Numerical Analysis - Part II

Anders C. Hansen

Lecture 10

The diffusion equation in two space dimensions

The diffusion equation in two space dimensions

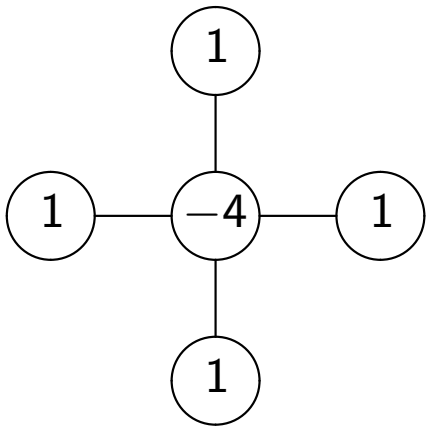
We are solving

$$\frac{\partial u}{\partial t} = \nabla^2 u, \qquad 0 \leq x, y \leq 1, \quad t \geq 0, \tag{1}$$

where $u = u(x, y, t)$, together with initial conditions at $t = 0$ and Dirichlet boundary conditions at $\partial\Omega$, where $\Omega = [0, 1]^2 \times [0, \infty)$. It is straightforward to generalize our derivation of numerical algorithms, e.g. by the method of lines.

Recall the five point formula

We have the *five-point method*



$$u_{i,j} = u_{i-1,j} + u_{i+1,j} + u_{i,j-1} + u_{i,j+1} - 4u_{i,j}\,,$$

discretising the two dimensional Laplacian.

The diffusion equation in two space dimensions

Thus, let $u_{\ell,m}(t) \approx u(\ell h, mh, t)$, where $h = \Delta x = \Delta y$, and let $u^n_{\ell,m} \approx u_{\ell,m}(nk)$ where $k = \Delta t$. The five-point formula results in

$$u'_{\ell,m} = \tfrac{1}{h^2}\big(u_{\ell-1,m} + u_{\ell+1,m} + u_{\ell,m-1} + u_{\ell,m+1} - 4u_{\ell,m}\big),$$

or in the matrix form

$$\mathbf{u}' = \tfrac{1}{h^2} A_* \mathbf{u}, \qquad \mathbf{u} = (u_{\ell,m}) \in \mathbb{R}^N, \tag{2}$$

where A_* is the block TST matrix of the five-point scheme:

$$A_* = \begin{bmatrix} H & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & H \end{bmatrix}, \quad H = \begin{bmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{bmatrix}.$$

Crank-Nicolson for 2D

Applying the trapezoidal rule to our semi-dicretization (2) we obtain the two-dimensional Crank-Nicolson method:

$$(I - \tfrac{1}{2}\mu A_*)\,\mathbf{u}^{n+1} = (I + \tfrac{1}{2}\mu A_*)\,\mathbf{u}^{n}\,, \tag{3}$$

in which we move from the n-th to the $(n+1)$-st level by solving the system of linear equations $B\mathbf{u}^{n+1} = C\mathbf{u}^{n}$, or $\mathbf{u}^{n+1} = B^{-1}C\mathbf{u}^{n}$. For stability, similarly to the one-dimensional case, the eigenvalue analysis implies that $A = B^{-1}C$ is normal and shares the same eigenvectors with B and C, hence

$$\lambda(A) = \frac{\lambda(C)}{\lambda(B)} = \frac{1 + \frac{1}{2}\mu\lambda(A_*)}{1 - \frac{1}{2}\mu\lambda(A_*)} \quad \Rightarrow \quad |\lambda(A)| < 1 \text{ as } \lambda(A_*) < 0$$

and the method is stable for all μ. The same result can be obtained through the Fourier analysis.

Splitting

We would like to find a fast solver to the system (3). The matrix $B = I - \frac{1}{2}\mu A_*$ has a structure similar to that of A_*, where

$$A_* = \begin{bmatrix} H & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & H \end{bmatrix}, \quad H = \begin{bmatrix} -4 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{bmatrix}.$$

so we may apply the Hockney method.

