Chapter III.
Numerical Methods for the Time-Dependent Schrödinger Equation

chap:num-tdse

This chapter deals with numerical methods for linear time-dependent Schrödinger equations, of low to moderate dimension (less than 10, say). Although the emphasis is on time-dependent aspects, we begin with a section on space discretization, where we describe the Galerkin and collocation approaches on the important examples of Hermite and Fourier bases, including their extension to higher dimensions using hyperbolic cross approximations and sparse grids for which the computational work grows only mildly with the dimension.

We then turn to time-stepping methods: polynomial approximations to the exponential of the Hamiltonian based on the Lanczos method or on Chebyshev polynomials, and splitting methods and their high-order refinements by composition and processing. We conclude the chapter with integrators for Schrödinger equations with a time-varying potential.

The time-dependent Schrödinger equation considered in this chapter (unless stated otherwise) is in $d \geq 1$ space dimensions, has $\hbar = 1$ and reads

$$i\,\frac{\partial \psi}{\partial t} = H\psi\,, \qquad H = T + V\,, \tag{0.1}$$

III:schrod-eq

with the kinetic energy operator $T = -\frac{1}{2\mu}\Delta$ with a positive mass parameter μ and a potential $V(x)$. In the final section we consider a time-dependent potential $V(x,t)$.

III.1 Space Discretization by Spectral Methods

We follow two tracks (among many possible) for the discretization of the Schrödinger equation in space: the Galerkin method with a basis of Hermite functions and collocation with trigonometric polynomials. Both cases are instances of spectral or pseudospectral methods, which are of common use in many application areas; see, e.g., Canuto, Hussaini, Quarteroni & Zang (2006), Fornberg (1996), Gottlieb & Orszag (1977), and Trefethen (2000). Both cases are studied here for the Schrödinger equation in one and several dimensions, with the extension to higher dimensions by hyperbolically reduced tensor product bases.

III.1.1 Galerkin Method, 1D Hermite Basis

Galerkin Method. We consider an approximation space $\mathcal{V}_K \subset L^2(\mathbb{R}^d)$ spanned by K basis functions $\varphi_0, \dots, \varphi_{K-1}$. We determine an approximate wave function $\psi_K(t) \in \mathcal{V}_K$ by the condition that at every instant t, its time derivative is determined by the condition

$$\frac{d\psi_K}{dt} \in \mathcal{V}_K \quad \text{such that} \quad \left\langle \varphi \,\middle|\, i\,\frac{d\psi_K}{dt} - H\psi_K \right\rangle = 0 \qquad \forall\, \varphi \in \mathcal{V}_K\,. \tag{1.1}$$

III:galerkin

This is, of course, the time-dependent variational principle (II.1.2) on the linear approximation space $\mathcal{V}_K$. In particular, we know from Sect. II.1 that norm, energy and symplectic structure are preserved. Writing the approximation as a linear combination of basis functions

$$\psi_K(t) = \sum_{k=0}^{K-1} c_k(t)\varphi_k \tag{1.2}$$

III:gal-sum

and inserting in (1.1), we obtain for the time-dependent coefficient vector $c = (c_k)$ the linear system of ordinary differential equations

$$i\,M_K \dot{c} = H_K c \tag{1.3}$$

III:gal-coeff

with the matrices

$$M_K = \big(\langle \varphi_j \,|\, \varphi_k\rangle\big)_{j,k=0}^{K-1}\,, \quad H_K = \big(\langle \varphi_j \,|\, H \,|\, \varphi_k\rangle\big)_{j,k=0}^{K-1}\,. \tag{1.4}$$

III:gal-matrix

The matrix M_K becomes the identity matrix in the case of an orthonormal basis, where $\langle \varphi_j \,|\, \varphi_k\rangle = \delta_{jk}$.

Hermite Basis in 1D. After a suitable rescaling and shift $x \to \alpha x + \beta$, this is the choice of basis functions

$$\varphi_k(x) = \frac{1}{\pi^{1/4}}\,\frac{1}{\sqrt{2^k k!}}\,H_k(x)\,e^{-x^2/2}\,. \tag{1.5}$$

III:hermite-formula

Here, $H_k(x)$ is the Hermite polynomial of degree k, which is the kth orthogonal polynomial with respect to the weight function e^{-x^2} on $\mathbb{R}$; see, e.g., Abramowitz & Stegun (1965). While formula (1.5) does not fail to impress, it is neither useful for computations nor for understanding the approximation properties of this basis. We therefore now turn to another way of writing the Hermite functions φ_k, which also provides some motivation for the use of this basis.

Ladder Operators. We recall the canonical commutator relation (I.4.8) between the one-dimensional position operator q given by $(q\psi)(x) = x\psi(x)$ and the momentum operator $p = -i\,d/dx$:

$$\frac{1}{i}\,[q,p] = 1.$$

It follows that the *ladder operators* defined by

$$A = \frac{1}{\sqrt{2}}\,(q + ip)\,, \quad A^\dagger = \frac{1}{\sqrt{2}}\,(q - ip) \tag{1.6}$$

III:ladder

satisfy the relations

$$A^\dagger A = \frac{1}{2}\,(p^2+q^2) - \frac{1}{2}\,, \qquad AA^\dagger = \frac{1}{2}\,(p^2+q^2) + \frac{1}{2}\,, \tag{1.7}$$ III:AdA

so that $A^\dagger A$ and $AA^\dagger$ have the same eigenfunctions as the Hamiltonian of the harmonic oscillator, $\frac{1}{2}\,(p^2+q^2)$. We also note

$$AA^\dagger = A^\dagger A + 1\,. \tag{1.8}$$ III:AdA-commute

Moreover, $A^\dagger$ is adjoint to A on the Schwartz space $\mathcal{S}$ of smooth rapidly decaying functions:

$$\langle A^\dagger\varphi \,|\, \psi\rangle = \langle \varphi \,|\, A\psi\rangle \qquad \forall\, \varphi,\psi \in \mathcal{S}\,. \tag{1.9}$$ III:A-adj

Harmonic Oscillator Eigenfunctions. We note that the Gaussian $\phi_0(x) = e^{-x^2/2}$ is in the kernel of A: $A\phi_0 = 0$. Moreover, it is checked that multiples of ϕ_0 are the only L^2 functions in the kernel of A, whereas $A^\dagger$ has only the trivial kernel 0. With (1.8) it follows that

$$AA^\dagger\phi_0 = A^\dagger A\phi_0 + \phi_0 = \phi_0\,,$$

and hence ϕ_0 is an eigenfunction of $AA^\dagger$ to the eigenvalue 1. Applying the operator $A^\dagger$ to both sides of this equation, we see that $\phi_1 = A^\dagger\phi_0$ is an eigenfunction of $A^\dagger A$ to the eigenvalue 1, and again by (1.8) an eigenfunction of $AA^\dagger$ to the eigenvalue 2. We continue in this way to construct successively $\phi_{k+1} = A^\dagger\phi_k$ for $k \geq 0$. We thus obtain eigenfunctions ϕ_k to $A^\dagger A$, with eigenvalue k, and to $AA^\dagger$, with eigenvalue $k+1$. These eigenfunctions are not yet normalized. To achieve this, we note that by (1.9),

$$\|A^\dagger\phi_k\|^2 = \langle A^\dagger\phi_k \,|\, A^\dagger\phi_k\rangle = \langle \phi_k \,|\, AA^\dagger\phi_k\rangle = (k+1)\,\|\phi_k\|^2\,.$$

We therefore obtain eigenfunctions to $AA^\dagger$ and $A^\dagger A$ of unit L^2 norm by setting

$$\varphi_0(x) = \frac{1}{\pi^{1/4}}\,e^{-x^2/2} \tag{1.10}$$ III:phi0

and

$$\varphi_{k+1} = \frac{1}{\sqrt{k+1}}\,A^\dagger\varphi_k \qquad \text{for} \quad k \geq 0\,. \tag{1.11}$$ III:raising

Since $A\varphi_{k+1} = \frac{1}{\sqrt{k+1}}\,AA^\dagger\varphi_k = \sqrt{k+1}\,\varphi_k$, we also have (replacing $k+1$ by k)

$$\varphi_{k-1} = \frac{1}{\sqrt{k}}\,A\varphi_k \qquad \text{for} \quad k \geq 0\,. \tag{1.12}$$ III:lowering

These relations explain the names of *raising operator* and *lowering operator* for $A^\dagger$ and A, respectively, and of *ladder operators* for both of them. Multiplying (1.11) by $\sqrt{k+1}$ and (1.12) by $\sqrt{k}$, summing the resulting formulas and using the definitions of A and $A^\dagger$, we obtain the three-term recurrence relation

$$\sqrt{k+1}\,\varphi_{k+1}(x) = \sqrt{2}\,x\,\varphi_k(x) - \sqrt{k}\,\varphi_{k-1}(x) \qquad \text{for} \quad k \geq 0\,, \tag{1.13}$$ III:hermite-rec

with $\varphi_{-1}(x) = 0$. This allows us to evaluate $\varphi_k(x)$ at any required point x. We state essential properties of these functions.

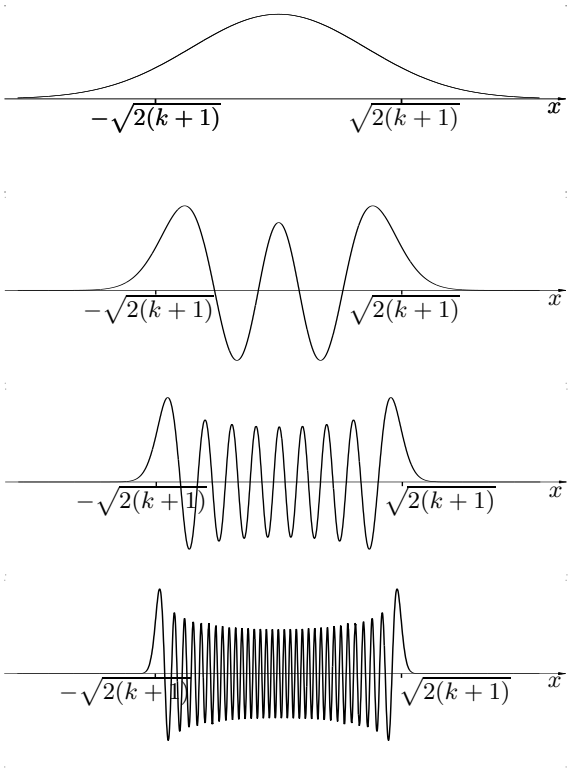

Fig. 1.1. Hermite functions φ_k for $k = 0, 4, 16, 64$.

III:thm:hermite

Theorem 1.1 (Hermite Functions). *The functions φ_k defined by (1.10) and (1.11) form a complete L^2-orthonormal set of functions, the eigenfunctions of the harmonic oscillator Hamiltonian $\frac{1}{2}(p^2 + q^2)$. They are identical to the Hermite functions given by (1.5).*

Proof. From the above construction it is clear that each φ_k is an oscillator eigenfunction to the eigenvalue $k + \frac{1}{2}$. As normalized eigenfunctions of a self-adjoint operator, the φ_k are orthonormal. It is also clear from the recurrence relation that φ_k is a polynomial of degree k times $e^{-x^2/2}$. By the orthonormality, this polynomial must be a multiple of the kth Hermite polynomial, which yields (1.5). For the proof of completeness we refer to Thaller (2000), Sect. 7.8. □

The completeness together with orthonormality yields that every function $f \in L^2(\mathbb{R})$ can be expanded as the series

$$f = \sum_{k=0}^{\infty} \langle \varphi_k \,|\, f \rangle \, \varphi_k \,, \tag{1.14}$$

III:hermite-series

where the convergence of the series is understood as convergence of the partial sums in the L^2 norm.

Approximation Properties. We denote by P_K the orthogonal projector onto $\mathcal{V}_K = \text{span}(\varphi_0, \ldots, \varphi_{K-1})$, given by

$$P_K f = \sum_{k<K} \langle \varphi_k \,|\, f \rangle \, \varphi_k \,.$$

This is the best approximation to f in $\mathcal{V}_K$ with respect to the L^2 norm. We have the following approximation result, for which we recall $A = \frac{1}{\sqrt{2}}(x + d/dx)$.

hermite-approx

Theorem 1.2 (Approximation by Hermite Functions). *For every integer $s \leq K$ and every function f in the Schwartz space $\mathcal{S}$,*

$$\|f - P_K f\| \leq \frac{1}{\sqrt{K(K-1)\ldots(K-s+1)}} \, \|A^s f\| \,.$$

Given sufficient smoothness and decay of the function, the approximation error thus decays as $\mathcal{O}(K^{-s/2})$ for growing K and any fixed s.

Proof. Using subsequently (1.14), (1.11) and (1.9) we obtain

$$\begin{aligned} f - P_K f &= \sum_{k \geq K} \langle \varphi_k \,|\, f \rangle \, \varphi_k \\ &= \sum_{k \geq K} \frac{1}{\sqrt{k(k-1)\ldots(k-s+1)}} \, \langle (A^\dagger)^s \varphi_{k-s} \,|\, f \rangle \, \varphi_k \\ &= \sum_{k \geq K} \frac{1}{\sqrt{k(k-1)\ldots(k-s+1)}} \, \langle \varphi_{k-s} \,|\, A^s f \rangle \, \varphi_k \,. \end{aligned}$$

By orthonormality, this yields

$$\begin{aligned} \|f - P_K f\|^2 &\leq \frac{1}{K(K-1)\ldots(K-s+1)} \sum_{j \geq 0} \big| \langle \varphi_j \,|\, A^s f \rangle \big|^2 \\ &= \frac{1}{K(K-1)\ldots(K-s+1)} \, \|A^s f\|^2 \,, \end{aligned}$$

which is the desired result. □

Since the set of linear combinations of shifted Gaussians is known to be dense in $L^2(\mathbb{R})$ (e.g., Thaller, 2000, p. 40), it is instructive to see the action of A^s on $e^{-(x-a)^2/2}$. A short calculation yields $A\, e^{-(x-a)^2/2} = \frac{1}{\sqrt{2}}\, a\, e^{-(x-a)^2/2}$ and hence

$$A^s \, e^{-(x-a)^2/2} = 2^{-s/2} \, a^s \, e^{-(x-a)^2/2} \,.$$

No surprise, the approximation of $e^{-(x-a)^2/2}$ by Hermite functions φ_k centered at 0 is slow to converge for large shifts $|a| \gg 1$. According to Theorem 1.2, the error becomes small from $K > \frac{e}{2} a^2$ onwards (on choosing $s = K$ and using Stirling's formula for $K!$).

