
$$

In view of the above inequalities and of the triangle inequality

$$
\sup_{j,k} |V(x_j, t_k) - v_j^k| \leq \|V - V^\varepsilon\|_\infty + \sup_{j,k} |V^\varepsilon(x_j, t_k) - v_j^{\varepsilon,k}| + \sup_{j,k} |v_j^{\varepsilon,k} - v_j^k|,
$$

it suffices to check that the second term in the above r.h.s. is of order of the r.h.s. of (1.239). We apply theorem 1.68 to the scheme with perturbed data. The norms of partial derivatives of $V^\varepsilon$ are estimated thanks to (1.238):

$$
\|D_x^2 V^\varepsilon\|_\infty = O(\varepsilon^{-3}); \quad \|D_x^2 V^\varepsilon\|_\infty = O(\varepsilon^{-1}); \quad \|D_x^4 V^\varepsilon\|_\infty = O(\varepsilon^{-3}).
$$

The conclusion follows. \hfill \Box
6.5. Error estimates in the case of a constant diffusion matrix. We will now extend the theory of error estimates based on the regularization by convolution in the case of variable coefficients. We will use, however, the important restriction that the diffusion coefficient is assumed to be constant.

**Theorem 1.77.** Let \((1.244)-(1.245)\) hold, \(b, r\) belong to \(V^{1/2}\), and the diffusion matrix \(a\) be constant. Let \(v\) be the solution of the explicit scheme \((1.94)\). Then again \((1.239)-(1.240)\) hold.

**Proof.** The proof is similar to the one of theorem \(1.76\) with some minor adaptations. Again only \((1.239)\) needs to be proved. We first extend \(b, r, f\) for \(t \in [-1, 0]\) by setting

\[
\phi(x, t) := \phi(x, 0), \quad t < 0, \quad \text{for } \phi = b, r, f,
\]

and by taking \(V\) as the solution of the PDE extended for \(t \in [-1, 0]\). We may assume that \(\psi(x, t) = 0\) if \(t \leq 0\). Then \(f^\varepsilon := f \ast \psi_\varepsilon\) and \(V^\varepsilon := f \ast \psi_\varepsilon\) are well-defined, for \((x, t) \in \mathbb{R} \times [0, T]\). Applying the convolution with \(\psi_\varepsilon\) on both sides of the first relation in \((1.25)\), we obtain that \(V^\varepsilon\) is solution of

\[
(1.248) \begin{cases}
V^\varepsilon_x + b(x, t)V^\varepsilon_x + \frac{1}{2}aV^\varepsilon_{xx} - r(x, t)V^\varepsilon + \hat{f}^\varepsilon(x, t) = 0, & (x, t) \in \mathbb{R} \times [0, T], \\
V^\varepsilon(x, T) = \hat{g}(x), & x \in \mathbb{R}.
\end{cases}
\]

where \(g_\varepsilon(x) := V^\varepsilon(x, T)\), the source term \(\hat{f}^\varepsilon(x, t)\) taking into account the variations of \(b\) and \(r\):

\[
(1.249) \quad \hat{f}^\varepsilon(x, t) := f^\varepsilon(x, t) + e_1^\varepsilon(x, t) - e_2^\varepsilon(x, t),
\]

with \(e_1^\varepsilon = (bV_x) \ast \psi_\varepsilon - b(V_x \ast \psi_\varepsilon)\) and \(e_2^\varepsilon = (rV) \ast \psi_\varepsilon - r(V \ast \psi_\varepsilon)\), i.e.,

\[
(1.250) \begin{cases}
e_1^\varepsilon(x, t) := \int \psi_\varepsilon(y, s)(b(x - y, t - s) - b(x, t))V_x(x - y, t - s)dyds, \\
e_2^\varepsilon(x, t) := \int \psi_\varepsilon(y, s)(r(x - y, t - s) - r(x, t))V(x - y, t - s)dyds.
\end{cases}
\]

Since \(V\) and \(V_x\) are bounded, it follows that \(\|e_1^\varepsilon\|_\infty + \|e_2^\varepsilon\|_\infty = O(\varepsilon)\), and hence, \(\|\hat{f}^\varepsilon - f\|_\infty + \|g^\varepsilon - g\|_\infty = O(\varepsilon)\). We conclude the proof as we did for the one of theorem \(1.76\).

**Exercise 1.78.** Derive similar estimates for the implicit or semi implicit schemes, and for problems with space dimension greater than one.

**Remark 1.79.** The above method has a straightforward extension to the case when the diffusion coefficient is variable, whenever the solution of a PDE has a Lipschitz spatial gradient (i.e., when \(V_{xx}\) is essentially bounded). However, this assumption is not always satisfied in applications. For variable diffusion coefficients one uses the “coefficients shaking” method introduced by Krylov, in the context of HJB equations; see e.g. Barles and Jakobsen [5], Bonnans et al. [11], Krylov [36].

6.6. Centered scheme. We now study the centered scheme \((1.98)\), assuming that \((1.101)\) holds, so that this scheme is simply monotonic by lemma \(1.27\). We briefly indicate where the changes in the analysis. Form the Taylor expansions (skipping the time argument)

\[
(1.251) \quad V(x \pm h_1) = V(x) \pm h_1 V_x(x) + h_1^2 \int_0^1 (1 - s)V_{xx}(x \pm sh_1)ds
\]
we deduce that
\[
\left| \frac{V(x+h_1)-V(x-h_1)}{2h_1} - V_x(x) \right| \leq h_1 \int_0^1 (1-s)|V_{xx}(x+sh_1) - V_{xx}(x-sh_1)|\,ds \\
\leq h^2_1 \|V_{xxx}\|_\infty.
\]
Following otherwise the analysis of the upwind scheme (theorem 1.68) we deduce that:

**Theorem 1.80.** For the centered scheme (1.98), assuming (1.101) holds and the solution \( V \) of the PDE to be smooth enough, we have the following error estimate:

\[
\sup_{j,k} \left| V(x_j,t_k) - v_j^k \right| \leq T \left( \|V_t\|_\infty h_0 + \|b\|_\infty \|V_{xxx}\|_\infty h_1^2 + \frac{1}{2} \|a\|_\infty \|V_{xxx}\|_\infty h_1^2 \right).
\]

If \( V \) is Hölder, following the analysis of theorem 1.76 we deduce that:

**Theorem 1.81.** Let (1.214)-(1.215) hold, and \((a,b,r)\) be constant. Let \( v \) be the solution of the centered scheme (1.98), assuming that (1.101) holds. Then, for some \( c > 0 \) not depending on \( h \) and \( \varepsilon \in (0,1) \):

\[
\sup_{j,k} \left| V(x_j,t_k) - v_j^k \right| \leq c \left( \varepsilon + \varepsilon^{-3} h_0 + \varepsilon^{-2} h_1^2 + \varepsilon^{-3} h_1^2 \right).
\]

In particular, when \( h_0 = O(h_1^2) \) (monotonicity condition), taking \( \varepsilon = O(h_1^{1/2}) \) we deduce that

\[
\sup_{j,k} \left| V(x_j,t_k) - v_j^k \right| = O \left( h_1^{1/2} \right).
\]