Observation 1 (Special structure of 5-point equations)

We wish to motivate and introduce a family of efficient solution methods for the 5-point equations: the *fast Poisson solvers*. Thus, suppose that we are solving $\nabla^2 u = f$ in a square $m \times m$ grid with the 5-point formula (all this can be generalized a great deal, e.g. to the nine-point formula). Let the grid be enumerated in *natural ordering*, i.e. by columns. Thus, the linear system $A\mathbf{u} = \mathbf{b}$ can be written explicitly in the block form

$$\underbrace{\begin{bmatrix} B & I & & \\ I & B & \ddots & \\ & \ddots & \ddots & I \\ & & I & B \end{bmatrix}}_{A} \begin{bmatrix} \mathbf{u}_1 \\ \mathbf{u}_2 \\ \vdots \\ \mathbf{u}_m \end{bmatrix} = \begin{bmatrix} \mathbf{b}_1 \\ \mathbf{b}_2 \\ \vdots \\ \mathbf{b}_m \end{bmatrix}, \qquad B = \begin{bmatrix} -4 & 1 & & \\ 1 & -4 & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -4 \end{bmatrix}_{m \times m},$$

where $\mathbf{u}_k, \mathbf{b}_k \in \mathbb{R}^m$ are portions of $\mathbf{u}$ and $\mathbf{b}$, respectively, and B is a TST-matrix which means *tridiagonal, symmetric* and *Toeplitz* (i.e., constant along diagonals).

Special structure of 5-point equations

Observation 2 (Special structure of 5-point equations)

By Exercise 4, its eigenvalues and orthonormal eigenvectors are given as

$$B\mathbf{q}_\ell = \lambda_\ell \mathbf{q}_\ell, \qquad \lambda_\ell = -4 + 2\cos\tfrac{\ell\pi}{m+1},$$

$$\mathbf{q}_\ell = \gamma_m \big(\sin\tfrac{j\ell\pi}{m+1}\big)_{j=1}^{m}, \qquad \ell = 1..m,$$

where $\gamma_m = \sqrt{\frac{2}{m+1}}$ is the normalization factor. Hence $B = QDQ^{-1} = QDQ$, where $D = \operatorname{diag}(\lambda_\ell)$ and $Q = Q^T = (q_{j\ell})$. Note that all $m \times m$ TST matrices share the same full set of eigenvectors, hence they all commute!

The Hockney method

Set $\mathbf{v}_k = Q\mathbf{u}_k$, $\mathbf{c}_k = Q\mathbf{b}_k$, therefore our system becomes

$$\begin{bmatrix} D & I & & \\ I & D & \ddots & \\ & \ddots & \ddots & I \\ & & I & D \end{bmatrix} \begin{bmatrix} \mathbf{v}_1 \\ \mathbf{v}_2 \\ \vdots \\ \mathbf{v}_m \end{bmatrix} = \begin{bmatrix} \mathbf{c}_1 \\ \mathbf{c}_2 \\ \vdots \\ \mathbf{c}_m \end{bmatrix}.$$

Let us by this stage reorder the grid *by rows, instead of by columns.*. In other words, we permute $\mathbf{v} \mapsto \widehat{\mathbf{v}} = P\mathbf{v}$, $\mathbf{c} \mapsto \widehat{\mathbf{c}} = P\mathbf{c}$, so that the portion $\widehat{\mathbf{c}}_1$ is made out of the first components of the portions $\mathbf{c}_1, \ldots, \mathbf{c}_m$, the portion $\widehat{\mathbf{c}}_2$ out of the second components and so on.

The Hockney method

This results in new system

$$\begin{bmatrix} \Lambda_1 & & \\ & \Lambda_2 & \\ & & \ddots \\ & & & \Lambda_m \end{bmatrix} \begin{bmatrix} \widehat{\mathbf{v}}_1 \\ \widehat{\mathbf{v}}_2 \\ \vdots \\ \widehat{\mathbf{v}}_m \end{bmatrix} = \begin{bmatrix} \widehat{\mathbf{c}}_1 \\ \widehat{\mathbf{c}}_2 \\ \vdots \\ \widehat{\mathbf{c}}_m \end{bmatrix}, \quad \Lambda_k = \begin{bmatrix} \lambda_k & 1 & & & \\ 1 & \lambda_k & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & 1 & \lambda_k \end{bmatrix}_{m \times m},$$

where $k = 1 \ldots m$.