Error of the Galerkin Method with Hermite Basis in 1D. We are now in the position to prove the following error bound. For a related result we refer to Faou & Gradinaru (2007).

ermite-galerkin

Theorem 1.3 (Galerkin Error). *Consider the Galerkin method with the one-dimensional Hermite basis $(\varphi_0, \dots, \varphi_{K-1})$, applied to a 1D Schrödinger equation (0.1) with a potential $V(x) = (1+x^2)B(x)$ with bounded B, with initial value $\psi_K(0) = P_K\psi(0)$. Then, if the exact solution is in $D(A^{s+2})$ for some integer $s \le K/2$, the error is bounded by*

$$\|\psi_K(t) - \psi(t)\| \le C\, K^{-s/2}\,(1+t) \max_{0\le\tau\le t} \|A^{s+2}\psi(\tau)\|\,,$$

where C is independent of K and t, is bounded by $C \le c\,2^{s/2}$ in dependence of s, and depends linearly on the bound of B.

Proof. (a) We write the Galerkin equation (1.1) as

$$i\dot\psi_K = P_K H P_K \psi_K$$

with the Hermitian matrix $P_K H P_K$, and the Schrödinger equation (0.1), acted on by P_K, as

$$iP_K\dot\psi = P_K H P_K\, P_K\psi + P_K H P_K^\perp \psi\,,$$

where $P_K^\perp = I - P_K$ is the complementary orthogonal projection. Subtracting the two equations and taking the inner product with $\psi_K - P_K\psi$ yields, by the same argument as in the proof of Theorem II.1.5,

$$\|\psi_K(t) - P_K\psi(t)\| \le \|\psi_K(0) - P_K\psi(0)\| + \int_0^t \|P_K H P_K^\perp \psi(\tau)\|\, d\tau\,.$$

We show in part (b) of the proof that

$$\|P_K H P_K^\perp \psi\| \le C\, K^{-s/2}\, \|A^{s+2}\psi\|\,. \tag{1.15}$$

III:skew-est

The result then follows together with Theorem 1.2, applied with $s+2$ instead of s, to estimate $\psi(t) - P_K\psi(t)$.

(b) It remains to prove (1.15). We recall that $H = \frac{1}{2\mu}p^2 + B(1+q^2)$. By (1.6) we have

$$p^2 = -\frac{1}{2}(A - A^\dagger)^2\,, \qquad q^2 = \frac{1}{2}(A + A^\dagger)^2\,.$$

With (1.11) and (1.12) this gives

$$\begin{aligned} p^2\varphi_k &= -\frac{1}{2}\big(\sqrt{k(k-1)}\,\varphi_{k-2} - (2k+1)\varphi_k + \sqrt{(k+2)(k+1)}\,\varphi_{k+2}\big) \\ q^2\varphi_k &= \frac{1}{2}\big(\sqrt{k(k-1)}\,\varphi_{k-2} + (2k+1)\varphi_k + \sqrt{(k+2)(k+1)}\,\varphi_{k+2}\big)\,. \end{aligned}$$

This yields, with $c_k = \langle\varphi_k \,|\, \psi\rangle$,

$$P_K p^2 P_K^\perp \psi = c_K\sqrt{K(K-1)}\,\varphi_{K-2} + c_{K+1}\sqrt{(K+1)K}\,\varphi_{K-1}\,.$$

Estimating the coefficients c_k as in the proof of Theorem 1.2 with $s+2$ instead of s, we obtain

$$\|P_K p^2 P_K^\perp \psi\| \le C\, K^{-s/2}\, \|A^{s+2}\psi\|\,.$$

Similarly, we get

$$\|q^2 P_K^\perp \psi\| \le C\, K^{-s/2}\, \|A^{s+2}\psi\|\,.$$

Together with the boundedness of B, these two estimates imply the bound (1.15). □

We remark that from Theorem II.1.5, we can alternatively obtain an a posteriori error bound $C\,K^{-s/2}\,t\,\max_{0\le\tau\le t}\big(\|A^{s+2}\psi_K(\tau)\|+\|A^{s+2}B\psi_K(\tau)\|\big)$, where the approximate solution ψ_K instead of the exact solution ψ appears in the estimate.

Computation of the Matrix Elements. To compute the entries of the matrix H_K of (1.4), we split into the harmonic oscillator and the remaining potential,

$$H = D + W \equiv \frac{1}{2\mu}(p^2+q^2) + \Big(V - \frac{1}{2\mu}\,q^2\Big)\,.$$

and consider the corresponding matrices

$$D_K = \big(\langle\varphi_j \,|\, D \,|\, \varphi_k\rangle\big)_{j,k=0}^{K-1}\,, \qquad W_K = \big(\langle\varphi_j \,|\, W \,|\, \varphi_k\rangle\big)_{j,k=0}^{K-1}\,.$$

By Theorem 1.1, D_K is diagonal with entries $d_k = (k+\frac12)/\mu$. To compute W_K, we use *Gauss–Hermite quadrature*, that is, Gaussian quadrature for the weight function e^{-x^2} over $\mathbb{R}$ (see, e.g., Gautschi 1997): for $M \ge K$, let x_i ($i=1,\dots,M$) be the zeros of the Mth Hermite polynomial $H_M(x)$. With the corresponding weights w_i or $\omega_i = w_i\, e^{x_i^2}$, the quadrature formula

$$\int_{-\infty}^{\infty} e^{-x^2}\, h(x)\,dx \approx \sum_{i=1}^{M} w_i\, h(x_i) \quad \text{or} \quad \int_{-\infty}^{\infty} f(x)\,dx \approx \sum_{i=1}^{M} \omega_i\, f(x_i)$$

is exact for all polynomials h of degree up to $2M-1$. If $f(x) = g(x)\cdot e^{-x^2/2}$ with a function $g \in L^2(\mathbb{R})$ for which the coefficients $c_k = \langle\varphi_k \,|\, g\rangle$ in the Hermite expansion (1.14) of g satisfy $|c_k| \le C\,(1+k)^{-r}$ with $r>1$, we then obtain that the quadrature error is bounded by $\mathcal{O}(M^{-r})$.

We thus approximate

$$\langle\varphi_j \,|\, W \,|\, \varphi_k\rangle \approx \sum_{i=1}^{M} \omega_i\, \varphi_j(x_i)\, W(x_i)\, \varphi_k(x_i)\,, \tag{1.16}$$

III:quad

using M evaluations of the potential for all K^2 matrix elements, and evaluating $\varphi_j(x_i)$ via the recurrence relation (1.13). To obtain all matrix elements with good accuracy, one would have to choose M distinctly larger than K, but in practice a popular choice is $M = K$. Though the lower right block in the matrix is then inaccurate, this does not impair the asymptotic accuracy of the overall numerical method for large K, since the inaccurate matrix elements only meet with the small coefficients that correspond to high-order Hermite functions. This observation can be turned into rigorous estimates with the arguments of the above proofs.

ubsec:hermite-d

III.1.2 Higher Dimensions: Hyperbolic Cross and Sparse Grids

We now turn to the Galerkin method with a tensor-product Hermite basis for the d-dimensional Schrödinger equation (0.1).

Full Tensor-Product Basis. The theory of the preceding section immediately extends to a full tensor-product basis of Hermite functions: for all multi-indices $k = (k_1, \dots, k_d)$ with integers $0 \le k_n < K$, take the product functions

$$\varphi_{(k_1,\dots,k_d)}(x_1, \dots, x_d) = \varphi_{k_1}(x_1) \dots \varphi_{k_d}(x_d)$$

or briefly

$$\varphi_k = \varphi_{k_1} \otimes \dots \otimes \varphi_{k_d} \tag{1.17}$$

III:phi-tensor

as the basis functions in the Galerkin method. While this is theoretically satisfactory, it is computationally infeasible in higher dimensions: the number of basis functions, the number of coefficients, the computational work all grow like K^d, exponentially with the dimension d to the large base K.

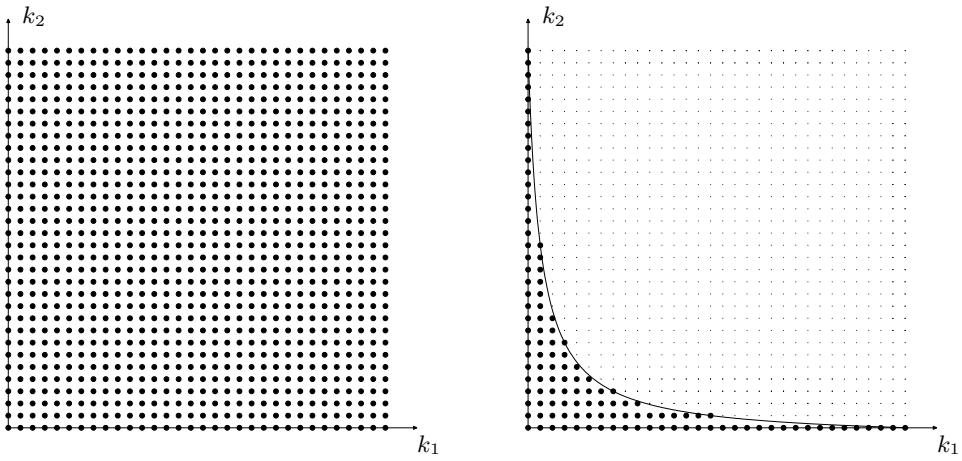

Fig. 1.2. Full and hyperbolically reduced tensor basis ($K = 32$).

III:fig:hyp

Hyperbolic Reduced Tensor-Product Basis. Instead of taking *all* tensor products with $k_j < K$, we only take a subset of multi-indices: for a bound K, let the hyperbolic multi-index set $\mathcal{K}$ be given as

$$\mathcal{K} = \mathcal{K}(d, K) = \left\{(k_1, \dots, k_d) \,:\, k_n \ge 0,\ \prod_{n=1}^{d} (1 + k_n) \le K\right\}. \tag{1.18}$$

III:hyp-set

This is illustrated for $d = 2$ and $K = 32$ in Fig. 1.2. Taking only the tensor products φ_k of (1.17) with $k \in \mathcal{K}$ as the basis functions in the Galerkin method greatly reduces their number:

III:lem:K **Lemma 1.4.** *The number $N(d,K)$ of multi-indices in $\mathcal{K}(d,K)$ is bounded by*

$$N(d,K) \le K\,(\log K)^{d-1}\,. \tag{1.19}$$ III:card-K

Proof. We clearly have $N(1,K) = K$. We then note

$$N(2,K) \le \frac{K}{1} + \frac{K}{2} + \frac{K}{3} \cdots + \frac{K}{K} \le K \log K\,,$$

where the terms in the sum correspond to $k_2 = 0, 1, \ldots, K-1$, respectively. In general, we have

$$N(d,K) \le N(d-1,K) + N(d-1,K/2) + \cdots + N(d-1,K/K)\,,$$

which by induction leads to the stated bound. □

Computations with the Galerkin method on the reduced tensor-product approximation space

$$\mathcal{V}_{\mathcal{K}} = \text{span}\,\{\varphi_k : k \in \mathcal{K}\} \tag{1.20}$$ III:hyp-space

thus appear to become feasible up to fairly large dimension d.

Approximation Properties. Can we still get decent approximations on this reduced space? As we show next, this is possible under more stringent regularity assumptions on the functions to be approximated. We denote by $P_{\mathcal{K}}$ the orthogonal projector onto $\mathcal{V}_{\mathcal{K}}$, given by

$$P_{\mathcal{K}} f = \sum_{k \in \mathcal{K}} \langle \varphi_k \,|\, f \rangle\, \varphi_k\,.$$

We let $A_n = \frac{1}{\sqrt{2}}(x_n + d/dx_n)$ and for a multi-index $\sigma = (\sigma_1, \ldots, \sigma_d)$, we denote $A^\sigma = A_1^{\sigma_1} \ldots A_d^{\sigma_d}$. We then have the following approximation result.

ermite-approx-d **Theorem 1.5 (Approximation by the Reduced Tensor Hermite Basis).** *For every fixed integer s and every function f in the Schwartz space $\mathcal{S}(\mathbb{R}^d)$,*

$$\|f - P_{\mathcal{K}} f\| \le C(s,d)\, K^{-s/2} \max_{|\sigma|_\infty \le s} \|A^\sigma f\|\,,$$

where the maximum is taken over all $\sigma = (\sigma_1, \ldots, \sigma_d)$ with $0 \le \sigma_n \le s$ for each n.