**Exercise 1.82.** Extend this analysis to the implicit centered scheme (1.107).

### 7. Discrete parabolic energy estimates

#### 7.1. Overview.
It is useful to check the stability of the solution of the numerical schemes in (weighted) \( L^2 \) norms. This is a desirable property, especially in the case of \( \theta \) schemes, or of some schemes for dimension greater than 1, for which simple monotonicity does not always hold. Finally, the notions introduced in this section give an introduction to the concepts of chapter 2. So, consider the following PDE

\[
V_t(x,t) + b(x,t)V_x(x,t) + \frac{1}{2}a(x,t)V_{xx}(x,t) = 0,
\]
with \( x \in \mathbb{R}, \ t \in [0,T], \ a(x,t) > 0 \), and the final condition

\[
V(x,T) = g(x), \quad x \in \mathbb{R}.
\]
We will recall the classical energy inequalities in the case of constant coefficients, and extend them to the case when \( a \) and \( b \) are Lipschitz functions of the space variables. We will conclude the section with the analysis of \( \theta \) schemes, and show that they satisfy energy estimates when \( \theta \geq 1/2 \).

As usual we denote by \( h_0 > 0 \) and \( h_1 > 0 \) the time and space steps, with \( h_0 = T/N, \ N \in \mathbb{N} \). We set \( x_j = jh_1, \ j \in \mathbb{Z} \), and \( t_k = kh_0, \ k = 0 \) to \( N \). The algorithm approximates \( V(x_j,t_k) \) by \( u_j^k \).
Over the space $\ell^2_\mathbb{Z}$ (set of sequences $(v_j)_{j \in \mathbb{Z}}$ with summable squares) we define the (scaled) norm and associated scalar product by
\[
\|v\|^2_h := h_1 \sum_j (v_j)^2; \quad (u, v)_h := h_1 \sum_j u_j v_j.
\]
We denote by $\ell^2_\mathbb{Z}$ the space $\ell^2_\mathbb{Z}$ endowed with the above norm, and assume that there exists $C_g > 0$ such that
\[
h_1 \sum_j |g(x_j)|^2 \leq C_g^2 < \infty, \quad \text{for all } h_1 > 0.
\]
The operators $\delta_1$, $D_h$ and $\Delta_h$ (discrete unit translation, gradient and Laplacian) are defined by
\[
(\delta_1 v)_j = v_{j+1}; \quad (D_h v)_j := \frac{v_{j+1} - v_j}{h_1}; \quad (\Delta_h v)_j := \frac{v_{j+1} + v_{j-1} - 2v_j}{h_1^2}.
\]
We adopt the usual abuse of notation $D_h v_j := (D_h v)_j$ and $\Delta_h v_j := (\Delta_h v)_j$. For $u$, $v$ in $\ell^2_\mathbb{Z}$ we easily check the discrete integration by parts formula
\[
-h_1 \sum_j (\Delta_h u)_j v_j = (D_h u, D_h v)_h, \quad \text{for all } u, v \text{ in } \ell^2_\mathbb{Z}.
\]

### 7.2. Constant coefficients.

In this section we assume that $a \geq 0$ and $b$ are constant; we may assume that $b \geq 0$. Consider the standard implicit finite differences (upwind) scheme
\[
\frac{u_j^{k+1} - u_j^k}{h_0} + bD_h u_j^k + \frac{1}{2} a \Delta_h u_j^k + f_j^k = 0; \quad j \in \mathbb{Z}, \ k = 0, \ldots, N - 1,
\]
with final condition
\[
u_j^N = g(x_j), \quad j \in \mathbb{Z}.
\]
This scheme is known to be simply monotonic.

#### 7.2.1. First discrete parabolic energy estimates.

We start by the simple case when the source term is equal to 0.

**Lemma 1.83.** Assume that $f = 0$. Then the solution of (1.261) satisfies
\[
\|u^k\|^2_h + h_0 a \|D_h u^k\|^2_h \leq \|u^{k+1}\|^2_h
\]
and
\[
\max_k \left( \max \left\{ \max_j \|u^k\|^2_h, h_0 a \sum_{k=0}^{N-1} \|D_h u^k\|^2_h \right\} \right) \leq \|u^N\|^2_h.
\]

**Proof.** Multiplying (1.261) by $h_0$, we obtain
\[
u_j^k - \frac{1}{2} h_0 a \Delta_h u_j^k = u_j^{k+1} + h_0 b D_h u_j^k, \quad j \in \mathbb{Z}, \ k = 0, \ldots, N - 1.
\]
Multiplying by $h_1 u_j^k$ and summing over $j$, we get with (1.260)
\[
\|u^k\|^2_h + \frac{1}{2} h_0 a \|D_h u^k\|^2_h \leq (u^{k+1}, u^k)_h + h_0 \frac{h_0}{h_1} (\delta_1 u^k - u^k, u^k)_h.
\]
Now \( \| \delta_1 u^k \|_h = \| u^k \|_h \), and so by the Cauchy-Schwarz inequality
\[
\begin{align*}
(\delta_1 u^k - u^k, u^k)_h & \leq \| \delta_1 u^k \|_h \| u^k \|_h - \| u^k \|^2_h = 0.
\end{align*}
\]
By the Young inequality,
\[
(1.268) \quad (u^{k+1}, u^k)_h \leq \frac{1}{2} (\| u^k \|^2_h + \| u^{k+1} \|^2_h).
\]
Combining relations (1.266)-(1.268), we get (1.263). Summing from \( k \) to \( N-1 \), we obtain
\[
(1.269) \quad \| u^k \|^2_h + h_0 a \sum_{\ell=k}^{N-1} \| D_h u^\ell \|^2_h \leq \| u^N \|^2_h,
\]
from which (1.264) easily follows.

In the sequel we need the following technical lemma;

**Lemma 1.84.** Let \( c > 0 \). For all \( \varepsilon > 0 \), For \( h_0 > 0 \) is small enough, then
\[
(1.270) \quad (1 - ch_0)^{-N} \leq e^{(c+\varepsilon)T}.
\]

**Proof.** For \( h_0 > 0 \) is small enough, we have that \( \log(1 - ch_0) \geq -(c+\varepsilon)h_0 \), and so,
\[
(1.271) \quad (1 - ch_0)^{-N} = e^{-N\log(1-ch_0)} \leq e^{N(c+\varepsilon)h_0} = e^{(c+\varepsilon)T},
\]
as was to be proved.