The Hockney method

These are m *uncoupled* systems, $\Lambda_k \widehat{\mathbf{v}}_k = \widehat{\mathbf{c}}_k$ for $k = 1...m$. Being *tridiagonal*, each such system can be solved fast, at the cost of $\mathcal{O}(m)$. Thus, the steps of the algorithm and their computational cost are as follows.

1. Form the products $\mathbf{c}_k = Q\mathbf{b}_k$, $k = 1...m$ ..... $\mathcal{O}(m^3)$
2. Solve $m \times m$ tridiagonal systems $\Lambda_k \widehat{\mathbf{v}}_k = \widehat{\mathbf{c}}_k$, $k = 1...m$ ..... $\mathcal{O}(m^2)$

3. Form the products $\mathbf{u}_k = Q\mathbf{v}_k$, $k = 1...m$ ..... $\mathcal{O}(m^3)$

Splitting

However, since the method (3) has a local truncation error $\mathcal{O}(k^3 + kh^2)$, we don't need an exact solution of the system: it would be enough to have one within the error.
Let us employ the notation

$$\Delta_x^2 u_{\ell,m} = u_{\ell-1,m} - 2u_{\ell,m} + u_{\ell+1,m}, \qquad \Delta_y^2 u_{\ell,m} = u_{\ell,m-1} - 2u_{\ell,m} + u_{\ell,m+1}\,.$$

Then the Crank-Nicolson method calculates $\mathbf{u}^{n+1}$ by solving the system

$$\left[I - \tfrac{1}{2}\mu(\Delta_x^2 + \Delta_y^2)\right] u_{\ell,m}^{n+1} = \left[I + \tfrac{1}{2}\mu(\Delta_x^2 + \Delta_y^2)\right] u_{\ell,m}^{n}\,, \quad \ell, m = 1 \ldots M. \tag{4}$$

Splitting

The local error is however preserved if we replace this formula by the difference equation

$$\left[I - \tfrac{1}{2}\mu\Delta_x^2\right]\left[I - \tfrac{1}{2}\mu\Delta_y^2\right] u_{\ell,m}^{n+1} = \left[I + \tfrac{1}{2}\mu\Delta_x^2\right]\left[I + \tfrac{1}{2}\mu\Delta_y^2\right] u_{\ell,m}^{n}\,, \quad (5)$$

which is called the split version of Crank-Nicolson. Indeed, the difference between two schemes is equal to

$$\begin{aligned} &\tfrac{1}{4}\mu^2\Delta_x^2\Delta_y^2(u_{\ell,m}^{n+1} - u_{\ell,m}^{n}) = \tfrac{k^2}{4}\tfrac{1}{h^2}\Delta_x^2\tfrac{1}{h^2}\Delta_y^2\left(k\tfrac{\partial}{\partial t}u_{\ell,m}^{n} + \mathcal{O}(k^2)\right) \\ &= \tfrac{k^3}{4}\left(\tfrac{\partial^2}{\partial x^2}\tfrac{\partial^2}{\partial y^2}\tfrac{\partial}{\partial t}u_{\ell,m}^{n} + \mathcal{O}(k+h^2)\right) = \mathcal{O}(k^3 + kh^2)\,, \end{aligned} \quad (6)$$

the same magnitude as of the local error.

Splitting

In the matrix form, (5) is equivalent to splitting the matrix A_* into the sum of two matrices A_x and A_y as

$$A_* = A_x + A_y,$$

$$A_x = \begin{bmatrix} -2I & I & & \\ I & \ddots & \ddots & \\ & \ddots & \ddots & I \\ & & I & -2I \end{bmatrix}, \; A_y = \begin{bmatrix} H & & & \\ & H & & \\ & & \ddots & \\ & & & H \end{bmatrix}, \; H = \begin{bmatrix} -2 & 1 & & \\ 1 & \ddots & \ddots & \\ & \ddots & \ddots & 1 \\ & & 1 & -2 \end{bmatrix}$$

and solving the uncoupled system

$$\left[I - \tfrac{1}{2}\mu A_x\right]\left[I - \tfrac{1}{2}\mu A_y\right]\mathbf{u}^{n+1} = \left[I + \tfrac{1}{2}\mu A_x\right]\left[I + \tfrac{1}{2}\mu A_y\right]\mathbf{u}^{n}.$$

as

$$B_x\mathbf{u}^{n+1/2} = C_xC_y\mathbf{u}^{n}, \qquad B_y\mathbf{u}^{n+1} = \mathbf{u}^{n+1/2}.$$