Proof. For every multi-index $k = (k_1, \ldots, k_d)$ we define the multi-index $\sigma(k)$ by the condition $k_n - \sigma(k)_n = (k_n - s)_+$ (with $a_+ = \max\{a, 0\}$) for all $n = 1, \ldots, d$, and note that $0 \le \sigma(k)_n \le s$. Similar to the proof of Theorem 1.2 we have

$$\begin{aligned} f - P_{\mathcal{K}} f &= \sum_{k \notin \mathcal{K}} \langle \varphi_k \,|\, f \rangle\, \varphi_k \\ &= \sum_{k \notin \mathcal{K}} a_{k,s}\, \langle (A^\dagger)^{\sigma(k)} \varphi_{k-\sigma(k)} \,|\, f \rangle\, \varphi_k \\ &= \sum_{k \notin \mathcal{K}} a_{k,s}\, \langle \varphi_{k-s} \,|\, A^{\sigma(k)} f \rangle\, \varphi_k\,, \end{aligned}$$

where the coefficients $a_{k,s}$ come about by (1.11) and are given as

$$a_{k,s} = \prod_{n=1}^{d} \frac{1}{\sqrt{(1+(k_n-1)_+)\dots(1+(k_n-s)_+)}}\,.$$

They satisfy, for $k \notin \mathcal{K}$,

$$|a_{k,s}|^2 \le \frac{c(s,d)}{K^s}\,, \tag{1.21}$$ III:a-coeff

because by the definition (1.18) of $\mathcal{K}$ we have the bound, for $k \notin \mathcal{K}$ and with $r = 1, \dots, s$,

$$\prod_{n=1}^{d} \big(1+(k_n-r)_+\big) \ge K \prod_{n=1}^{d} \frac{1+(k_n-r)_+}{1+k_n} \ge K\,(r+1)^{-d}\,.$$

By orthonormality, (1.21) yields

$$\|f - P_{\mathcal{K}} f\|^2 \le \frac{c(s,d)}{K^s} \sum_k \big|\langle \varphi_k \,|\, A^{\sigma(k)} f\rangle\big|^2\,.$$

Since there are s^d different possible values of $\sigma(k)$, a crude estimation yields

$$\|f - P_{\mathcal{K}} f\|^2 \le \frac{s^d\, c(s,d)}{K^s} \max_{|\sigma|_\infty \le s} \|A^\sigma f\|^2\,,$$

which is the stated result. □

We note that for a shifted d-dimensional Gaussian $e^{-|x-a|^2/2}$, we have that $A^\sigma e^{-|x-a|^2/2} = (a/\sqrt{2})^\sigma e^{-|x-a|^2/2}$, and so we now need $K \gg \prod_{n=1}^{d}(1+|a_n|^2)$ to obtain good approximation.

Error of the Galerkin Method with Reduced Tensor Hermite Basis. With the proof of Theorem 1.3 we then obtain the following result from Theorem 1.5.

ite-galerkin-d **Theorem 1.6 (Galerkin Error).** *Consider the Galerkin method with the hyperbolically reduced tensor Hermite basis applied to a d-dimensional Schrödinger equation (0.1) with a potential $V(x) = (1+|x|^2)B(x)$ with bounded B, with initial value $\psi_{\mathcal{K}}(0) = P_{\mathcal{K}}\psi(0)$. Then, for any fixed integer s the error is bounded by*

$$\|\psi_{\mathcal{K}}(t) - \psi(t)\| \le C(s,d)\, K^{-s/2}\,(1+t) \max_{0\le\tau\le t} \max_{|\sigma|_\infty \le s+2} \|A^\sigma \psi(\tau)\|$$

with the maximum over all $\sigma = (\sigma_1, \dots, \sigma_d)$ with $0 \le \sigma_n \le s+2$ for each n. □

Numerical Integration Using Sparse Grids. The matrix elements $\langle \varphi_j \,|\, H \,|\, \varphi_k\rangle$ for $j, k \in \mathcal{K}$ contain high-dimensional integrals. These can be approximated by numerical integration on sparse grids, following Smolyak (1963), Zenger (1991), Gerstner & Griebel

(1998) and using an adaptation that takes care of the increasingly oscillatory behaviour of the high-order Hermite functions.

We describe Smolyak's sparse grid quadrature when based on one-dimensional Gauss–Hermite quadrature in every coordinate direction. For $\ell = 0, 1, 2, \ldots$, let x_i^ℓ denote the zeros of the Hermite polynomial of degree 2^ℓ, and let w_i^ℓ be the corresponding weights and $\omega_i^\ell = w_i^\ell \, e^{(x_i^\ell)^2}$, so that we have the one-dimensional 2^ℓ-point Gauss–Hermite quadrature formula

$$Q_\ell f = \sum_{i=1}^{2^\ell} \omega_i^\ell \, f(x_i^\ell) \approx \int_{-\infty}^{\infty} f(x)\, dx \,.$$

We introduce the difference formulas between successive levels,

$$\Delta_\ell f = Q_\ell f - Q_{\ell-1} f \,,$$

and for the lowest level we set $\Delta_0 f = Q_0 f$. The full tensor quadrature approximation at level L to a d-dimensional integral $\int_{\mathbb{R}^d} f(x_1, \ldots, x_d)\, dx_1 \ldots dx_d$ reads

$$Q_L \otimes \ldots \otimes Q_L f = \sum_{i_1=1}^{2^L} \ldots \sum_{i_d=1}^{2^L} \omega_{i_1}^L \ldots \omega_{i_d}^L \, f(x_{i_1}^L, \ldots, x_{i_d}^L) \,,$$

which can be rewritten as

$$Q_L \otimes \ldots \otimes Q_L f = \sum_{\ell_1=0}^{L} \ldots \sum_{\ell_d=0}^{L} \Delta_{\ell_1} \otimes \ldots \otimes \Delta_{\ell_d} f \tag{1.22}$$

III:Q-full

and uses $(2^L)^d$ grid points at which f is evaluated. This number is substantially reduced in *Smolyak's algorithm*, which neglects all contributions from the difference terms with $\ell_1 + \ldots + \ell_d > L$ and thus arrives at the quadrature formula

$$\sum_{\ell_1 + \ldots + \ell_d \le L} \Delta_{\ell_1} \otimes \ldots \otimes \Delta_{\ell_d} f \approx \int_{\mathbb{R}^d} f(x_1, \ldots, x_d)\, dx_1 \ldots dx_d \,. \tag{1.23}$$

III:smolyak

Here, f is evaluated only at the points of the *sparse grid*

$$\Gamma_L^d = \{(x_{i_1}^{\ell_1}, \ldots, x_{i_d}^{\ell_d}) : \, \ell_1 + \ldots + \ell_d \le L\} \,,$$

which has only $\mathcal{O}(2^L \cdot L^{d-1})$ points; as an illustration see Fig. III.1.2 for $L = 5$ and $d = 2$. If $f(x) = g(x) \cdot e^{-|x|^2/2}$ with a function $g \in L^2(\mathbb{R}^d)$ for which the coefficients $c_m = \langle \varphi_m \,|\, g \rangle$ in the Hermite expansion of g satisfy

$$|c_m| \le C \prod_{n=1}^{d} (1 + m_n)^{-r} \tag{1.24}$$

III:cm

with $r > 1$, then the contribution of the omitted terms with $\ell_1 + \ldots + \ell_d > L$ and hence the quadrature error can be shown, by a tedious exercise, to be bounded by $\mathcal{O}((2^L)^{-r})$.

sparse-hermite

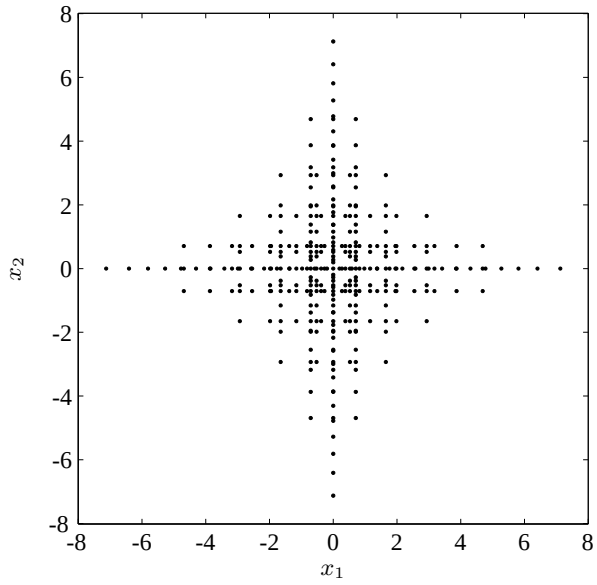

Fig. 1.3. Gauss–Hermite sparse grid ($L = 5$, $d = 2$).

Remark. A disadvantage of Gauss–Hermite quadrature formulas is the fact that they are not nested: the quadrature points of level $\ell - 1$ are not a subset of those of level ℓ. As an alternative, which will not be explored here, one might consider transformation to a finite interval and using the trapezoidal rule or Clenshaw-Curtis quadrature there. With a nested quadrature, the sparse grid contains approximately half as many grid points as for the case of a non-nested basic quadrature formula with the same number of quadrature points. It is not clear if the otherwise excellent properties of Gauss–Hermite quadrature are indeed offset by nested quadratures for suitably truncated or transformed integrals.

Computation of the Matrix Elements. The integrand f_{jk} in the matrix element

$$\langle \varphi_j \,|\, W \,|\, \varphi_k \rangle = \int_{\mathbb{R}^d} \varphi_j(x)\, W(x)\, \varphi_k(x)\, dx \equiv \int_{\mathbb{R}^d} f_{jk}(x)\, dx$$

becomes highly oscillatory for multi-indices j and k with large components. In this situation, an estimate of the type (1.24) cannot be expected to hold true with a constant that is uniform in j and k, but rather (with $a_+ = \max\{a, 0\}$)

$$|c_m(j,k)| \le C \prod_{n=1}^{d} \big(1 + (m_n - j_n - k_n)_+\big)^{-r} \tag{1.25}$$

III:cjkm

for the Hermite coefficients $c_m(j,k)$ of $g_{jk}(x) = f_{jk}(x)\, e^{|x|^2/2}$. This suggests a modification of Smolyak's algorithm in which terms in the sum (1.22) are discarded only

if they are of size $\mathcal{O}((2^L)^{-r})$ under condition (1.25). Such an adaptation of the algorithm reads as follows: for a pair of multi-indices j and k, let $\widehat{\ell}_1, \ldots, \widehat{\ell}_d$ be such that $c \cdot 2^{\widehat{\ell}_n - 1} < \max\{j_n, k_n\} \le c \cdot 2^{\widehat{\ell}_n}$ for a chosen constant c. We discard only terms with

$$(\ell_1 - \widehat{\ell}_1)_+ + \ldots + (\ell_d - \widehat{\ell}_d)_+ > L\,.$$

In the case of a hyperbolically reduced multi-index set (1.18), we have actually

$$\widehat{\ell}_1 + \ldots + \widehat{\ell}_d \le 2\log_2 K + \alpha d\,,$$

where $\alpha \in \mathbb{R}$ depends only on c. Such a modification can thus be implemented by increasing L in dependence of K by $2\log_2 K$. The number of evaluations of the potential on the resulting sparse grid thus becomes $\mathcal{O}(K^2 \cdot 2^L \cdot (L + 2\log_2 K)^{d-1})$ and hence is essentially quadratic in K of (1.18). The choice of L depends on the smoothness and growth properties of the potential.

III.1.3 Collocation Method, 1D Fourier Basis

sec:fourier-1d

Truncation, Periodization, Rescaling. We start from the one-dimensional Schrödinger equation (0.1) on the real line. If we expect the wavefunction to be negligible outside an interval $[a, b]$ on the considered time interval, we may replace the equation on the whole real line by that on the finite interval with periodic boundary conditions. After a rescaling and shift $x \to \alpha x + \beta$ we may assume that the space interval is $[-\pi, \pi]$:

$$i\frac{\partial \psi}{\partial t}(x,t) = -\frac{1}{2\mu}\frac{\partial^2 \psi}{\partial x^2}(x,t) + V(x)\psi(x,t)\,, \qquad x \in [-\pi, \pi]\,, \tag{1.26}$$

III:schrod-1d

with periodic boundary conditions: $\psi(-\pi, t) = \psi(\pi, t)$ for all t.

Collocation by Trigonometric Polynomials. We look for an approximation to the wave function $\psi(x,t)$ by a trigonometric polynomial at every instant t,

$$\psi(x,t) \approx \psi_K(x,t) = \sum_{k=-K/2}^{K/2-1} c_k(t)\, e^{ikx}\,, \qquad x \in [-\pi, \pi]\,, \tag{1.27}$$

III:trig-pol

where K is a given even integer. We might determine the unknown Fourier coefficients $c_k(t)$ by a Galerkin method on the space of trigonometric polynomials as in the previous section. Here, we consider instead the approach by *collocation*, which requires that the approximation satisfy the Schrödinger equation in a finite number of grid points, as many points as there are unknown coefficients. We thus choose the K equidistant grid points $x_j = j \cdot 2\pi/K$ with $j = -K/2, \ldots, K/2 - 1$ and require that

$$i\frac{\partial \psi_K}{\partial t}(x_j,t) = -\frac{1}{2\mu}\frac{\partial^2 \psi_K}{\partial x^2}(x_j,t) + V(x_j)\psi(x_j,t) \quad (j = -K/2, \ldots, K/2 - 1). \tag{1.28}$$

III:coll

This condition is equivalent to a system of ordinary differential equations for the coefficients $c_k(t)$, as we show next.