**Lemma 1.85.** Let \( f^k \in \ell^2_h \) for all \( k \) and \( \varepsilon > 0 \). If \( h_0 \) is small enough, the solution of (1.261) satisfies
\[
(1.272) \quad \max_k \| u^k \|^2_h + h_0 a \sum_{k=0}^{N-1} \| D_h u^k \|^2_h \leq e^{T+\varepsilon} \left( \| u^N \|^2_h + h_0 \sum_{k=0}^{n-1} \| f^k \|^2_h \right).
\]

**Proof.** Repeating the arguments of the proof of lemma 1.85 we obtain that
\[
(1.273) \quad \| u^k \|^2_h + h_0 a \| D_h u^k \|^2_h \leq \| u^{k+1} \|^2_h + 2h_0 (f^k, u^k)_h \\
\leq \| u^{k+1} \|^2_h + h_0 (\| f^k \|^2_h + \| u^k \|^2_h),
\]
so that in particular
\[
(1.274) \quad (1 - h_0) \| u^k \|^2_h \leq \| u^{k+1} \|^2_h + h_0 \| f^k \|^2_h,
\]
and so with lemma 1.84 when \( h_0 > 0 \) is small enough:
\[
(1.275) \quad \| u^k \|^2_h \leq (1 - h_0)^{k-N} \left( \| u^N \|^2_h + h_0 \sum_{\ell=k}^{N-1} \| f^\ell \|^2_h \right), \\
\leq e^{T+\varepsilon} \left( \| u^N \|^2_h + h_0 \sum_{\ell=0}^{N-1} \| f^\ell \|^2_h \right),
\]
as was to be proved.

### 7.2.2. A refined first discrete parabolic energy estimates

Consider the function space
\[
(1.276) \quad V_h := \{ u \in \ell^2_h ; \quad D_h u \in \ell^2_h \},
\]
which is an Hilbert space when endowed with the norm
\[
(1.277) \quad \| u \|_{V_h} := (\| u \|^2_h + \| D_h u \|^2_h)^{1/2}.
\]
We can obtain an estimate of the l.h.s. of (1.272) by assuming only that \( f^k \) belongs to the dual of \( V_h \) for all \( k \). (rather than to \( \ell^2_h \) and of course that \( u^N \in \ell^2_h \). We leave the proof as an exercise.
7.2.3. Second discrete parabolic energy estimates.

Lemma 1.86. Assume that \( f^k \in \ell_2^2 \) for all \( k \), and \( u^N \in V_h \). Then the solution of (1.261) satisfies for some \( c > 0 \), not depending on \( (f, g) \):

\[
\max_k \| D_h u^k \|_h + h_0^{-1} \sum_{k=0}^{N-1} \| u^{k+1} - u^k \|_h^2 \leq c \left( \| u^N \|_V^2 + h_0 \sum_{k=0}^{n-1} \| f^k \|_h^2 \right).
\]

Proof. Multiplying (1.261) by \((u_j^{k+1} - u_j^k)\), summing over \( j \) and using (1.260), we obtain that

\[
\frac{1}{h_0} \| u_j^{k+1} - u_j^k \|_h^2 + \frac{a}{2} \| D_h u_j^k \|_h^2 = \frac{1}{2} a (D_h u_j^k, D_h u_j^{k+1})_h + (f^k, u_j^k - u_j^{k+1})_h + b(D_h u_j^k, u_j^k - u_j^{k+1})_h.
\]

Use

\[
(D_h u_j^k, D_h u_j^{k+1})_h \leq \frac{1}{2} (\| D_h u_j^k \|_h^2 + \| D_h u_j^{k+1} \|_h^2),
\]

\[
(f^k, u_j^k - u_j^{k+1})_h \leq \frac{1}{2} h_0 \| f^k \|_h^2 + \frac{1}{2 h_0} \| u_j^k - u_j^{k+1} \|_h^2,
\]

\[
b(D_h u_j^k, u_j^k - u_j^{k+1})_h \leq h_0 b^2 \| D_h u_j^k \|_h^2 + \frac{1}{4 h_0} \| u_j^k - u_j^{k+1} \|_h^2.
\]

It follows that

\[
\frac{1}{4 h_0} \| u_j^{k+1} - u_j^k \|_h^2 + \left( \frac{a}{4} - h_0 b^2 \right) \| D_h u_j^k \|_h^2 \leq \frac{a}{4} \| D_h u_j^{k+1} \|_h^2 + \frac{1}{2 h_0} \| f^k \|_h^2.
\]

Summing inequalities (1.283) from \( k \) to \( N - 1 \), we get

\[
\frac{1}{4 h_0} \sum_{\ell=k}^{N-1} \| u_j^{\ell+1} - u_j^{\ell} \|_h^2 + \frac{a}{4} \| D_h u_j^{\ell} \|_h^2 \leq \frac{a}{4} \| D_h u_j^N \|_h^2 + h_0 b^2 \sum_{\ell=k}^{N-1} \| D_h u_j^{\ell} \|_h^2 + \frac{1}{2 h_0} \sum_{\ell=k}^{N-1} \| f^\ell \|_h^2.
\]

The conclusion follows since we can estimate the r.h.s. thanks to lemma 1.272 \( \square \)

7.3. Varying coefficients. We now assume that \( \inf a = \alpha > 0 \) and that \( a = a(x, t) \) and \( b = b(x, t) \) are uniformly Lipschitz w.r.t. the space variable: for all \( t \in [0, T] \) and \( x, x' \) in \( \mathbb{R} \), we have that

\[
\inf a = \alpha > 0, \quad \text{and for all } x, x' \text{ in } \mathbb{R}:
\]

\[
|a(x', t) - a(x, t)| \leq L_a |x' - x|,
\]

\[
|b(x', t) - b(x, t)| \leq L_b |x' - x|.
\]

Setting \( a_j^k = a(x_j, t_k) \) and \( b_j^k = b(x_j, t_k) \), the expression of the upwind implicit discretization scheme is now

\[
\frac{u_j^{k+1} - u_j^k}{h_0} + (b_j^k)_+ D_h u_j^k + (b_j^k)_- D_h u_{j-1}^k + \frac{1}{2} a_j^k \Delta_h u_j^k = 0; \quad j \in \mathbb{Z}, \quad k = 0, \ldots, N-1,
\]

with the same final condition (1.262).
Lemma 1.87. Set $c_{ab} := \frac{1}{2}(L_a)^2/\alpha + 2L_a$. Let $\varepsilon > 0$. If $(h_0, h_1)$ is small enough, then the solution of the scheme (1.285) satisfies

$$
\begin{align*}
&\text{(i)} \quad \max_k \|u^k\|^2_h \leq e^{c_{ab}(\varepsilon + \varepsilon^2)} \|u^N\|^2_h, \\
&\text{(ii)} \quad \frac{1}{2}h_0^2 \sum_{k=0}^{N-1} \|D_h u^k\|^2_h \leq \left(1 + Tc_{ab}e^{c_{ab}(\varepsilon + \varepsilon^2)}\right) \|u^N\|^2_h.
\end{align*}
$$

Proof. Multiplying (1.285) by $2h_0h_1u^k_j$ and summing over $j$, we obtain with (1.268):

$$
\|u^k\|^2_h - h_0h_1 \sum_j a^j_k u^k_j \Delta_h u^k_j \leq \|u^{k+1}\|^2_h + 2h_0h_1 \sum_j \left((b^j_k)^+D_h u^k_j + (b^j_k)^-D_h u^k_{j-1}\right) u^k_j.
$$