Splitting

The matrix

$$B_y = I - \frac{1}{2}\mu A_y$$

is block diagonal, and solving $B_y\mathbf{u} = \mathbf{v}$ is just solving one and the same tridiagonal system $B\mathbf{u}_i = \mathbf{v}_i$ with different right-hand sides. Matrix $B_x = I - \frac{1}{2}\mu A_x$ is of the same form up to a permutation (reodering of the grid), so solving $B_x\mathbf{v} = \mathbf{b}$ is again a fast procedure.

The general diffusion equation

Consider the general diffusion equation

$$\begin{aligned}\frac{\partial u}{\partial t} &= \nabla^\top \left(a(x,y)\nabla u\right) + f(x,y) \\ &= \frac{\partial}{\partial x}\left(a(x,y)\frac{\partial u}{\partial x}\right) + \frac{\partial}{\partial y}\left(a(x,y)\frac{\partial u}{\partial y}\right) + f(x,y),\end{aligned} \tag{7}$$

where $a(x,y) > \alpha > 0$ and $f(x,y)$ are given, together with initial conditions on $[0,1]^2$ and Dirichlet boundary conditions along $\partial[0,1]^2 \times [0,\infty)$. Replace each space derivative by *central differences* at midpoints,

$$\frac{\mathrm{d}g(\xi)}{\mathrm{d}\xi} \approx \frac{g(\xi + \frac{1}{2}h) - g(\xi - \frac{1}{2}h)}{h},$$

resulting in the ODE system

$$\begin{aligned}u'_{\ell,m} \quad = \quad \frac{1}{h^2}\Big[&a_{\ell-\frac{1}{2},m}u_{\ell-1,m} + a_{\ell+\frac{1}{2},m}u_{\ell+1,m} + a_{\ell,m-\frac{1}{2}}u_{\ell,m-1} + a_{\ell,m+\frac{1}{2}}u_{\ell,m+1} \\ &- \left(a_{\ell-\frac{1}{2},m} + a_{\ell+\frac{1}{2},m} + a_{\ell,m-\frac{1}{2}} + a_{\ell,m+\frac{1}{2}}\right)u_{\ell,m}\Big] + f_{\ell,m}.\end{aligned} \tag{8}$$

The general diffusion equation

Assuming zero boundary conditions and $f \equiv 0$, we have a system $\mathbf{u}' = A\mathbf{u}$, and we may solve it again by Crank–Nicolson, and apply the split

$$A = A_x + A_y.$$

Here, A_x and A_y are again constructed from the contribution of discretizations in the x- and y-directions respectively, namely A_x includes all the $a_{\ell\pm\frac{1}{2},m}$ terms, and A_y consists of the remaining $a_{\ell,m\pm\frac{1}{2}}$ components. Arguments similar to what we used in moving from (4) to (5) justify the use of the split version in this general case as well.

Intermezzo – Linear systems of ODEs

With greater generality, let us consider the ODE system

$$\mathbf{y}' = A\mathbf{y}, \qquad \mathbf{y}(0) = \mathbf{y}_0. \tag{9}$$

We define formally a *matrix exponential* by Taylor series, $\mathrm{e}^B := \sum_{k=0}^{\infty} \frac{1}{k!} B^k$, and easily verify by formal differentiation that $\mathrm{d}\mathrm{e}^{tA}/\mathrm{d}t = A\mathrm{e}^{tA}$, therefore $\mathbf{y}(t) = \mathrm{e}^{tA}\mathbf{y}_0$ is a soluton.
One observes that one-step methods for solving (9) are approximating a matrix exponential. Thus, with $k = \Delta t$,

Euler: $\mathbf{y}_n = (I + kA)^n \mathbf{y}_0$, $\qquad 1 + z = \mathrm{e}^z + \mathcal{O}(z^2)$;

TR: $\mathbf{y}_n = \left[\left(I - \frac{1}{2}kA\right)^{-1} \left(I + \frac{1}{2}kA\right) \right]^n \mathbf{y}_0$, $\qquad \frac{1+\frac{1}{2}z}{1-\frac{1}{2}z} = \mathrm{e}^z + \mathcal{O}(z^3)$.