Discrete Fourier Transform. Let $\mathcal{F}_K : \mathbb{C}^K \to \mathbb{C}^K$ denote the *discrete Fourier transform* of length K, defined by

$$\widehat{v} = \mathcal{F}_K v \quad \text{with} \quad \widehat{v}_k = \frac{1}{K} \sum_{j=-K/2}^{K/2-1} e^{-ikj\cdot 2\pi/K}\, v_j \quad (k = -K/2, \dots, K/2-1). \tag{1.29}$$

III:dft

The inverse transform is then $\mathcal{F}_K^{-1} = K\mathcal{F}_K^*$, that is,

$$v = \mathcal{F}_K^{-1}\widehat{v} \quad \text{with} \quad v_j = \sum_{k=-K/2}^{K/2-1} e^{ijk\cdot 2\pi/K}\, \widehat{v}_k \qquad (j = -K/2, \dots, K/2-1). \tag{1.30}$$

III:dftinv

The familiar *fast Fourier transform* (FFT) algorithm (see, e.g., the informative Wikipedia article on this topic) computes either transform with $\mathcal{O}(K \log K)$ complex multiplications and additions, instead of the K^2 operations needed for a naive direct computation from the definition.

Differential Equations for the Fourier Coefficients and Grid Values. From (1.27) we note that the vector of grid values of ψ_K is the inverse discrete Fourier transform of the coefficient vector:

$$\big(\psi_K(x_j,t)\big) = \mathcal{F}_K^{-1}\big(c_k(t)\big)\,. \tag{1.31}$$

III:c-psi

This relation and differentiation of (1.27) yield that the collocation condition (1.28) is equivalent to the following differential equation for the vector $c = (c_k)$ of Fourier coefficients: with the diagonal matrices $D_K = \frac{1}{2\mu}\,\mathrm{diag}(k^2)$ and $V_K = \mathrm{diag}\big(V(x_j)\big)$,

$$i\dot{c} = D_K c + \mathcal{F}_K V_K \mathcal{F}_K^{-1} c\,. \tag{1.32}$$

III:coll-c

Alternatively, by taking the inverse Fourier transform on both sides of (1.32) and recalling (1.31), we obtain a system of differential equations for the grid values $u_j(t) = \psi_K(x_j,t)$: for the vector $u = (u_j)$,

$$i\dot{u} = \mathcal{F}_K^{-1} D_K \mathcal{F}_K u + V_K u\,. \tag{1.33}$$

III:coll-u

We observe that the matrices on the right-hand sides of (1.32) and (1.33) are all Hermitian, because $\sqrt{K}\mathcal{F}_K$ is a unitary transformation.

Approximation by Trigonometric Interpolation. For a continuous 2π-periodic function f we denote by $\mathcal{I}_K f$ the trigonometric polynomial with K Fourier modes ranging from $-K/2$ to $K/2-1$ which interpolates f in the K equidistant grid points $x_j = j\cdot 2\pi/K$:

$$\mathcal{I}_K f(x) = \sum_{k=-K/2}^{K/2-1} c_k\, e^{ikx} \quad \text{with} \quad (c_k) = \mathcal{F}_K\big(f(x_j)\big)\,.$$

III:thm:ipol **Theorem 1.7 (Interpolation Error).** *Suppose that f is a 2π-periodic function for which the s-th derivative $\partial_x^s f \in L^2$, for some $s \geq 1$. Then, the interpolation error is bounded in L^2 by*

$$\|f - \mathcal{I}_K f\| \leq C\, K^{-s}\, \|\partial_x^s f\|\,,$$

where C depends only on s.

Proof. We write the Fourier series of f and the trigonometric interpolation polynomial as

$$f(x) = \sum_{k=-\infty}^{\infty} a_k\, e^{ikx}\,, \qquad \mathcal{I}_K f(x) = \sum_{k=-K/2}^{K/2-1} c_k\, e^{ikx}\,.$$

From the interpolation condition it is verified that the coefficients are related by the *aliasing formula*

$$c_k = \sum_{\ell=-\infty}^{\infty} a_{k+\ell K}\,.$$

Using Parseval's formula and the Cauchy–Schwarz inequality, we thus obtain

$$\begin{aligned}
\|f - \mathcal{I}_K f\|^2 &= \sum_{k=-K/2}^{K/2-1} \Big(\Big| \sum_{\ell\neq 0} a_{k+\ell K} \Big|^2 + \sum_{\ell\neq 0} |a_{k+\ell K}|^2 \Big) \\
&\leq \sum_{k=-K/2}^{K/2-1} \Big(\sum_{\ell\neq 0} (k+\ell K)^{-2s} \cdot \sum_{\ell\neq 0} (k+\ell K)^{2s} |a_{k+\ell K}|^2 \\
&\qquad + \sum_{\ell\neq 0} (k+\ell K)^{-2s} \cdot (k+\ell K)^{2s} |a_{k+\ell K}|^2 \Big) \\
&\leq C^2\, K^{-2s} \sum_{k=-\infty}^{\infty} |k^s a_k|^2 = C^2\, K^{-2s}\, \|\partial_x^s f\|^2\,,
\end{aligned}$$

which is the desired result. □

In the same way it is shown that for every integer $m \geq 1$,

$$\|\partial_x^m (f - \mathcal{I}_K f)\| \leq C\, K^{-s}\, \|\partial_x^{s+m} f\|\,. \tag{1.34}$$ III:ipol-diff

Error of the Collocation Method with Fourier Basis in 1D. We obtain the following error bound.

thm:coll-error **Theorem 1.8 (Collocation Error).** *Suppose that the exact solution $\psi(t) = \psi(\cdot, t)$ has $\partial_x^{s+2}\psi(t) \in L^2$ for every $t \geq 0$, for some $s \geq 1$. Then, the error of the Fourier collocation method (1.28) with initial value $\psi_K(x,0) = \mathcal{I}_K \psi(x,0)$ is bounded in L^2 by*

$$\|\psi_K(t) - \psi(t)\| \leq C\, K^{-s}\, (1+t) \max_{0\leq\tau\leq t} \|\partial_x^{s+2}\psi(\tau)\|\,,$$

where C depends only on s.

Proof. The error analysis is based on reformulating method (1.28) as an equation with continuous argument: by interpolation on both sides of (1.28),

$$i\,\frac{\partial \psi_K}{\partial t}(x,t) = -\frac{1}{2\mu}\frac{\partial^2 \psi_K}{\partial x^2}(x,t) + \mathcal{I}_K(V\psi_K)(x,t)\,, \qquad x \in [-\pi,\pi]\,. \tag{1.35}$$

III:coll-cont

On the other hand, using that $\mathcal{I}_K V\psi = \mathcal{I}_K V \mathcal{I}_K \psi$, we obtain that the interpolant to the solution satisfies the equation

$$i\,\frac{\partial \mathcal{I}_K\psi}{\partial t}(x,t) = -\frac{1}{2\mu}\frac{\partial^2 \mathcal{I}_K\psi}{\partial x^2}(x,t) + (\mathcal{I}_K V \mathcal{I}_K\psi)(x,t) + \delta_K(x,t)\,, \tag{1.36}$$

III:coll-ipol

with the defect

$$\delta_K = -\frac{1}{2\mu}\Big(\mathcal{I}_K\frac{\partial^2\psi}{\partial x^2} - \frac{\partial^2 \mathcal{I}_K\psi}{\partial x^2}\Big)\,.$$

The error $\varepsilon_K = \psi_k - \mathcal{I}_K\psi$ thus satisfies the equation

$$i\,\frac{\partial \varepsilon_K}{\partial t} = -\frac{1}{2\mu}\frac{\partial^2 \varepsilon_K}{\partial x^2} + \mathcal{I}_K(V\varepsilon_K) - \delta_K\,.$$

In terms of the Fourier coefficients $e = (e_k)$ and $d = (d_k)$ given by

$$\varepsilon_K(x,t) = \sum_{k=-K/2}^{K/2-1} e_k(t)\,e^{ikx}\,, \qquad \delta_K(x,t) = \sum_{k=-K/2}^{K/2-1} d_k(t)\,e^{ikx}\,,$$

this reads, as in (1.32):

$$i\dot e = D_K e + \mathcal{F}_K V_K \mathcal{F}_K^{-1} e - d\,,$$

with Hermitian matrices on the right-hand side, since $\mathcal{F}_K$ is unitary. Forming the Euclidean inner product with e, taking the real part and integrating we obtain, by the same argument as in the proof of Theorem II.1.5,

$$\|e(t)\| \le \|e(0)\| + \int_0^t \|d(\tau)\|\,d\tau\,.$$

By Parseval's formula, this is the same as

$$\|\varepsilon_K(t)\| \le \|\varepsilon_K(0)\| + \int_0^t \|\delta_K(\tau)\|\,d\tau\,.$$

We estimate $\delta_K(\tau)$ using Theorem 1.7 for $\partial_x^2\psi(\cdot,\tau)$ and (1.34) with $m = 2$:

$$\|\delta_K(\tau)\| \le CK^{-s}\,\|\partial_x^{s+2}\psi(\cdot,\tau)\|\,.$$

Recalling that $\varepsilon_K = \psi_K - \mathcal{I}_K\psi$ and using Theorem 1.7 to estimate the interpolation error $\mathcal{I}_K\psi - \psi$, we obtain the stated result. □

Comparison with the Fourier Galerkin Method. If we use the Galerkin method (1.1) with the basis e^{-ikx} $(k = -K/2, \dots, K/2-1)$, then we obtain equations for the coefficients that are very similar to (1.32):

$$i\dot{c} = D_K c + \widehat{V}_K c\,. \tag{1.37}$$ III:gal-c

Here, $\widehat{V}_K$ is the matrix with the entry $\frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ijx}\,V(x)\,e^{ikx}\,dx$ at position (j,k). In the collocation method (1.32), this integral is simply replaced by the trapezoidal sum approximation $\frac{1}{K}\sum_l e^{-ikx_l}\,V(x_l)\,e^{imx_l}$, with no harm to the error of the method as Theorem 1.8 shows.

III.1.4 Higher Dimensions: Hyperbolic Cross and Sparse Grids

The above results extend immediately to a full tensor-grid approximation in higher dimensions. The number of grid points and Fourier coefficients to be dealt with is then K^d in dimension d with K grid points in each direction. An approach to a reduced computational cost uses a hyperbolically reduced tensor basis of exponentials and an associated sparse grid, leading to a discretization working with $\mathcal{O}(K(\log K)^{d-1})$ grid points and Fourier coefficients. The construction is based on a discrete Fourier transform on sparse grids given by Hallatschek (1992).

Hyperbolic Cross. Instead of considering the full tensor product basis $e^{ik\cdot x} = e^{ik_1x_1}\dots e^{ik_dx_d}$ with $-K/2 \le k_n \le K/2-1$, we consider a reduced set of multi-indices $k = (k_1,\dots,k_d)$, which is constructed as follows. We order the set of integers into different levels by setting $\mathbb{Z}_0 = \{0\}$, $\mathbb{Z}_1 = \{-1\}$, $\mathbb{Z}_2 = \{-2,1\}$, $\mathbb{Z}_3 = \{-4,-3,2,3\}$, and in general

$$\mathbb{Z}_\ell = \{k \in \mathbb{Z} : -2^{\ell-1} \le k < -2^{\ell-2} \text{ or } 2^{\ell-2} \le k < 2^{\ell-1}\}\,. \tag{1.38}$$ III:Zl

This yields a partition of the integers into different levels as indicated in the following diagram of the line of integers:

…	-8	-7	-6	-5	-4	-3	-2	-1	0	1	2	3	4	5	6	7	…
…	4	4	4	4	3	3	2	1	0	2	3	3	4	4	4	4	…

We then define the *hyperbolic cross*

$$\begin{aligned}\mathcal{K} = \mathcal{K}_L^d = \{(k_1,\dots,k_d) :\ &\text{There are } \ell_1,\dots,\ell_d \text{ with } \ell_1+\dots+\ell_d \le L \\ &\text{such that } k_n \in \mathbb{Z}_{\ell_n} \text{ for } n = 1,\dots,d\}\,.\end{aligned} \tag{1.39}$$ III:hyp-cross

We will work with the basis of exponentials $e^{ik\cdot x}$ with $k \in \mathcal{K}$. As in Lemma 1.4 it is seen that $\mathcal{K}$ has $\mathcal{O}(2^L \cdot L^{d-1})$ elements.

Sparse Grid. As we now show, the wave vectors in the hyperbolic cross are in a bijective correspondence with a set of grid points in $[0,2\pi]^d$. Consider first the hierarchical ordering of grid points in the interval $[0,2\pi)$ obtained by setting $X_0 = \{0\}$, $X_1 = \{\pi\}$, $X_2 = \{\frac{\pi}{2}, \frac{3\pi}{2}\}$, $X_3 = \{\frac{\pi}{4}, \frac{3\pi}{4}, \frac{5\pi}{4}, \frac{7\pi}{4}\}$, and in general

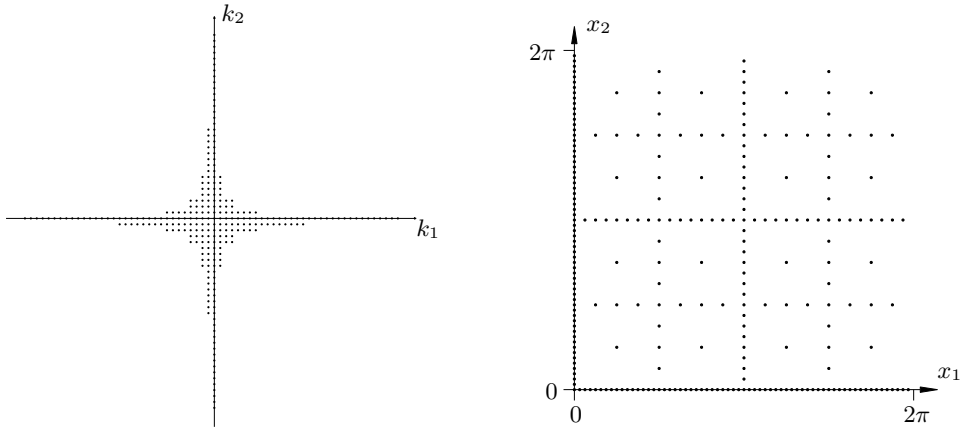

Fig. 1.4. Hyperbolic cross and sparse grid (L=6).

$$X_\ell = \Big\{ (2j-1)\frac{2\pi}{2^\ell} \,:\, j = 1, \ldots, 2^{\ell-1} \Big\} .$$

Clearly, each grid point in X_ℓ is in a one-to-one correspondence with an integer in $\mathbb{Z}_\ell$. We define the *sparse grid* corresponding to the hyperbolic cross $\mathcal{K}$ as

$$\begin{aligned} \Gamma = \Gamma_L^d = \{ (x_1, \ldots, x_d) :\ & \text{There are } \ell_1, \ldots, \ell_d \text{ with } \ell_1 + \ldots + \ell_d \le L \\ & \text{such that } \ x_n \in X_{\ell_n} \ \text{for } \ n = 1, \ldots, d \} . \end{aligned} \tag{1.40}$$

III:sparse-grid

We will use quadrature and trigonometric interpolation on this grid.