Use

$$
-h_1 \sum_j a^j_k \Delta_h u^k_j \leq -\sum_j (D_h u^k_j - D_h u^k_{j-1})a^j_k u^k_j,
$$

$$
= \sum_j D_h u^k_j (a^j_{j+1} u_{j+1}^k - a^j_k u^k_j),
$$

$$
= h_1 \sum j a^j_k |D_h u^k_j|^2 + h_1 \sum_j \frac{a^j_{j+1} - a^j_k}{h_1} D_h u^k_j u^k_{j+1},
$$

$$
\geq \alpha \|D_h u^k\|^2_h - L_a \|D_h u^k\|_h \|u^k\|_h.
$$

Also, since $2(\gamma - \delta) \leq \gamma^2 - \delta^2$ for all $\gamma$, $\delta$:

$$
2 \sum_j (b^j_k)^+(u^k_{j+1} - u^k_j) u^k_j \leq \sum_j (b^j_k)^+ \left((u^k_{j+1})^2 - (u^k_j)^2\right),
$$

$$
= \sum_j \left( (b^j_{j-1})^+ - (b^j_k)^+ \right) (u^k_j)^2,
$$

$$
2 \sum_j (b^j_k)^-(u^k_j - u^k_{j-1}) u^k_j \leq 2 \sum_j \left| (b^j_k)^- \right| \left((u^k_{j-1})^2 - (u^k_j)^2\right),
$$

$$
= \sum_j \left( \left| (b^j_{j+1})^- \right| - \left| (b^j_k)^- \right| \right) (u^k_j)^2.
$$

Combining the above inequalities, we get

$$
\|u^k\|^2_h + \alpha h_0 \|D_h u^k\|^2_h \leq \|u^{k+1}\|^2_h + h_0 L_a \|D_h u^k\|_h \|u^k\|_h + 2h_0 L_a \|u^k\|^2_h.
$$

Since

$$
L_a \|D_h u^k\|_h \|u^k\|_h \leq \frac{1}{2} \alpha \|D_h u^k\|^2_h + \frac{(L_a)^2}{2 \alpha} \|u^k\|^2_h,
$$

we obtain that

$$
(1 - h_0 c_{ab}) \|u^k\|^2_h + \frac{1}{2} \alpha h_0 \|D_h u^k\|^2_h \leq \|u^{k+1}\|^2_h.
$$

In particular, for $(h_0, h_1)$ small enough, we obtain by with lemma 1.84 that

$$
\|u^k\|^2_h \leq (1 - h_0 c_{ab})^{-N} \|u^N\|^2_h \leq e^{c_{ab}(\varepsilon + \varepsilon^2)} \|u^N\|^2_h.
$$
That is, (1.286)(i) holds. Summing (1.293) over \( k \), we obtain
\[
(1.295) \quad \|u_0\|_h^2 + \frac{1}{2} \alpha h_0 \sum_{k=0}^{N-1} \|D_h u_k\|_h^2 \leq \|u_N\|_h^2 + h_0 c_{ab} \sum_{k=0}^{N-1} \|u_k\|_h^2 \leq \|u_N\|_h^2 + h_0 c_{ab} \sup_k \|u_k\|_h^2,
\]
and the conclusion follows with (1.286)(i).

**7.4. Centered schemes.** Assume now that, instead of (1.285), we use the implicit centered scheme
\[
(1.296) \quad \frac{u_{j+1}^k - u_j^k}{h_0} + b_j^k \frac{u_{j+1}^k - u_{j-1}^k}{2h_1} + \frac{1}{2} a_j^k \Delta_h u_j^k = 0; \quad j \in \mathbb{Z}, \ k = 0, \ldots, N - 1,
\]
with the same final condition (1.262).

**Lemma 1.88.** Let (1.284) hold. Then there exists \( C > 0 \) such that, if \( (h_0, h_1) \) is small enough, we have that
\[
(1.297) \quad \max_k \left( \max_{h_0} \|u^k\|_h^2, h_0 \sum_{k=0}^{N-1} \|D_h u^k\|_h^2 \right) \leq C \|u_N\|_h^2.
\]

**Proof.** The proof is similar to the one of lemma 1.87. It suffices to analyze the contribution of the first order term when multiplying (1.296) by \( u_j^k \) and summing over \( j \).

We have that
\[
(1.298) \quad h_1 \sum_j b_j^k \frac{u_{j+1}^k - u_{j-1}^k}{2h_1} u_j^k \leq \frac{1}{2} h_1 \|b\|_\infty \sum_j \left( \frac{|u_{j+1}^k - u_j^k|}{h_1} + \frac{|u_j^k - u_{j-1}^k|}{h_1} \right) |u_j^k| \\
\leq \|b\|_\infty \|D_h u_j^k\|_h \|u^k\|_h \\
\leq \frac{1}{4} \alpha \|D_h u_j^k\|_h^2 + \frac{1}{2} \|b\|_\infty \|u^k\|_h^2.
\]
The conclusion follows in the same way as before.

**7.5. \( \theta \) schemes.** We now study the energy estimates for the \( \theta \) schemes introduced in section 2.7. For the sake of simplicity we limit ourself to the case of constant coefficients \( b = 0 \) and \( a > 0 \). Let \( \theta \in [0, 1] \). The \( \theta \)-scheme may be written in the form
\[
(1.299) \quad \frac{u_{j+1}^{k+1} - u_j^k}{h_0} + \frac{1}{2} a \Delta_h w_j^k = 0, \quad \text{where} \quad w_j^k := \theta u_j^k + (1 - \theta) u_{j+1}^{k+1},
\]
with final condition
\[
(1.300) \quad u_j^N = g(x_j), \quad j \in \mathbb{Z}.
\]

**Lemma 1.89.** There exists \( C > 0 \) such that, if \( \theta \geq \frac{1}{2} \) and \( (h_0, h_1) \) is small enough, then
\[
(1.301) \quad \max_k \left( \max_{h_0} \|u^k\|_h^2, h_0 \sum_{k=0}^{N-1} \|D_h w^k\|_h^2 \right) \leq C \|u_N\|_h^2.
\]
Proof. Multiplying (1.299) by $h_0 h_1 w_j^k$ and summing over $j$, we get

\begin{equation}
(u^k, w^k)_h + \frac{1}{2} a h_0 \| D_h w^k \|^2_h = (u^{k+1}, w^k)_h.
\end{equation}

Substituting $w^k = \theta u^k + (1 - \theta) u_j^{k+1}$, we obtain

\begin{equation}
\theta \| u^k \|^2_h + \frac{1}{2} a h_0 \| D_h u^{k+1} \|^2_h = (1 - \theta) \| u^{k+1} \|^2_h + (2 \theta - 1) (u^k, u^{k+1})_h.
\end{equation}

Using $(u^k, u^{k+1})_h \leq \frac{1}{2} (\| u^k \|^2_h + \| u^{k+1} \|^2_h)$, for $\theta \geq \frac{1}{2}$, it follows that:

\begin{equation}
\| u^k \|^2_h + a h_0 \| D_h u^{k+1} \|^2_h \leq \| u^{k+1} \|^2_h.
\end{equation}

We conclude as we did in the proof of lemma 1.85. \hfill \square

Remark 1.90. As discussed in remark 2.7, the simple monotonicity of the $\theta$ scheme is obtained under quite restrictive conditions. So the above result is important since it gives a sound basis for this method when $\theta \in [\frac{1}{2}, 1]$.