Splitting methods – The philosophy

Recall that, for $z_1, z_2 \in \mathbb{C}$, we have $\mathrm{e}^{z_1+z_2} = \mathrm{e}^{z_1}\mathrm{e}^{z_2}$ and had this been true for the matrices, i.e. that $\mathrm{e}^{tA} = \mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$, we could have approximated each component of the exponent of $A = A_x + A_y$ with the trapezoidal rule, say, to produce

$$\mathbf{u}^{n+1} = \left(I - \tfrac{1}{2}\mu A_x\right)^{-1}\left(I + \tfrac{1}{2}\mu A_x\right)\left(I - \tfrac{1}{2}\mu A_y\right)^{-1}\left(I + \tfrac{1}{2}\mu A_y\right)\mathbf{u}^n\,, \quad \mu = k/h^2\,, \tag{10}$$

and since both $I - \frac{1}{2}\mu A_x$ and $I - \frac{1}{2}\mu A_y$ are tridiagonal, this system can be solved very cheaply.

Splitting methods – The philosophy

Unfortunately, the assumption that $\mathrm{e}^{t(B+C)} = \mathrm{e}^{tB}\mathrm{e}^{tC}$ is, in general, false. Not all hope is lost, though, and we will demonstrate that, suitably implemented, splitting is a powerful technique to reduce drastically the expense of numerical solution.

Splitting methods – The philosophy

Comparing the Taylor expansions of $\mathrm{e}^{t(B+C)}$ with $\mathrm{e}^{tB}\mathrm{e}^{tC}$ we obtain

$$\mathrm{e}^{tB}\mathrm{e}^{tC} = \mathrm{e}^{t(B+C)} + \tfrac{1}{2}t^2(BC - CB) + \mathcal{O}(t^3). \tag{11}$$

In particular, $\mathrm{e}^{tB}\mathrm{e}^{tC} = \mathrm{e}^{t(B+C)}$ for all $t \geq 0$ if and only if B and C commute. The good news is, however, that approximating $\mathrm{e}^{\Delta t(B+C)}$ with $\mathrm{e}^{\Delta tB}\mathrm{e}^{\Delta tC}$ incurs an error of $\mathcal{O}((\Delta t)^2)$. So, if r is a rational function such that $r(z) = \mathrm{e}^z + \mathcal{O}(z^2)$, then

$$\mathbf{u}^{n+1} = r(\mu A_x)r(\mu A_y)\mathbf{u}^n \tag{12}$$

produces an error of $\mathcal{O}((\Delta t)^2)$. The choice $r(z) = (1 + \frac{1}{2}z)/(1 - \frac{1}{2}z)$ results in a *split Crank–Nicolson* scheme, whose implementation reduces to a solution of tridiagonal algebraic linear systems.

Splitting methods – Strang splitting

It is easy to prove that

$$\mathrm{e}^{t(B+C)} = \tfrac{1}{2}\left(\mathrm{e}^{tB}\mathrm{e}^{tC} + \mathrm{e}^{tC}\mathrm{e}^{tB}\right) + \mathcal{O}(t^3),$$
$$\mathrm{e}^{t(B+C)} = \mathrm{e}^{\frac{1}{2}tB}\mathrm{e}^{tC}\mathrm{e}^{\frac{1}{2}tB} + \mathcal{O}(t^3),$$

the second formula is called the *Strang splitting*. Thus, as long as $r(z) = \mathrm{e}^z + \mathcal{O}(z^3)$, the time-stepping formula

$$\mathbf{u}^{n+1} = r\left(\tfrac{1}{2}\mu A_x\right) r\left(\mu A_y\right) r\left(\tfrac{1}{2}\mu A_x\right) \mathbf{u}^n$$

carries a local error of $\mathcal{O}((\Delta t)^3)$.