Smolyak's Sparse-Grid Quadrature. We consider the trapezoidal (or rectangle) rule approximation to the one-dimensional integral $\frac{1}{2\pi}\int_0^{2\pi} g(x)\,dx$ of a 2π-periodic function g,

$$Q_\ell g = 2^{-\ell} \sum_{j=0}^{2^\ell - 1} g\Big(j \frac{2\pi}{2^\ell} \Big) = 2^{-\ell} \sum_{m=0}^{\ell} \sum_{x \in X_m} g(x) ,$$

and the difference between two levels,

$$\Delta_\ell g = Q_\ell g - Q_{\ell-1} g , \qquad \Delta_0 g = Q_0 g .$$

As in Section III.1.2, we consider Smolyak's quadrature for a multi-variate function $f(x_1, \ldots, x_d)$, which uses values of f only on the sparse grid $\Gamma = \Gamma_L^d$:

$$S_\Gamma f = S_L^d f = \sum_{\ell_1 + \ldots + \ell_d \le L} \Delta_{\ell_1} \otimes \ldots \otimes \Delta_{\ell_d} f . \tag{1.41}$$

III:sparse-smolyak

It has the following useful property.

III:lem:exp

Lemma 1.9. *Smolyak's quadrature (1.41) is exact for the exponentials $e^{ik\cdot x}$ for all multi-indices k in the hyperbolic cross $\mathcal{K}_L^d$.*

Proof. We first note that the one-dimensional trapezoidal rule Q_ℓ gives the exact value 0 for exponentials e^{ikx} whenever k is not an integral multiple of 2^ℓ, and it gives the correct value 1 for $k = 0$. With the formula

$$S_L^d f = \sum_{\ell=0}^{L} \Delta_\ell \otimes S_{L-\ell}^{d-1} f \,,$$

the result then follows by induction over the dimension. □

III:rem:exp **Remark 1.10.** Unlike the full-grid case, the quadrature S_L^d is not exact for products $e^{-ijx}e^{ikx}$ with $j, k \in \mathcal{K}_L^d$. The problem arises with terms such as $j = (-2^{L-1}, 0, 0, \ldots, 0)$ and $k = (0, -2^{L-1}, 0, \ldots, 0)$. Since $k - j \in \mathcal{K}_{2L}^d$ for $j, k \in \mathcal{K}_L^d$, we note that such products are integrated exactly by S_{2L}^d, hence with roughly the squared number of grid points. (Cf. the similar situation in the Hermite case discussed at the end of Section III.1.2.)

Sparse-Grid Trigonometric Interpolation. The one-dimensional trigonometric interpolation of a 2π-periodic function f on a grid of 2^ℓ equidistant grid points is given as

$$I_\ell g(x) = \sum_{k=-2^{\ell-1}}^{2^{\ell-1}-1} c_k^\ell \, e^{ikx} \qquad \text{with} \quad c_k^\ell = Q_\ell(e^{-ikx} g) \,.$$

We let $\Delta_\ell = I_\ell - I_{\ell-1}$ denote the difference operators between successive levels (with $\Delta_0 = I_0$). The trigonometric interpolation of a multivariate function f on the full tensor grid with 2^L grid points in every coordinate direction can then be written as

$$\sum_{\ell_1=0}^{L} \ldots \sum_{\ell_d=0}^{L} \Delta_{\ell_1} \otimes \ldots \otimes \Delta_{\ell_d} f(x_1, \ldots, x_d) \,.$$

Hallatschek (1992) introduces the corresponding operator with evaluations of f only on the sparse grid $\Gamma = \Gamma_L^d$ as

$$\mathcal{I}_\Gamma f(x_1, \ldots, x_d) = \sum_{\ell_1 + \ldots + \ell_d \le L} \Delta_{\ell_1} \otimes \ldots \otimes \Delta_{\ell_d} f(x_1, \ldots, x_d) \qquad (1.42)$$

III:sparse-ipol-operator

and notes the following important property.

em:sparse-ipol **Lemma 1.11.** *$\mathcal{I}_\Gamma f$ interpolates f on the sparse grid Γ.*

Proof. This follows from the observation that the terms omitted from the full-grid interpolation operator all vanish on the sparse grid. □

Sparse Discrete Fourier Transform. We observe that $\mathcal{I}_\Gamma f(x)$ for $x = (x_1, \ldots, x_d)$ is a linear combination of exponentials $e^{ik \cdot x}$ with k in the hyperbolic cross $\mathcal{K} = \mathcal{K}_L^d$:

$$\mathcal{I}_\Gamma f(x) = \sum_{k\in\mathcal{K}} c_k\, e^{ik\cdot x}\,.$$

This defines a discrete Fourier transform

$$\mathcal{F}_\Gamma : \mathbb{C}^\Gamma \to \mathbb{C}^\mathcal{K} :\ \big(f(x)\big)_{x\in\Gamma} \mapsto \big(c_k\big)_{k\in\mathcal{K}}\,. \tag{1.43}$$ III:sparse-dft

With the map that determines the grid values of a trigonometric polynomial from its coefficients,

$$\mathcal{T}_\mathcal{K} : \mathbb{C}^\mathcal{K} \to \mathbb{C}^\Gamma :\ \big(c_k\big)_{k\in\mathcal{K}} \mapsto \Big(\sum_{k\in\mathcal{K}} c_k e^{ik\cdot x}\Big)_{x\in\Gamma}\,, \tag{1.44}$$ III:sparse-idft

we have from the interpolation property that $\mathcal{T}_\mathcal{K}\mathcal{F}_\Gamma f = f$ for all $f = (f(x))_{x\in\Gamma}$, and hence $\mathcal{F}_\Gamma$ is invertible and

$$\mathcal{F}_\Gamma^{-1} = \mathcal{T}_\mathcal{K}\,. \tag{1.45}$$ III:sparse-inverse

Both $\mathcal{F}_\Gamma$ and its inverse can be implemented with $\mathcal{O}(2^L \cdot L^d)$ operations, using one-dimensional FFTs and hierarchical bases; see Hallatschek (1992) and Gradinaru (2007).

There is no discrete Parseval formula for $\mathcal{F}_\Gamma$, but by Remark 1.10, the following restricted Parseval relation is still valid: with the inner product $\langle f \,|\, g\rangle_\Gamma = S_\Gamma(\overline{f}g)$ on Γ and the Euclidean inner product $\langle\cdot\,|\,\cdot\rangle_\mathcal{K}$ on $\mathcal{K}$,

$$\langle \mathcal{F}_\Gamma^{-1}c \,|\, \mathcal{F}_\Gamma^{-1}d\rangle_\Gamma = \langle c \,|\, d\rangle_\mathcal{K} \qquad \text{if } c_k = d_k = 0 \text{ for } k \in \mathcal{K}_L^d \setminus \mathcal{K}_{L/2}^d\,. \tag{1.46}$$ III:sparse-parseval

Approximation by Sparse-Grid Trigonometric Interpolation. Error bounds are given by Hallatschek (1992) in the maximum norm, and by Gradinaru (2008) in L^2 and related norms. The L^2 error bound reads

$$\|\mathcal{I}_\Gamma f - f\| \le C(d,s)\,(L+1)^{d-1}\,(2^L)^{-s}\,\|\partial_{x_1}^{s+1}\ldots\partial_{x_d}^{s+1} f\|\,. \tag{1.47}$$ III:sparse-ipol-error

The estimate is obtained by carefully estimating the terms $\Lambda_{\ell_1}\otimes\ldots\otimes\Lambda_{\ell_d} f$ that have been omitted in (1.42).

Collocation of the Schrödinger Equation on Sparse Grids. Gradinaru (2008) studies the collocation method, which approximates the solution by a trigonometric polynomial with coefficients on the hyperbolic cross,

$$\psi_\mathcal{K}(x,t) = \sum_{k\in\mathcal{K}} c_k(t)\, e^{ik\cdot x}\,, \tag{1.48}$$ III:sparse-psiK

and requires the Schrödinger equation to hold in the points of the sparse grid. This yields the system for the Fourier coefficients $c = (c_k)_{k\in\mathcal{K}}$,

$$i\dot{c} = D_\mathcal{K} c + \mathcal{F}_\Gamma V_\Gamma \mathcal{F}_\Gamma^{-1} c\,, \tag{1.49}$$ III:sparse-ode

where $(D_\mathcal{K}c)_k = \frac{1}{2\mu}\,|k|^2 c_k$ for $k\in\mathcal{K}$, and V_Γ is the diagonal matrix with entries $V(x)$ for $x\in\Gamma$. Gradinaru (2008) shows that the error of the collocation method over bounded

time intervals is bounded by $\mathcal{O}(L^{d-1}\,(2^L)^{-s})$ if mixed derivatives up to order $s+2$ in each coordinate direction are bounded in L^2.

An unpleasant feature in (1.49) is the fact that the matrix $\mathcal{F}_\Gamma V_\Gamma \mathcal{F}_\Gamma^{-1}$ is not Hermitian, since the sparse-grid Fourier transform $\mathcal{F}_\Gamma$ is not a scalar multiple of a unitary operator, unlike the full tensor-grid case. This can give numerical artefacts such as the loss of conservation of norm and in theory may lead to an exponential, instead of linear, error growth in time, with a rate that is given by a bound of the skew-Hermitian part of $\mathcal{F}_\Gamma V_\Gamma \mathcal{F}_\Gamma^{-1}$. Moreover, some of the time-stepping methods considered in the subsequent sections are not applicable in the case of non-Hermitian matrices.

Discretizations on Sparse Grids Having Hermitian Matrices. Are there methods with similar complexity and approximation properties to the sparse-grid collocation method but which have a Hermitian matrix? We start from the interpretation of the collocation method as a Galerkin method with trapezoidal rule approximation of the integrals in the matrix elements, as noted at the end of Section III.1.3, and aim for a multi-dimensional, sparse-grid extension that approximates the matrix elements by Smolyak's quadrature.

We consider the inner product on $\mathbb{C}^\Gamma$ defined by Smolyak's quadrature on the sparse grid,

$$\langle f \,|\, g\rangle_\Gamma = S_\Gamma(\overline{f}g)\,,$$

and the Euclidean inner product $\langle\cdot\,|\,\cdot\rangle_\mathcal{K}$ on $\mathbb{C}^\mathcal{K}$. With respect to these inner products, we take the adjoint $(\mathcal{F}_\Gamma^{-1})^*$ of $\mathcal{F}_\Gamma^{-1}$:

$$\langle \mathcal{F}_\Gamma^{-1} a \,|\, f\rangle_\Gamma = \langle a \,|\, (\mathcal{F}_\Gamma^{-1})^* f\rangle_\mathcal{K} \qquad \forall\, f\in\mathbb{C}^\Gamma,\ a\in\mathbb{C}^\mathcal{K}.$$

Then, $(\mathcal{F}_\Gamma^{-1})^* f = \big(S_\Gamma(e^{-ik\cdot x} f)\big)_{k\in\mathcal{K}}$, and we obtain that

$$(\mathcal{F}_\Gamma^{-1})^* V_\Gamma \mathcal{F}_\Gamma^{-1} = \Big(S_\Gamma\big(e^{-ij\cdot x} V(x)\, e^{ik\cdot x}\big)\Big)_{j,k\in\mathcal{K}}$$

is the Hermitian matrix that contains the sparse-grid quadrature approximations to the Galerkin matrix elements.