Exercice 1.91. Extend the analysis to the case when $b \neq 0$.

Exercice 1.92. Extend the analysis to the case when $a$ and $b$ are not constant.

7.6. Generalized finite differences. The space $\ell^2_{\mathbb{Z}^n}$ is endowed with the norm $\| w \|_2 := \left( \sum_{j \in \mathbb{Z}^n} |w_j|^2 \right)^{1/2}$. We study the implicit generalized finite differences algorithm (1.168) without the source term that adds no essential difficulty, that we write in the form

\begin{equation}
v_j^k = \frac{1}{2} h_0 \sum_{\xi \in \Xi} \eta_{j, \xi}^k \left( v_{j+\xi}^k + v_{j-\xi}^k - 2 v_j^k \right) + v_j^{k+1}, \quad j \in \mathbb{Z}^n, \quad k = 0 : N - 1.
\end{equation}

We need the following condition for the coefficients of the decomposition:

\begin{equation}
\left| \eta_{j, \xi}^k - \eta_{j+\xi, \xi}^k \right| \leq L_{\xi} \left( \eta_{j, \xi}^k \right)^{1/2}, \quad \text{for all } j \in \mathbb{Z}^n \text{ and } \xi \in \Xi.
\end{equation}

Remark 1.93. For a one dimensional problem the stencil reduces to $\xi = 1$ and $\eta_{j, 1}^k = a_j^k / h_1^2$. Then condition (1.306) is equivalent to the Lipschitz condition in (1.284).

Lemma 1.94. If (1.306) holds, then there exists $c > 0$, not depending on $v^N$, such that

\begin{equation}
\max_k \| v^k \|^2_2 + h_0 \sum_{k=0}^{N-1} \sum_{j \in \mathbb{Z}^n} \eta_{j, \xi}^k \left( v_{j+\xi}^k - v_j^k \right)^2 \leq c \| v^N \|^2_2.
\end{equation}

Proof. Multiplying (1.305) by $2 v_j^k$, summing over $j$ and using the Cauchy-Schwarz inequality in $\ell^2_{\mathbb{Z}^n}$, we obtain:

\begin{equation}
\| v^k \|^2_2 \leq h_0 \sum_{j \in \mathbb{Z}^n} \sum_{\xi \in \Xi} \eta_{j, \xi}^k \left( v_{j+\xi}^k + v_{j-\xi}^k - 2 v_j^k \right) v_j^k + \| v^{k+1} \|^2_2.
\end{equation}
Given $\xi \in \Xi$ and $j \in \mathbb{Z}$, setting $j(i) := j + i\xi$, we have that:

(1.309) \[ A_j := \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i)+\xi} + v^k_{j(i)-\xi} - 2v^k_{j(i)} \right) v^k_{j(i)} \]

\[ = \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} + v^k_{j(i-1)} - 2v^k_{j(i)} \right) v^k_{j(i)} \]

\[ = \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right) v^k_{j(i)} - \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i)} - v^k_{j(i-1)} \right) v^k_{j(i)} \]

\[ = \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right) v^k_{j(i)} - \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right) v^k_{j(i+1)} \]

\[ = -\sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right)^2 + \sum_{i \in \mathbb{Z}} \left( \eta^k_{j(i), \xi} - \eta^k_{j(i+1), \xi} \right) \left( v^k_{j(i+1)} - v^k_{j(i)} \right) v^k_{j(i+1)}. \]

Using now (1.306), we get that

(1.310) \[ \left( \eta^k_{j(i), \xi} - \eta^k_{j(i+1), \xi} \right) \left( v^k_{j(i+1)} - v^k_{j(i)} \right) v^k_{j(i+1)} \leq \frac{1}{2} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right)^2 + \frac{1}{2} L^2 \xi_0 v^k_{j(i)} \]

and so

(1.311) \[ A_j \leq -\frac{1}{2} \sum_{i \in \mathbb{Z}} \eta^k_{j(i), \xi} \left( v^k_{j(i+1)} - v^k_{j(i)} \right)^2 + \frac{1}{2} L^2 \xi_0 \sum_{i \in \mathbb{Z}} v^k_{j(i)} \]

Given $\xi \in \Xi$, fix $q$ such that $\xi_q \neq 0$ and set

(1.312) \[ H_q := \{ j \in \mathbb{Z}^n; 0 \leq j_q < \xi_q \}. \]

Then $\mathbb{Z}^n = \bigcup \{ j(i); i \in \mathbb{Z}, j \in H_q \}$, and so,

(1.313) \[ \sum_{j \in \mathbb{Z}^n} \eta^k_{j, \xi} \left( v^k_{j+\xi} + v^k_{j-\xi} - 2v^k_j \right) v^k_j \]

\[ = \sum_{j \in \mathbb{Z}^n} A_j \leq -\frac{1}{2} \sum_{j \in \mathbb{Z}^n} \eta^k_{j, \xi} \left( v^k_{j+\xi} - v^k_{j(i)} \right)^2 + \frac{1}{2} L^2 \xi_0 \sum_{i \in \mathbb{Z}} \| v^k \|^2 \]

Setting $C_\Xi := \frac{1}{2} \sum_{\xi \in \Xi} L^2 \xi_0$, we deduce with (1.308) that

(1.314) \[ (1 - h_0 C_\Xi) \| v^k \|^2 + \frac{1}{2} h_0 \sum_{j \in \mathbb{Z}^n} \eta^k_{j, \xi} \left( v^k_{j+\xi} - v^k_j \right)^2 \leq \| v^{k+1} \|^2, \]

and the conclusion follows using classical techniques.

We next establish some elementary properties of the solution of this PDE, whenever it exists.

8. Scaling

When the final condition has exponential growth, as in the case of the (reformulation of the) European option, the theory of the past sections, cannot be applied since it uses spaces of either bounded of square summable functions. We will see how to scale properly the problem in order to derive results on well-posedness of the numerical scheme. For the sake of simplicity we confine ourself to the case of the heat equation

(1.315) \[ \ddot{u}_t(x,t) + \frac{1}{2} \ddot{u}_{xx}(x,t) = 0, \quad x \in \mathbb{R}, \quad t \in [0, T], \]
with final condition
\[ \bar{u}(x, T) = g(x), \quad x \in \mathbb{R}, \]
where \( g \) has at most exponential growth:
\[ |g(x)| \leq c_1 e^{c_2 |x|}, \quad \text{for all } x \in \mathbb{R}, \]
for some \( c_1 > 0 \) and \( c_2 > 0 \). A possibility (in the more general case when \( x \in \mathbb{R}^n \)) is to consider \textit{weighted} \( L^2 \) \textit{spaces} of the form
\[ L^{2, \rho}(\mathbb{R}^n) := \{ f : \mathbb{R}^n \to \mathbb{R} \text{ measurable}; \int_{\mathbb{R}^n} f^2(x)\rho(x)dx < \infty \}, \]
where \( \rho \) is a positive measurable function defined on \( \mathbb{R}^n \), i.e., \( \rho(x) > 0 \) a.e., endowed with the norm
\[ \|f\|_{2, \rho}^2 := \int_{\mathbb{R}^n} f^2(x)\rho(x)dx. \]
We say that \( \rho \) is \textit{adapted} to \( g \) if \( g \in L^{2, \rho}(\mathbb{R}) \), i.e., if \( \int_{\mathbb{R}} g^2(x)\rho(x)dx < \infty \). In the analysis we need the following properties:

**Definition 1.95.** We say that \( \rho : \mathbb{R}^n \to \mathbb{R} \) is a \textit{balanced weight function} if it of class \( C^2 \), with positive values, and such that
\[ \rho(x) - 1 D\rho(x) \quad \text{and} \quad \rho(x)^{-1} D^2\rho(x) \]
are Lipschitz and bounded.