Splitting methods – Stability

As far as stability is concerned, we observe that both A_x and A_y are symmetric, hence normal, therefore so are $r(\mu A_x)$ and $r(\mu A_y)$. Then Euclidean ℓ_2-norm equals the spectral radius, therefore for the splitting (12), we have

$$\|\mathbf{u}^{n+1}\| \leq \|r(\mu A_x)\| \cdot \|r(\mu A_y)\| \cdot \|\mathbf{u}^n\| = \rho[r(\mu A_x)] \cdot \rho[r(\mu A_y)] \cdot \|\mathbf{u}^n\|.$$

It is easy to verify by Gershgorin theorem that the eigenvalues of the matrices A_x and A_y are nonpositive, hence provided that r fulfils $|r(z)| < 1$ for $z \in \mathbb{C}$ with $\operatorname{Re} z < 0$, it is then true that

$$\rho[r(\mu A_x)], \rho[r(\mu A_y)] \leq 1.$$

This proves $\|\mathbf{u}^{n+1}\| \leq \|\mathbf{u}^n\| \leq \cdots \leq \|\mathbf{u}^0\|$, hence stability.

Splitting of inhomogeneous systems

Recall our goal, namely fast methods for the two-dimensional diffusion equation. Our exposition so far has been contrived, because of the assumption that the boundary conditions are zero. In general, the linear ODE system is of the form

$$\mathbf{u}' = A\mathbf{u} + \mathbf{b}, \qquad \mathbf{u}(0) = \mathbf{u}^0, \tag{13}$$

where $\mathbf{b}$ originates in boundary conditions (and in a forcing term $f(x, y)$ in the original PDE (7)).

Splitting of inhomogeneous systems

Note that our analysis should accommodate $\mathbf{b} = \mathbf{b}(t)$, since boundary conditions might vary in time! The *exact* solution of (13) is provided by the *variation of constants* formula

$$\mathbf{u}(t) = \mathrm{e}^{tA}\mathbf{u}(0) + \int_0^t \mathrm{e}^{(t-s)A}\mathbf{b}(s)\,\mathrm{d}s, \qquad t \geq 0,$$

therefore

$$\mathbf{u}(t_{n+1}) = \mathrm{e}^{\Delta t A}\mathbf{u}(t_n) + \int_{t_n}^{t_{n+1}} \mathrm{e}^{(t_{n+1}-s)A}\mathbf{b}(s)\,\mathrm{d}s\,.$$

Splitting of inhomogeneous systems

The integral can be frequently evaluated explicitly, e.g. when $\mathbf{b}$ is a linear combination of polynomial and exponential terms. For example, $\mathbf{b}(t) \equiv \mathbf{b} = \text{const}$ yields

$$\mathbf{u}(t_{n+1}) = \mathrm{e}^{\Delta t A}\mathbf{u}(t_n) + A^{-1}\left(\mathrm{e}^{\Delta t A} - I\right)\mathbf{b}.$$

This, unfortunately, is not a helpful observation, since, even if we split the exponential e^{tA}, how are we supposed to split $A^{-1} = (B + C)^{-1}$?

Splitting of inhomogeneous systems

The remedy is not to evaluate the integral explicitly but, instead, to use quadrature. For example, the trapezoidal rule $\int_0^k g(\tau)\,\mathrm{d}\tau = \frac{1}{2}k[g(0)+g(k)] + \mathcal{O}(k^3)$ gives

$$\mathbf{u}(t_{n+1}) \;\approx\; \mathrm{e}^{\Delta t A}\mathbf{u}(t_n) + \tfrac{1}{2}\Delta t[\mathrm{e}^{\Delta t A}\mathbf{b}(t_n) + \mathbf{b}(t_{n+1})],$$

with a local error of $\mathcal{O}((\Delta t)^3)$. We can now replace exponentials with their splittings. For example, Strang's splitting results in

$$\mathbf{u}^{n+1} \;=\; r\big(\tfrac{1}{2}\Delta t B\big)\, r\big(\Delta t C\big)\, r\big(\tfrac{1}{2}\Delta t B\big)\big[\mathbf{u}^n + \tfrac{1}{2}\Delta t\mathbf{b}^n\big] + \tfrac{1}{2}\Delta t\mathbf{b}^{n+1}.$$

As before, everything reduces to (inexpensive) solution of tridiagonal systems!