Instead of (1.49) we would like to determine the coefficients of (1.48) from

$$i\dot{c} = D_\mathcal{K} c + (\mathcal{F}_\Gamma^{-1})^* V_\Gamma \mathcal{F}_\Gamma^{-1} c\,. \tag{1.50}$$

III:sparse-ode-symm

This method can be rewritten as a quasi-Galerkin method on the hyperbolic-cross space $\mathcal{V}_\mathcal{K} = \operatorname{span}\{e^{ik\cdot x} :\, k\in\mathcal{K}\}$: determine $\psi_\mathcal{K}(t)\in\mathcal{V}_\mathcal{K}$ (i.e., of the form (1.48)) such that

$$\Big\langle \varphi_\mathcal{K} \,\Big|\, i\,\frac{\partial\psi_\mathcal{K}}{\partial t}\Big\rangle = \Big\langle \varphi_\mathcal{K} \,\Big|\, -\frac{1}{2\mu}\Delta\psi_\mathcal{K}\Big\rangle + \Big\langle \varphi_\mathcal{K}\,\Big|\, V\psi_\mathcal{K}\Big\rangle_\Gamma \qquad \forall\,\varphi_\mathcal{K}\in\mathcal{V}_\mathcal{K}\,. \tag{1.51}$$

III:sparse-qgal

Here, the last inner product is the discrete inner product on the sparse grid instead of the usual L^2 inner product. Unfortunately, it appears that this does *not* give a convergent discretization for the hyperbolic cross $\mathcal{K}=\mathcal{K}_L^d$ and the sparse grid $\Gamma=\Gamma_L^d$ of the same level L. We describe three ways to cope with this difficulty:

1. *Discrete Galerkin Method with a Simplified Mass Matrix:* We replace the L^2 inner products in (1.51) by the discrete inner product on $\varGamma = \varGamma_L^d$. Then we obtain a standard Galerkin method with a discrete inner product. The associated orthogonal projection to $\mathcal{V}_\mathcal{K}$ is just the interpolation $\mathcal{I}_\varGamma$. Optimal error bounds are then obtained with the standard proof for Galerkin methods, as in Theorem 1.3. However, since the exponentials $e^{ik\cdot x}, k \in \mathcal{K}$, do not form an orthonormal basis with respect to the discrete inner product, there are now non-diagonal matrices

$$M_\mathcal{K} = \big(m_{jk}\big)_{j,k\in\mathcal{K}} = \big(\langle e^{ij\cdot x} \,|\, e^{ik\cdot x}\rangle_\varGamma\big)_{j,k\in\mathcal{K}}\,, \qquad T_\mathcal{K} = \frac{1}{2\mu}\big(j\cdot k\, m_{jk}\big)_{j,k\in\mathcal{K}}$$

in the differential equations for the coefficients:

$$M_\mathcal{K}\dot{c} = T_\mathcal{K} c + (\mathcal{F}_\varGamma^{-1})^* V_\varGamma \mathcal{F}_\varGamma^{-1} c\,.$$

By (1.46), the mass matrix partitioned into blocks corresponding to $\mathcal{K}_{L/2}^d$ and $\mathcal{K}_L^d \setminus \mathcal{K}_{L/2}^d$ takes the form

$$M_\mathcal{K} = \begin{pmatrix} I & B^T \\ B & N \end{pmatrix}$$

with sparse matrices B and N. An approximate Choleski factor of $M_\mathcal{K}$ is given by

$$C = \begin{pmatrix} I & 0 \\ B & I \end{pmatrix} \quad \text{with} \quad C^{-1} = \begin{pmatrix} I & 0 \\ -B & I \end{pmatrix} \quad \text{and} \quad CC^T = \begin{pmatrix} I & B^T \\ B & I + BB^T \end{pmatrix},$$

where only the lower diagonal block differs from that in $M_\mathcal{K}$. Replacing $M_\mathcal{K}$ by CC^T, we obtain for $b = C^T c$

$$\dot{b} = C^{-1} T_\mathcal{K} C^{-T} b + C^{-1} (\mathcal{F}_\varGamma^{-1})^* V_\varGamma \mathcal{F}_\varGamma^{-1} C^{-T} b\,.$$

Since only the lower diagonal block of $M_\mathcal{K}$ has been changed, we can still get error bounds as for the full Galerkin method, but with 2^{-L} replaced by $2^{-L/2}$.

2. *Discrete Galerkin Method with Refined Sparse Grid.* By Lemma 1.9, the mass matrix becomes the identity matrix if we choose the finer grid

$$\varGamma = \varGamma_{2L}^d$$

with $2L$ instead of L levels and thus, alas, roughly the squared number of grid points. In that case, the L^2 inner products (1.51) are equal to the discrete inner products on $\varGamma$, and we obtain a standard Galerkin method with a discrete inner product. The associated orthogonal projection to $\mathcal{V}_\mathcal{K}$ is $P_\mathcal{K}\mathcal{I}_\varGamma$, where $P_\mathcal{K}$ is the orthogonal projection with respect to the L^2 inner product. Optimal error bounds are then obtained with the standard proof for Galerkin methods, as in Theorem 1.3.

3. *Galerkin Method with an Approximated Potential.* We use the standard Galerkin method with L^2 inner products, and compute the matrix elements of the potential, $\langle e^{ij\cdot x} \,|\, V \,|\, e^{ik\cdot x}\rangle$, *exactly* for an approximated potential $V(x) \approx \sum_{m\in\mathcal{M}} v_m e^{im\cdot x}$ (possibly over a coarser hyperbolic cross $\mathcal{M} \subset \mathcal{K}$), noting that $\langle e^{ij\cdot x} \,|\, e^{im\cdot x} e^{ik\cdot x}\rangle \neq 0$ only for $j = k + m$. This requires $\mathcal{O}(\#\mathcal{M}\cdot\#\mathcal{K})$ operations for computing a matrix-vector product.

III.2 Polynomial Approximations to the Matrix Exponential

After space discretization, we are left with a linear system of differential equations

$$i\dot{y} = Ay \tag{2.1}$$ III:lin-ode

with a Hermitian matrix A of large dimension and of large norm, such as (1.3) or (1.32) or (1.50). The solution to the initial value $y(0) = y_0$ is given by the matrix exponential

$$y(t) = e^{-itA} y_0 \,. \tag{2.2}$$ III:matrix-exp

We study time stepping methods that advance the approximate solution[1] from time t^n to $t^{n+1} = t^n + \Delta t$, from y^n to y^{n+1}. In the present section we consider methods that require only multiplications of the matrix A with vectors, and hence are given by polynomial approximations $P(\Delta t A)$ to the exponential:

$$y^{n+1} = P(\Delta t A)\, y^n \,.$$

We consider in detail the *Chebyshev method*, where the polynomial is chosen *a priori* from given information on the extreme eigenvalues of A, and the *Lanczos method*, where the polynomial is determined by a Galerkin method on the Krylov subspace, which consists of the products of all polynomials of $\Delta t A$ of a given degree with the starting vector.

We mention in passing that there are further interesting methods that require only matrix-vector products with A: the *Leja point method* has similar approximation properties to the Chebyshev method but in contrast to the Chebyshev method, higher-degree polynomials of the family are constructed by reusing the computations for the lower-degree polynomials, cf. Caliari, Vianello & Bergamaschi (2004); explicit *symplectic methods* preserve the symplectic structure of the differential equation, see Gray & Manolopoulos (1996) and Blanes, Casas & Murua (2006).

III.2.1 Chebyshev Method

A near-optimal polynomial approximation to the exponential is given by its truncated Chebyshev expansion. We describe this approach, which in the context of Schrödinger equations was put forward by Tal-Ezer & Kosloff (1984), and give an error analysis based on Bernstein's theorem on polynomial approximations to analytic functions on an interval. We refer to Rivlin (1990) for background information on Chebyshev polynomials and to Markushevich (1977), Chap. III.3, for the polynomial approximation theory based on Faber polynomials.

Chebyshev Polynomials. For every non-negative integer k, the function defined by

$$T_k(x) = \cos(k\theta) \quad \text{with} \quad \theta = \arccos x \in [0, \pi], \quad \text{for } x \in [-1, 1] \tag{2.3}$$ III:cheb-cos

[1] The time step number n will always be indicated as superscript in the notation.

is in fact a polynomial of degree k, named the kth *Chebyshev polynomial*. This fact is seen from the recurrence relation

$$T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)\,, \qquad k \geq 1\,, \tag{2.4}$$ III:cheb-rec

starting from $T_0(x) = 1$ and $T_1(x) = x$, which is obtained from the trigonometric identity $\cos((n+1)\theta) + \cos((n-1)\theta) = 2\cos\theta\,\cos(n\theta)$. The Chebyshev polynomials are orthogonal polynomials with respect to the weight function $(1-x^2)^{-1/2}$ on $[-1,1]$:

$$\int_{-1}^{1} T_j(x)\,T_k(x)\,\frac{dx}{\sqrt{1-x^2}} = 0 \qquad \text{for} \quad j \neq k\,, \tag{2.5}$$ III:cheb-orth

as is seen by substituting $x = \cos\theta$ and $dx/\sqrt{1-x^2} = d\theta$ and using the orthogonality of the complex exponentials.

Another useful formula is

$$2\,T_k(x) = \bigl(x + \sqrt{x^2-1}\bigr)^k + \bigl(x - \sqrt{x^2-1}\bigr)^k, \tag{2.6}$$ III:cheb-sqrt-formula

again verified by substituting $x = \cos\theta$. The *Joukowski transform*

$$w = \Phi(z) = z + \sqrt{z^2-1}\,, \qquad z = \Psi(w) = \frac{1}{2}\Bigl(w + \frac{1}{w}\Bigr) \tag{2.7}$$ III:cheb-joukowski

is the conformal map between the exterior of the interval $[-1,1]$ and the exterior of the unit disk, $|w| > 1$. (The branch of the square root is chosen such that $\sqrt{z^2-1} \sim z$ for $z \to \infty$.) The level sets $\Gamma_r = \{z : |\Phi(z)| = r\} = \{\Psi(w) : |w| = r\}$ for $r > 1$ are ellipses with foci ± 1, major semi-axis $r + r^{-1}$ and minor semi-axis $r - r^{-1}$. Since the Laurent expansion at ∞ of $(z - \sqrt{z^2-1})^k$ contains only powers z^{-j} with $j \geq k$, the integral of that function over a closed contour Γ encircling the interval $[-1,1]$ vanishes by Cauchy's theorem. With Cauchy's integral formula we thus obtain from (2.6)

$$2\,T_k(x) = \frac{1}{2\pi i}\int_\Gamma \frac{\Phi(z)^k}{z-x}\,dz\,, \qquad x \in [-1,1], \tag{2.8}$$ III:cheb-faber

which establishes an important relationship between the Chebyshev polynomials and the conformal map: the Chebyshev polynomials are the *Faber polynomials* for the interval $[-1,1]$; cf. Markushevich (1977), Sect. III.3.14.

Chebyshev and Fourier Series. Given a (smooth) complex-valued function $f(x)$ on the interval $-1 \leq x \leq 1$, we expand the 2π-periodic, symmetric function

$$g(\theta) = f(\cos\theta)$$

as a Fourier series:

$$g(\theta) = \sum_{k=-\infty}^{\infty} c_k\,e^{ik\theta} \qquad \text{with} \quad c_k = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{-ik\theta}\,g(\theta)\,d\theta$$

or in fact, by the symmetry $g(-\theta) = g(\theta)$,

$$g(\theta) = c_0 + 2\sum_{k=1}^{\infty} c_k \cos(k\theta) \qquad \text{with} \quad c_k = \frac{1}{\pi}\int_0^{\pi} \cos(k\theta)\, g(\theta)\, d\theta\,.$$

Substituting $x = \cos\theta$ and $dx/\sqrt{1-x^2} = d\theta$, we obtain the *Chebyshev expansion*

$$f(x) = c_0 + 2\sum_{k=1}^{\infty} c_k\, T_k(x) \qquad \text{with} \quad c_k = \frac{1}{\pi}\int_{-1}^{1} T_k(x)\, f(x)\, \frac{dx}{\sqrt{1-x^2}}\,. \tag{2.9}$$

III:cheb-series

Chebyshev Approximation of Holomorphic Functions. We study the approximation of a holomorphic function $f(x)$ by the truncated series with m terms,

$$\Pi_m f(x) = c_0 + 2\sum_{k=1}^{m-1} c_k\, T_k(x)\,,$$

which is a polynomial of degree $m-1$. The following is a version of a theorem by Bernstein (1912); see Markushevich (1977), Sect. III.3.15. Here $\Phi(z) = z + \sqrt{z^2-1}$ is again the conformal map (2.7) from the complement of the interval $[-1,1]$ to the exterior of the unit disk, and $\Psi(w) = \frac{1}{2}(w + \frac{1}{w})$ is the inverse map.

:thm:bernstein

Theorem 2.1 (Chebyshev Approximation). *Let $r > 1$, and suppose that $f(z)$ is holomorphic in the interior of the ellipse $|\Phi(z)| < r$ and continuous on the closure. Then, the error of the truncated Chebyshev series is bounded by*

$$|f(x) - \Pi_m f(x)| \le 2\,\mu(f,r)\,\frac{r^{-m}}{1-r^{-1}} \qquad \textit{for} \quad -1 \le x \le 1,$$

with the mean value $\mu(f,r) = \frac{1}{2\pi r}\int_{|w|=r} |f(\Psi(w))|\cdot|dw|$.

Proof. We start from the Cauchy integral formula over the ellipse $\Gamma_r = \{z : |\Phi(z)| = r\} = \{\Psi(w) : |w| = r\}$ and substitute $z = \Psi(w)$:

$$f(x) = \frac{1}{2\pi i}\int_{\Gamma_r} \frac{f(z)}{z-x}\,dz = \frac{1}{2\pi i}\int_{|w|=r} f(\Psi(w))\,\frac{\Psi'(w)}{\Psi(w)-x}\,dw\,. \tag{2.10}$$

III:cheb-f-int

We expand in negative powers of w,

$$\frac{\Psi'(w)}{\Psi(w)-x} = \sum_{k=0}^{\infty} a_k(x)\, w^{-k-1} \qquad \text{for} \quad |w| > 1, \tag{2.11}$$

III:cheb-res

where the Taylor coefficients at ∞ are given as

$$a_k(x) = \frac{1}{2\pi i}\int_{|w|=r} w^k\,\frac{\Psi'(w)}{\Psi(w)-x}\,dw = \frac{1}{2\pi i}\int_{\Gamma_r} \frac{\Phi(z)^k}{z-x}\,dz\,.$$

By (2.8), these coefficients turn out to be simply

$$a_k(x) = 2\,T_k(x).$$

Inserting (2.11) into (2.10) therefore yields

$$f(x) - \Pi_m f(x) = \frac{1}{2\pi i} \int_{|w|=r} f(\Psi(w)) \cdot 2 \sum_{k=m}^{\infty} T_k(x)\, w^{-k-1}\, dw\,.$$

Since $|T_k(x)| \le 1$ for $-1 \le x \le 1$, we have for $|w| = r > 1$

$$\Big| \sum_{k=m}^{\infty} T_k(x)\, w^{-k-1} \Big| \le \sum_{k=m}^{\infty} r^{-k-1} = \frac{r^{-m-1}}{1 - r^{-1}}\,,$$

and the result follows. □

Chebyshev Approximation of Complex Exponentials. The complex exponential $e^{i\omega x}$ is an entire function, and we can choose r in Theorem 2.1 dependent on m to balance the growth of $\mu(e^{i\omega z}, r)$ with r against the decay of r^{-m}. This gives the following corollary showing superlinear convergence after a stagnation up to $m \approx |\omega|$. Since the polynomial must capture the extrema and zeros of $\cos(\omega x)$ and $\sin(\omega x)$ for a uniform approximation, it is obvious that at least a degree m proportional to $|\omega|$ is needed to obtain an error uniformly smaller than 1. Once this barrier is surmounted, the error decays very rapidly with growing degree m.