**Example 1.96.** A possible choice of weight function, adapted to the case of exponential growth as in (1.317), is \( \rho(x) = e^{-\mu \varphi(x)} \), for \( \mu > 0 \) large enough, with \( \varphi(x) = (|x|^2 + 1)^{1/2} \). Then (remembering that \( D\varphi(x) \) is an horizontal vector)
\[ \left\{ \begin{array}{l} \rho(x)^{-1} D\rho(x) = -\mu D\varphi(x); \\ \rho(x)^{-1} D^2\rho(x) = \mu^2 D\varphi(x) D\varphi(x)^\top - \mu D^2\varphi(x). \end{array} \right. \]
Relation (1.320) is satisfied, since \( D\varphi(x) \) and \( D^2\varphi(x) \) are Lipschitz and bounded.

One may think of various other examples of balanced weight functions. The following lemma may help in designing new ones.

**Lemma 1.97.** Let \( \rho \) and \( \eta \) be balanced weight functions over \( \mathbb{R}^n \), and \( \alpha \in \mathbb{R}, \alpha \neq 0 \). Then \( \psi := \rho \eta \) and \( \varphi = \rho^\alpha \) are balanced weight functions.

**Proof.** We have that
\[ \frac{D\psi}{\psi} = \frac{D\rho}{\rho} + \frac{D\eta}{\eta}; \quad \frac{D^2\psi}{\psi} = \frac{D^2\rho}{\rho} + \frac{D^2\eta}{\eta} + 2 \frac{D\rho}{\rho} \left( \frac{D\eta}{\eta} \right)^\top; \]
and
\[ \frac{D\varphi}{\varphi} = \alpha \frac{D\rho}{\rho}; \quad \frac{D^2\varphi}{\varphi} = \alpha \frac{D^2\rho}{\rho} + \alpha (\alpha - 1) \left( \frac{D\rho}{\rho} \right)^\top. \]
The conclusion follows. \( \square \)

**Example 1.98.** A polynomial without real root is a balanced weight function; by the above lemma, the same holds for the power of quotient of such polynomials. Such weight functions may be useful for terminal conditions with polynomial growth.
In the sequel $\rho$ will be a balanced weight function. We set
\begin{equation}
(1.324) \quad \eta(x) := \rho(x)^{-1/2}; \quad \bar{v}(x, t) := \bar{u}(x, t)/\eta(x).
\end{equation}
By the above lemma, $\eta$ is a balanced weight function.

**Lemma 1.99.** Let $\bar{u}$ be smooth. Then $\bar{v} = \bar{u}/\eta$ is solution of the following PDE
\begin{equation}
(1.325) \quad \bar{v}_t(x, t) + \frac{1}{2} \eta''(x) \bar{v}(x, t) + \frac{\eta'(x)}{\eta(x)} \bar{v}_x(x, t) + \frac{1}{2} \bar{v}_{xx}(x, t) = 0, \quad x \in \mathbb{R}, \quad t \in [0, T],
\end{equation}
with final condition
\begin{equation}
(1.326) \quad \bar{v}(x, T) = g(x)/\eta(x), \quad x \in \mathbb{R}.
\end{equation}

**Proof.** Since $\bar{u}(x, t) = \eta(x) \bar{v}(x, t)$, we have that $\bar{u}_t(x, t) = \eta(x) \bar{v}_t(x, t)$, as well as
\begin{equation}
(1.327) \quad \left\{ \begin{array}{l}
\bar{u}_x(x, t) = \eta'(x) \bar{v}(x, t) + \eta(x) \bar{v}_x(x, t); \\
\bar{u}_{xx}(x, t) = \eta''(x) \bar{v}(x, t) + \eta(x) \bar{v}_{xx}(x, t) + 2\eta'(x) \bar{v}_x(x, t).
\end{array} \right.
\end{equation}
Substituting these expressions in (1.315) and dividing by $\eta(x)$, the result follows. $\square$

Let us now see how we can relate the new PDE with the *standard implicit scheme* for solving (1.315):
\begin{equation}
(1.328) \quad \frac{u_{j}^{k+1} - u_{j}^{k}}{h_0} + \frac{1}{2} \frac{u_{j+1}^{k} + u_{j-1}^{k} - 2u_{j}^{k}}{h_1^2} = 0, \quad j \in \mathbb{Z}, \quad k = 0, \ldots, N - 1,
\end{equation}
where $h_0 := T/N$, $h_1 > 0$, $x_j := jh_1$, $t_k := kh_0$ are “as usual”, and $u_j^k$ is an approximation of $\bar{u}(x_j, t_k)$, with final condition
\begin{equation}
(1.329) \quad u_j^N = g(x_j), \quad j \in \mathbb{Z}.
\end{equation}
We set $\eta_j := \eta(x_j)$ and $v_j^k := u_j^k/\eta_j$. Dividing (1.328) by $\eta_j$, we obtain that
\begin{equation}
(1.330) \quad \frac{v_j^{k+1} - v_j^k}{h_0} + \frac{1}{2} \frac{h_1^2}{\eta_j} \left( \eta_{j+1} v_{j+1}^{k+1} + \eta_{j-1} v_{j-1}^{k+1} - 2\eta_j v_j^{k} \right) = 0.
\end{equation}
For all $w = (w_j)_{j \in \mathbb{Z}}$, recalling that $\Delta_h w_j = h_1^2 (w_{j+1} + w_{j-1} - 2w_j)$, we have that
\begin{equation}
(1.331) \quad \left\{ \begin{array}{l}
\eta_{j+1} w_{j+1} + \eta_{j-1} w_{j-1} - 2\eta_j w_j = \\
h_1^2 (\Delta_h \eta_j) w_j + \frac{1}{2} (\eta_{j+1} - \eta_{j-1})(w_{j+1} - w_{j-1}) + \frac{1}{2} h_1^2 (\eta_{j+1} + \eta_{j-1}) \Delta_h w_j,
\end{array} \right.
\end{equation}
as can be checked by reordering the r.h.s. Applying this identity with $w_j = v_j^k$ and combining with (1.330) we deduce that
\begin{equation}
(1.332) \quad \frac{v_j^{k+1} - v_j^k}{h_0} + \frac{1}{2} \frac{\Delta_h \eta_j}{\eta_j} v_j^k + \frac{\eta_{j+1} - \eta_{j-1}}{2h_1 \eta_j} \frac{v_{j+1}^{k} - v_{j-1}^{k}}{2h_1} + \frac{1}{2} \frac{\eta_{j+1} + \eta_{j-1}}{2\eta_j} \Delta_h v_j^k = 0.
\end{equation}
As expected, this is a consistent approximation of the PDE (1.325). Denote the coefficients by
\begin{equation}
(1.333) \quad \hat{\beta}_j^k := \frac{1}{2} \frac{\Delta_h \eta_j}{\eta_j}, \quad \hat{\gamma}_j^k := \frac{\eta_{j+1} - \eta_{j-1}}{2h_1 \eta_j}, \quad \hat{\rho}_j^k := \frac{\eta_{j+1} + \eta_{j-1}}{2\eta_j}.
\end{equation}
Consider the following hypotheses, where $1$ is the constant function with value $1$:
\begin{equation}
(1.334) \quad \| \hat{\beta} \|_\infty + \| \hat{\gamma} \|_\infty < C; \quad \text{for some } C > 0,
\end{equation}
as well as
\begin{equation}
\hat{\beta}, \hat{\gamma}, \hat{\nu} \text{ are uniformly Lipschitz, and } \|\hat{\nu} - 1\|_{\infty} \to 0, \text{ when } h_1 \to 0.
\end{equation}

**Lemma 1.100.** Let (1.334)-(1.335) hold. Then the estimate (1.297) holds.

**Proof.** It suffices to adapt to the present setting the proof of lemma 1.88. □

**Remark 1.101.** The results of section 6.5 do not apply to the present setting since the second order term of the scheme has a varying coefficient, but we might apply them to the discretization of (1.325).

**Remark 1.102.** The practitioner has the choice either to discretize the heat equation itself, of the PDE (1.325) satisfied by \( v \). The latter allows to manipulate bounded numbers, and to deduce the analysis of error estimates from theorem 1.77. On the other hand, solving a PDE with constant coefficients may help to speed up the computations.

### 9. American and Bermudean options

#### 9.1. Overview.

The backward PDE that we have studied, with a given final condition, correspond to European options to be exercised at the final time. Two major variants are

- **Bermudean options** for which the option can be exercised at some given *exercise times* \( 0 = t_1 < \cdots < t_n = T \), with payoff \( g(x,t) \) for each exercise time \( t \). In the case of a call (resp. put) option, the exercise will take place at the first exercise time for which the present value of the option is not greater (resp. less) than the payoff.
- **American options** are similar, but the exercise can take place at any time \( t \) in \( [0,T] \).

In practice we may choose a step time \( h_0 = T/N \) and approximate the American option by the Bermudean one with exercise times \( t_i = ih_0, i = 0 \) to \( N \). Between two exercise times the option value will follow the same PDE as for a European option. So, we can integrate (backwards) over each time step, say between \( t_i \) and \( t_{i+1} \), using the schemes previously studied for European options, and then, in the case of a call (resp. put) option, take as value for the option at time \( t_i \) the maximum (resp. minimum) of the current value and of the payoff.

We detail the computations in the case of a call option, when the related PDE is the Black and Scholes one, after performing the logarithmic transformation, i.e., (1.8). Denoting by \( x \) the space variable, and setting
\begin{equation}
a(x,t) := \sigma(x,t)^2; \quad b(x,t) := r(x,t) - \frac{1}{2}\sigma(x,t)^2,
\end{equation}
this equation reads:
\begin{equation}
V_t(x,t) + b(x,t)V_x(x,t) + \frac{1}{2}a(x,t)V_{xx}(x,t) - r(x,t)V(x,t) = 0, \quad x \in \mathbb{R}, \ t \in (0,T),
\end{equation}
with final condition \( V(x,T) = g(x), \) for all \( x \in \mathbb{R} \).

#### 9.2. Explicit and implicit finite differences.
9.2.1. **Explicit centered finite differences.** We denote the gain function by \( s(x,t) \) and as usual set \( s_j^k = s(jh_1, kh_0) \). For the sake of simplicity we state the centered (rather than upwind) explicit scheme corresponding to i.e., in ordered form, with the notation of section 3.2.2 (compare to (1.99)),

\[
\left\{
\begin{array}{ll}
\hat{v}_{j}^{k-1} &= \left( 1 - h_0 r - \frac{h_0}{h_1^2} a_j^k \right) v_j^k + \frac{1}{2} \left( \frac{h_0}{h_1} b_j^k + \frac{h_0}{h_1^2} a_j^k \right) v_{j+1}^k \\
+ \frac{1}{2} \left( \frac{h_0}{h_1} b_j^k + \frac{h_0}{h_1^2} a_j^k \right) v_{j-1}^k ,
\end{array}
\right.
\]

(1.338)

\[
v_j^k - \hat{v}_{j}^{k-1} = \max(v_j^{k-1}, s_j^{k-1}), \quad j \in \mathbb{Z}, \quad k = 1 : N,
\]

\[
v_j^N = g_j, \quad j \in \mathbb{Z},
\]

The difference on the first relation above with (1.98) is due to the contribution of the zero order term of the PDE, that changes the coefficient of \( v_j^k \). So, this scheme is simply monotonic whenever (compare to (1.100)):

(1.339)

(i) \( h_0 \| r \|_\infty + \frac{h_0}{h_1^2} \| a \|_\infty \leq 1; \) (ii) \( h_1 |b(x,t)| \leq a(x,t), \) for all \( (x,t) \in \mathbb{R} \times [0,T] \).

When \( h_1 \) is small enough, the second relation and satisfied and the maximal value for \( h_0 \) is close to \( h_1^2/a \). In any case the time step must be of the order of the square of the space step.

It is instructive to write (1.338) in a different way. First, the second relation is equivalent to \( \min(v_j^{k-1} - \hat{v}_{j-1}^{k-1}, v_j^{k-1} - s_j^{k-1}) = 0 \) and therefore also to

(1.340)

\[
\min(v_j^{k-1} - \hat{v}_{j}^{k-1})/h_0, v_j^{k-1} - s_j^{k-1}) = 0.
\]

Using the first relation in (1.338), we obtain (compare to (1.98)):

(1.341)

\[
\min \left( \frac{v_j^{k-1} - v_j^k}{h_0} - b_j^k \frac{v_{j+1}^k - v_{j-1}^k}{2h_1} - \frac{1}{2} a_j^k \frac{v_{j+1}^k + v_{j-1}^k - 2v_j^k}{h_1^2} + rv_j^k, v_j^{k-1} - s_j^{k-1} \right) = 0.
\]

This suggests that the value of an American option is solution of the PDE

(1.342)

\[
\min (-v_j(x,t) - bv_j(x,t), v(x,t) - s_j(x,t)) = 0.
\]

This is indeed the case in an appropriate sense, see lemma 2.48 and remark 2.49.