II:thm:cheb-exp

Theorem 2.2 (Eventual Superlinear Convergence to $\mathbf{e^{i\omega x}}$). *The error of the Chebyshev approximation $p_{m-1}(x)$ of degree $m-1$ to the complex exponential $e^{i\omega x}$ with real ω is bounded by*

$$\max_{-1\le x\le 1} |p_{m-1}(x) - e^{i\omega x}| \le 4 \left(e^{1-(\omega/2m)^2}\, \frac{|\omega|}{2m} \right)^m \quad \textit{for} \quad m \ge |\omega|\,. \tag{2.12}$$

III:cheb-exp-error

Proof. We have $\mu(e^{i\omega z}, r) \le \max_{z\in I_r} |e^{i\omega z}| = e^{|\omega|(r-r^{-1})/2}$, where the maximum is attained at $z = \pm\frac{1}{2}(ir + \frac{1}{ir})$ on the minor semi-axis. Theorem 2.1 thus gives us the bound

$$\max_{-1\le x\le 1} |p_{m-1}(x) - e^{i\omega x}| \le \frac{2r^{-m}}{1 - r^{-1}}\, e^{|\omega|(r-r^{-1})/2}\,.$$

Choosing $r = 2m/|\omega| \ge 2$ then yields the stated result, which could be slightly refined. □

The Chebyshev coefficients of $e^{i\omega x}$ are given explicitly by Bessel functions of the first kind: by formula (9.1.21) in Abramowitz & Stegun (1965),

$$c_k = \frac{1}{\pi} \int_0^{\pi} e^{i\omega\cos\theta}\, \cos(k\theta)\, d\theta = i^k J_k(\omega)\,. \tag{2.13}$$

III:cheb-bessel

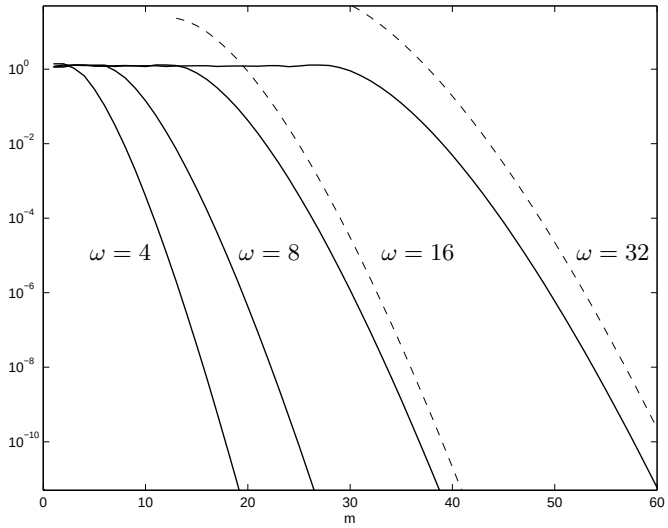

Fig. 2.1. Chebyshev approximation of $e^{i\omega x}$. Maximum error on $[-1, 1]$ versus degree, for $\omega = 4, 8, 16, 32$. Dashed: Error bounds of Theorem 2.2.

From $e^{i\omega x}$ with $-1 \le x \le 1$, uniform polynomial approximation of $e^{-i\xi}$ for $\alpha \le \xi \le \beta$ is obtained by transforming

$$x = \frac{2}{\beta - \alpha}\Big(\xi - \frac{\alpha+\beta}{2}\Big), \quad \xi = \frac{\alpha+\beta}{2} + x\,\frac{\beta-\alpha}{2}.$$

We then approximate $e^{-i\xi} = e^{-i(\alpha+\beta)/2}\,e^{-ix(\beta-\alpha)/2}$ using $e^{-ix(\beta-\alpha)/2} \approx c_0 + 2\sum_{k=1}^{m-1} c_k\,T_k(x)$ with $c_k = i^k J_k\big(-\frac{\beta-\alpha}{2}\big) = (-i)^k J_k\big(\frac{\beta-\alpha}{2}\big)$, so that

$$e^{-i\xi} \approx e^{-i(\alpha+\beta)/2}\left(c_0 + 2\sum_{k=1}^{m-1} c_k\,T_k\Big(\frac{2}{\beta-\alpha}\Big(\xi - \frac{\alpha+\beta}{2}\Big)\Big)\right) \quad \text{for} \quad \alpha \le \xi \le \beta\,.$$

Chebyshev Method for the Matrix Exponential Operator. Let A be a Hermitian matrix all of whose eigenvalues are known to lie in the interval $[a, b]$. As proposed by Tal-Ezer & Kosloff (1984), we approximate the action of the matrix exponential on a vector v by

$$e^{-i\Delta t A}v \approx e^{-i\Delta t(a+b)/2}\left(c_0 v + 2\sum_{k=1}^{m-1} c_k\,T_k\Big(\frac{2}{(b-a)}\Big(A - \frac{(a+b)}{2}I\Big)\Big)v\right) \quad (2.14)$$

III:cheb-exp-A

with $c_k = (-i)^k J_k(\Delta t(b-a)/2)$. We denote the right-hand side of (2.14) as $P_{m-1}(\Delta t\,A)v$ and observe that it is in fact a function of the product $\Delta t\,A$. The actual way to compute (2.14) is by a recursive algorithm proposed by Clenshaw (1962) for the evaluation of truncated Chebyshev expansions of functions:

Clenshaw Algorithm: let $X = \frac{2}{(b-a)}\Big(A - \frac{(a+b)}{2} I\Big)$, set $d_{m+1} = d_m = 0$ and

$$d_k = c_k v + 2X d_{k+1} - d_{k+2} \qquad \text{for} \quad k = m-1, m-2, \ldots, 0\,.$$

The approximation (2.14) is then given as

$$P_{m-1}(\Delta t\, A)v = d_0 - d_2\,.$$

This identity is readily verified using the Chebyshev recurrence relation (2.4) for the terms in the sum, descending from the terms of highest degree. The algorithm requires m matrix-vector multiplications to compute $P_{m-1}(\Delta t\, A)v$ and needs to keep only three vectors in memory.

thm:cheb-method

Theorem 2.3 (Error of the Chebyshev Method). *Let A be a Hermitian matrix with all its eigenvalues in the interval $[a, b]$, and let v be a vector of unit Euclidean norm. Then, the error of the Chebyshev approximation (2.14) is bounded in the Euclidean norm by*

$$\|P_{m-1}(\Delta t\, A)v - e^{-i\Delta t A}v\| \le 4\left(e^{1-(\omega/2m)^2}\,\frac{\omega}{2m}\right)^m \qquad \textit{for} \quad m \ge \omega$$

with $\omega = \Delta t\,(b-a)/2$.

Proof. For a diagonal matrix A, the estimate follows immediately from Theorem 2.2 and the linear transformation between the intervals $[\Delta t\, a, \Delta t\, b]$ and $[-1, 1]$. Since every Hermitian matrix A can be unitarily transformed to diagonal form, we obtain the result as stated. □

Step Size Restriction. The condition $m \ge \omega$ can be read as a restriction of the step size for given degree m:

$$\Delta t \le \frac{2m}{b-a}\,.$$

This can also be viewed as saying that at least one matrix-vector multiplication is needed on every time interval of length $1/(b-a)$. In the treatment of the Schrödinger equation, this length shrinks as the spatial discretization is refined: for illustration, consider Fourier collocation in one space dimension, with K Fourier modes. For the matrix $A = D_K + \mathcal{F}_K V_K \mathcal{F}_K^{-1}$ of (1.32), the eigenvalues lie in the interval $[a, b]$ with

$$a = \min_x V(x)\,, \qquad b = \frac{1}{2\mu}\frac{K^2}{4} + \max_x V(x)\,.$$

For large K, or small $\Delta x = 2\pi/K$, we have that $\omega = \Delta t(b-a)/2$ is approximately proportional to $\Delta t\, K^2$, or

$$\omega \sim \frac{\Delta t}{\Delta x^2}\,.$$

The condition $m \ge 2\omega$ for the onset of error reduction therefore translates into a step-size restriction

$$\Delta t \le C\, m\, \Delta x^2\,, \tag{2.15}$$

III:cheb-dtdx

and the number of matrix-vector multiplications to cover a fixed time interval is thus inversely proportional to Δx^2.

III.2.2 Lanczos Method

A different approach to approximately computing $e^{-i\Delta t A}v$ using only the action of A on vectors is based on a Galerkin approximation to $i\dot{y} = Ay$ on the Krylov space spanned by $v, Av, \ldots, A^{m-1}v$. A suitable basis for this space is given by the Lanczos iteration, named after Lanczos (1950), which has become a classic in numerical linear algebra primarily because of its use for eigenvalue problems and solving linear systems; see, e.g., Golub & Van Loan (1996), Chap. 9, and Trefethen & Bau (1997), Chap. VI. The use of the Lanczos method for approximating $e^{-i\Delta t A}v$ was first proposed by Park & Light (1986), properly in the context of approximating the evolution operator of the Schrödinger equation. Krylov subspace approximation to the matrix exponential operator has since been found useful in a variety of application areas — and has been honourably included as the twentieth of the "Nineteen dubious ways to compute the exponential of a matrix" by Moler & Van Loan (2003). Error analyses, both for the Hermitian and non-Hermitian case, have been given by Druskin & Knizhnerman (1995), Hochbruck & Lubich (1997), and Saad (1992).

Krylov Subspace and Lanczos Basis. Let A be an $N \times N$ Hermitian matrix, and let v be a non-zero complex N-vector. The mth *Krylov subspace* of $\mathbb{C}^N$ with respect to A and v is

$$\mathcal{K}_m(A, v) = \text{span}(v, Av, A^2v, \ldots, A^{m-1}v)\,, \tag{2.16}$$

III:krylov-space

that is, the space of all polynomials of A up to degree $m-1$ acting on the vector v.

The *Hermitian Lanczos method* builds an orthonormal basis of this space by Gram-Schmidt orthogonalization: beginning with $v_1 = v/\|v\|$, it constructs v_{k+1} recursively for $k = 1, 2, \ldots$ by orthogonalizing Av_k against the previous v_j and normalizing:

$$\tau_{k+1,k}\, v_{k+1} = Av_k - \sum_{j=1}^{k} \tau_{jk}\, v_j \tag{2.17}$$

III:krylov-lanczos-iter

with $\tau_{jk} = v_j^* A v_k$ for $j \le k$, and with $\tau_{k+1,k} > 0$ determined such that v_{k+1} is of unit Euclidean norm — unless the right-hand side is zero, in which case the dimension of $\mathcal{K}_m(A, v)$ is k for $m \ge k$ and the process terminates.

By the mth step, the method generates the $N \times m$ matrix $V_m = (v_1 \ldots v_m)$ having the orthonormal Lanczos vectors v_k as columns, and the $m \times m$ matrix $T_m = (\tau_{jk})$ with $\tau_{jk} = 0$ for $j - k > 1$. Because of (2.17), these matrices are related by

$$AV_m = V_m T_m + \tau_{m+1,m} v_{m+1} e_m^T, \tag{2.18}$$

III:krylov-AV

where $e_m^T = (0 \ldots 0\, 1)$ is the mth unit vector. By the orthonormality of the Lanczos vectors v_k, this equation implies

$$T_m = V_m^* A V_m\,, \tag{2.19}$$

III:krylov-T

which shows in particular that T_m is a Hermitian matrix, and hence a tridiagonal matrix: $\tau_{jk} = 0$ for $|j - k| > 1$. The sum in (2.17) therefore actually contains only the two terms

for $j = k-1, k$. For a careful practical implementation, error propagation and the loss of orthogonality due to rounding errors are a concern for larger m; see Golub & Van Loan (1996), Sect. 9.2.

Galerkin Method on the Krylov Subspace. Following Park & Light (1986), we consider the Galerkin method (1.1) for the approximation of the initial value problem

$$i\dot{y} = Ay\,, \quad y(0) = v \qquad \text{with} \quad \|v\| = 1$$

on the Krylov subspace $\mathcal{K}_m(A, v)$ with $m \ll N$ ($m \leq 20$, say): we determine an approximation $u_m(t) \in \mathcal{K}_m(A, v)$ with $u_m(0) = v$ such that at every instant t, the time derivative satisfies

$$\langle w_m \,|\, i\dot{u}_m(t) - Au_m(t)\rangle = 0 \qquad \forall\, w_m \in \mathcal{K}_m(A, v)\,.$$

Writing $u_m(t)$ in the Lanczos basis,

$$u_m(t) = \sum_{k=1}^{m} c_k(t)\, v_k = V_m c(t) \qquad \text{with} \quad c(t) = \big(c_k(t)\big)\,,$$

we obtain for the coefficients the linear differential equation

$$i\dot{c}(t) = T_m c(t)\,, \qquad c(0) = e_1 = (1, 0, \ldots, 0)^T$$

with the Lanczos matrix $T_m = (v_j^* A v_k)_{j,k=1}^m$ of (2.19). Clearly, the solution is given by $c(t) = e^{-itT_m} e_1$. The Galerkin approximation $u_m(t) = V_m c(t)$ at time Δt is thus the result of the

$$\textit{Lanczos method:} \qquad e^{-i\Delta t A} v \approx V_m e^{-i\Delta t T_m}\, e_1\,. \tag{2.20}$$

III:krylov-exp

For the small tridiagonal Hermitian matrix T_m, the exponential is readily computed from a diagonalization of T_m. The algorithm needs to keep all the Lanczos vectors in memory, which may not be feasible for large problems. In such a situation, the Lanczos iteration may be run twice with only four vectors in memory: in a first run for computing T_m, and in a second run (without recomputing the already known inner products) for forming the linear combination of the Lanczos vectors according to (2.20).