9.2.2. **Implicit centered finite differences.** We need to replace the first line of (1.339) by the expression corresponding to (1.107), plus the contribution of the zero order term, i.e., for \( j \in \mathbb{Z} \) and \( k = 1 : N \):

(1.343)

\[
\frac{v_j^{k+1} - \hat{v}_{j}^k}{h_0} + b_j^k \frac{v_{j+1}^{k+1} - \hat{v}_{j-1}^k}{2h_1} + \frac{1}{2} a_j^k \frac{v_{j+1}^{k+1} + v_{j-1}^{k+1} - 2v_j^{k+1}}{h_1^2} + rv_j^{k+1} = 0.
\]

As in the case of the explicit scheme we can interpret this scheme as a discretization of (1.342) by noting that it is equivalent to

(1.344)

\[
\min \left( \frac{v_j^k - v_j^{k+1}}{h_0} - b_j^k \frac{v_{j+1}^{k+1} - \hat{v}_{j-1}^k}{2h_1} - \frac{1}{2} a_j^k \frac{v_{j+1}^{k+1} + v_{j-1}^{k+1} - 2v_j^{k+1}}{h_1^2} + rv_j^{k+1} \right) = 0.
\]
9.3. Fully implicit finite differences. The above expression suggests to consider the following scheme:

\[
(1.345) \quad \min \left( \frac{v^k_j - v^{k+1}_j}{h_0} - b_j^k \frac{v^{k+1}_j - v^{j-1}_j}{2h_1} - \frac{1}{2} a_j^k \frac{v^{j+1}_j + v^{j-1}_j - 2v^k_j}{h_1^2} + rv^k_j, v^k_j - s^k_j \right) = 0.
\]

Whereas for (1.344) the main task is to solve the linear equation corresponding to the implicit step, we now have to solve a nonlinear system. The algorithm remains invariant if we multiply the first argument of the minimum by \(h_0\); reordering the terms we get that

\[
(1.346) \quad \min \left( (1 + rh_0 + \frac{h_0}{h_1^2} a_j^k) v^k_j - (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{k+1}_j \right.
\]

\[
\left. - (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{j-1}_j - v^{k+1}_j, v^k_j - s^k_j \right) = 0.
\]

We next multiply the first argument of the minimum by

\[
(1.347) \quad \beta_j^k := (1 + rh_0 + \frac{h_0}{h_1^2} a_j^k)^{-1}
\]

so that the coefficient of \(v^k_j\) is one. Using that

\[
(1.348) \quad \min(v^k_j - F_j^k, v^k_j - s^k_j) = 0 \quad \text{iff} \quad v^k_j = \max(F_j^k, s^k_j),
\]

we obtain the equivalent expression

\[
(1.349) \quad v^k_j = \max \left( (\beta_j^k (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{k+1}_j + \beta_j^k (- (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{j-1}_j + \beta_j^k v^{k+1}_j, s^k_j) \right).
\]

Consider the operators: \(\ell^\infty \to \ell^\infty, \mathcal{T}\) and \(\mathcal{T}\), defined by

\[
(1.350) \quad (\mathcal{T} v)_j := \beta_j^k \left( (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{k+1}_j + (- (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{j-1}_j + v^{k+1}_j) \right)
\]

\[
(\mathcal{T} v)_j := \max(v^k_j, s^k_j).
\]

If (1.339)(ii) holds, then \(\mathcal{T}\) is a contraction (same arguments as in the study of implicit algorithms, section 2.4) and \(\mathcal{T} v\) is nonexpansive (it has Lipschitz constant 1). Therefore the composition \(\mathcal{T} \circ \mathcal{T}\) is a contraction. Since the scheme amounts to find a fixed point of \(\mathcal{T} \circ \mathcal{T}\), its solution is well-defined (provided that the terminal condition \(g\) is bounded, otherwise we should use weighted spaces).

Value iterations. We can compute numerically an approximation of the solution by iterating the fixed-point operator. Let us analyze how efficient is this approach. We cannot then solve exactly the equation of the scheme, but rather obtain an approximation such that (compare to (1.349)):

\[
(1.351) \quad v^k_j = \max \left( (\beta_j^k (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{k+1}_j + \beta_j^k (- (b_j^k \frac{h_0}{h_1} + \frac{1}{2} \frac{h_0}{h_1^2} a_j^k) v^{j-1}_j + \beta_j^k v^{k+1}_j, s^k_j) + h_0 f^k_j, \right.
\]

for some perturbation \(f^k_j\) that should be small enough. By arguments similar to those of the proof of lemma 1.14, it is not difficult to show that the mapping \(f \mapsto v\) is (in \(L^\infty\) spaces) has Lipschitz constant \(T\). For the sake of simplicity we make the optimistic
assumption that the error due to the scheme is of order of the step size $h_0$, and that $h_0 = O(h_0^2)$. Therefore, if

\begin{equation}
\|f\|_\infty = O(h_0),
\end{equation}

the value computed by the algorithm will still approximate the solution of the original problem with an error of $O(h_0)$.

At step $k$ of the algorithm, as starting point for the sequence generated by the fixed-point operator, we choose the value $v^{k+1}$. Let us assume that by doing so we are at distance $O(h_0)$ of the solution of this perturbed scheme, given $v^{k+1}$ (since if the solution is smooth, its variation over a time step is of order of $h_0$). The contraction coefficient being of order $O(1 + \lambda h_0)^{-1}$, we need a number of iterations $q$ such that $(1 + \lambda h_0)^{-q} h_0 = O(h_0^2)$, that is $q \lambda h_0 = O(\log(1/h_0))$, and finally since the number of time steps is $N = T/h_0$

$q = O(N \log N)$. Since there are $N$ time steps the algorithm need a total of $O(N^2 \log N)$ fixed-point iterations. If $N_x$ is the number of space point where we effectively make a computation, and $N = O(N_x^2)$, the number of operations is of order of $N_x^5$ (neglecting the logarithmic term). If $N_x = 100$ we get then $10^{10}$ operations which is already a large number. Clearly the method is not practicable.

**Remark 1.103.** For the standard implicit algorithm (1.344) we can solve the tridiagonal linear system in (1.343) in $O(N_x)$ operations, so that the total number of operations is $O(N_x^3)$. This is optimal since this is the order of the number of points in the grid.
Bibliography


Index

θ scheme, 23

cone
  positive, 82

assembly
  in finite elements, 124
balanced weight function, 63, 74
basket options, 30

centered schemes, 27
characteristics, 24
compatibility
  with a stencil, 40
cone
  pointed, 82
consistency error, 50
convolution, 11
  discrete, 16
Courant-Friedrich-Lewy condition, 25
diameter, 121
diffusion
  diagonal, 31
  diagonally dominant, 33
diffusion matrix, 30
  scaled, 35
  dual cone, 39
  dual norm, 72
elliptic operator, 74
finite differences
  generalized, 36
Gelfand triple, 77, 78
heat kernel, 12
interpolation operator, 122
monotone operator, 89
multiindex, 53

multiplier, 100

norm
  Frobenius, 40
option
  call, 9
  digital, 9
  put, 9
  trivial, 9
order relation, 81
  induced, 83
primality mapping, 86
quickhull algorithm, 41
radius, 121
reflexive, 72
Ritz projection, 128

semi coercive, 76, 78
semi implicit schemes, 28
simplex, 122
  face, 122
  standard, 122
splitting, 29
stencil, 36
  diagonal, 36
  diagonally dominant, 39
stochastic matrix, 16
strike, 9
transport, 28
triangulation, 121
underlying, 9
upwind, 24
upwind scheme, 26
weak convergence, 72
weighted spaces, 61