By the interpretation of (2.20) as a Galerkin method, we know from Sect. II.1 that norm and energy are preserved.

A Posteriori Error Bound and Stopping Criterion. From Theorem II.1.5 with the Krylov subspace as approximation space we have the error bound

$$\|u_m(t) - y(t)\| \leq \int_0^t \operatorname{dist}\big(Au_m(s), \mathcal{K}_m(A, v)\big)\, ds\,.$$

By (2.18) we have

$$Au_m(s) = AV_m\, e^{-isT_m}\, e_1 = V_m T_m\, e^{-isT_m}\, e_1 + \tau_{m+1,m}\, v_{m+1}\, e_m^T\, e^{-isT_m}\, e_1$$

and therefore

$$\operatorname{dist}\bigl(Au_m(s), \mathcal{K}_m(A,v)\bigr) = \tau_{m+1,m}\, \bigl|\bigl[e^{-isT_m}\bigr]_{m,1}\bigr|\,,$$

where $[\cdot]_{m.1}$ denotes the $(m,1)$ element of a matrix. This gives us the following computable error bound.

hm:krylov-apost

Theorem 2.4 (A Posteriori Error Bound). *Let A be a Hermitian matrix, and v a vector of unit Euclidean norm. Then, the error of the mth Lanczos approximation to $e^{-i\Delta t A}v$ is bounded by*

$$\|V_m e^{-i\Delta t T_m}\, e_1 - e^{-i\Delta t A} v\| \le \tau_{m+1,m} \int_0^{\Delta t} \bigl|\bigl[e^{-isT_m}\bigr]_{m,1}\bigr|\, ds\,. \qquad \square$$

If we approximate the integral on the right-hand side by the right-endpoint rectangle rule, we arrive at a *stopping criterion* for the Lanczos iteration (for given Δt) or alternatively at a *step-size selection criterion* (for given m),

$$\Delta t\, \tau_{m+1,m} \bigl|\bigl[e^{-i\Delta t\, T_m}\bigr]_{m,1}\bigr| \le \mathit{tol}$$

for an error tolerance *tol*, or without the factor Δt for an error tolerance per unit step. This criterion has previously been considered with different interpretations by Saad (1992) and Hochbruck, Lubich & Selhofer (1998). In view of Theorem 2.4, a safer choice would be to take a quadrature rule with more than one function evaluation. With a diagonalized T_m, this is inexpensive to evaluate.

Lanczos Method for Approximating $f(A)v$. The following lemma follows directly from the Lanczos relations (2.18) and (2.19).

III:lem:lanczos

Lemma 2.5. *Let A be a Hermitian matrix and v a vector of unit norm.*
(a) If all eigenvalues of A are in the interval $[a,b]$, then so are those of T_m.
(b) For every polynomial p_{m-1} of degree at most $m-1$, it holds that

$$p_{m-1}(A)v = V_m\, p_{m-1}(T_m)\, e_1\,. \tag{2.21}$$

III:krylov-p

Proof. (a) If θ is an eigenvalue of T_m to the eigenvector w of unit norm, then $u = V_m w$ is again of unit norm, and by (2.19), $\theta = w^* T_m w = u^* A u$, which is in $[a,b]$.

(b) Clearly, $v = V_m e_1$. From (2.18) it follows by induction over $k = 1, 2, \ldots$ that

$$A^k V_m e_1 = V_m T_m^k e_1$$

as long as the lower left entry $e_m^T T_m^{k-1} e_1 = 0$. Since T_m^{k-1} is a matrix with $k-1$ subdiagonals, this holds for $k \le m-1$. $\square$

For any complex-valued function f defined on $[a,b]$, we have $f(A)$ given via the diagonalization $A = U\,\mathrm{diag}(\lambda_j)U^*$ as $f(A) = U\,\mathrm{diag}(f(\lambda_j))U^*$. Justified by (a) and motivated by (b), we can consider the approximation

$$f(A)v \approx V_m\, f(T_m)\, e_1\,. \tag{2.22}$$

III:krylov-f

For $f(x) = e^{-i\Delta t\, x}$ this is (2.20). Lemma 2.5 immediately implies the following useful approximation result.

III:thm:krylov-f

Theorem 2.6 (Optimality of the Lanczos Method). *Let f be a complex-valued function defined on an interval $[a,b]$ that contains the eigenvalues of the Hermitian matrix A, and let v be a vector of unit norm. Then, the error of the Lanczos approximation to $f(A)v$ is bounded by*

$$\|V_m f(T_m)e_1 - f(A)v\| \le 2 \inf_{p_{m-1}} \max_{x\in[a,b]} |p_{m-1}(x) - f(x)|\,,$$

where the infimum is taken over all polynomials of degree at most $m-1$.

Proof. By Lemma 2.5 (b), we have for every polynomial p_{m-1} of degree at most $m-1$,

$$V_m f(T_m)e_1 - f(A)v = V_m\big(f(T_m) - p_{m-1}(T_m)\big)e_1 - \big(f(A) - p_{m-1}(A)\big)v.$$

Diagonalization of A and T_m and Lemma 2.5 (a) show that each of the two terms to the right is bounded by $\max_{x\in[a,b]} |f(x) - p_{m-1}(x)|$. □

Error Bound of the Lanczos Method for the Matrix Exponential Operator. Combining Theorems 2.6 and 2.2, together with the linear transformation from the interval $[a,b]$ to $[-1,1]$, yields the following result.

thm:krylov-exp

Theorem 2.7 (Eventual Superlinear Error Decay). *Let A be a Hermitian matrix all of whose eigenvalues are in the interval $[a,b]$, and let v be a vector of unit Euclidean norm. Then, the error of the Lanczos method (2.20) is bounded by*

$$\|V_m e^{-i\Delta t\, T_m} e_1 - e^{-i\Delta t\, A} v\| \le 8\left(e^{1-(\omega/2m)^2}\,\frac{\omega}{2m}\right)^m \qquad \text{for} \quad m \ge \omega$$

with $\omega = \Delta t\,(b-a)/2$. □

III.3 Splitting and Composition Methods

The methods of the previous section have the attractive feature that they only require matrix-vector products with the discretized Hamiltonian A of (2.1). However, the maximum permitted step size is inversely proportional to the norm of A, which leads to a time step restriction to $\Delta t = \mathcal{O}(\Delta x^2)$, as we recall from (2.15). The splitting methods considered in this section can achieve good accuracy with no such restriction, provided that the wave function has sufficient spatial regularity.

III.3.1 Splitting Between Kinetic Energy and Potential

We consider the Schrödinger equation

$$i\dot\psi = H\psi \qquad \text{with} \quad H = T + V\,, \tag{3.1}$$

III:split-schrod

where T and V are the kinetic energy operator and the potential, respectively, or the corresponding discretized operators. We will assume no bound on the self-adjoint operator or matrix T. In our theoretical results we will assume bounds of the potential V, but the method to be described can work well under weaker assumptions. On the practical side, the basic assumption is that the equations

$$i\dot{\theta} = T\theta \qquad \text{and} \qquad i\dot{\phi} = V\phi$$

can both be solved more easily than the full equation (3.1). As we have seen in Chap. I, on the analytical level this is definitely the case in the non-discretized Schrödinger equation: the free Schrödinger equation (only T) is solved by Fourier transformation, and the equation with only the potential V is solved by multiplying the initial data with the scalar exponential $e^{-iV(x)}$ at every space point x. This situation transfers, in particular, to the Fourier collocation method of Section III.1.3, where solving the differential equations for the kinetic and potential parts in (1.32) or (1.33) is done trivially, using the exponentials of diagonal matrices and FFTs.

Strang Splitting. We consider time stepping from an approximation ψ^n at time t^n to the new approximation ψ^{n+1} at time $t^{n+1} = t^n + \Delta t$ by

$$\psi^{n+1} = e^{-i\frac{\Delta t}{2}V}\, e^{-i\Delta t\, T}\, e^{-i\frac{\Delta t}{2}V}\, \psi^n\,. \tag{3.2}$$

III:split-strang

This symmetric operator splitting was apparently first studied by Strang (1968) and independently by Marchuk (1968) in the context of dimensional splitting of advection equations. It was proposed, in conjunction with the Fourier method in space, for nonlinear Schrödinger equations by Hardin & Tappert (1973) and rediscovered for the linear Schrödinger equation, in the disguise of the Fresnel equation of laser optics, by Fleck, Morris & Feit (1976). The scheme was introduced to chemical physics by Feit, Fleck & Steiger (1982). In combination with Fourier collocation in space, the method is usually known as the *split-step Fourier method* in the chemical and physical literature.

Algorithm of the Split-Step Fourier method. In the notation of Sect. III.1.3, we recall the differential equation (1.33) for the vector $u = (u_j)$ of grid values $u_j(t) = \psi_K(x_j, t)$:

$$i\dot{u} = \mathcal{F}_K^{-1} D_K \mathcal{F}_K u + V_K u$$

with the diagonal matrices $D_K = \frac{1}{2\mu}\,\mathrm{diag}(k^2)$ and $V_K = \mathrm{diag}\big(V(x_j)\big)$, where k and j range from $-N/2$ to $N/2-1$. With method (3.2), a time step is computed in a way that alternates between pointwise operations and FFTs, overwriting the approximation at time t^n by that at time t^{n+1}:

1. replace $u_j := e^{-i\frac{\Delta t}{2}V(x_j)} u_j \quad (j = -N/2, \ldots, N/2-1)$
2. FFT: $u := \mathcal{F}_K u$
3. replace $u_k := e^{-i\Delta t k^2/(2\mu)} u_k \quad (k = -N/2, \ldots, N/2-1)$
4. inverse FFT: $u := \mathcal{F}_K^{-1} u$
5. replace $u_j := e^{-i\frac{\Delta t}{2}V(x_j)} u_j \quad (j = -N/2, \ldots, N/2-1).$

The exponentials in Substep 5 and Substep 1 of the next time step can be combined into a single exponential if the output at time t^{n+1} is not needed.

Unitarity, Symplecticity, Time-Reversibility. The Strang splitting has interesting structure-preserving properties. For self-adjoint T and V, the exponentials $e^{-i\Delta t\, T}$ and $e^{-i\frac{\Delta t}{2}V}$ are unitary (they preserve the norm) and symplectic (they preserve the canonical symplectic two-form $\omega(\xi,\eta) = -2\,\mathrm{Im}\,\langle\xi\,|\,\eta\rangle$, see Theorem II.1.2), and so does their composition. The time-step operator of the Strang splitting is thus both unitary and symplectic. We remark that neither holds for the Chebyshev method, whereas the Lanczos method is unitary, but symplectic only in the restriction to the Krylov subspace, which changes from one time step to the next. Moreover, the Strang splitting is time-reversible: a step of the method starting from ψ^{n+1} with negative step size $-\Delta t$ leads us back to the old ψ_n, or more formally, exchanging $n \leftrightarrow n+1$ and $\Delta t \leftrightarrow -\Delta t$ in the method gives the same method again. We note that neither the Chebyshev method nor the Lanczos method are time-reversible.

III.3.2 Error Bounds for the Strang Splitting

For bounded T and V, Taylor expansion of the exponentials readily shows

$$e^{-i\frac{\Delta t}{2}V}\, e^{-i\Delta t\, T}\, e^{-i\frac{\Delta t}{2}V} = e^{-i\Delta t(T+V)} + \mathcal{O}\big(\Delta t^3(\|T\| + \|V\|)^3\big)\,.$$

However, such an error bound is of no use when T or V are of large norm. Since $\|T\| \sim (\Delta x)^{-2}$ (as in (2.15)), this error bound would indicate a small error only for $\Delta t \ll \Delta x^2$, whereas numerical experiments clearly indicate that the error of the Strang splitting for initial data of moderately bounded energy is bounded independently of Δx for a given Δt. For problems with smooth potential and smooth initial data the error is numerically observed to be $\mathcal{O}(\Delta t^3)$ uniformly in Δx after one step of the method, and $\mathcal{O}(t^n \Delta t^2)$ at time t^n after n steps, uniformly in n and Δx.

In the following we present an error analysis from Jahnke & Lubich (2000), which explains this favourable behaviour of the splitting method. Here we assume that T and V are self-adjoint operators on a Hilbert space $\mathcal{H}$, and T is positive semi-definite. We require no bound for T, but we assume a (moderate) bound of V:

$$\|V\psi\| \le B\|\psi\| \qquad \forall \psi \in \mathcal{H}\,. \tag{3.3}$$

III:split-V-bound

We introduce the norms

$$\begin{aligned} \|\varphi\|_1 &= \langle\varphi\,|\,T + I\,|\,\varphi\rangle^{1/2} \\ \|\varphi\|_2 &= \langle\varphi\,|\,(T+I)^2\,|\,\varphi\rangle^{1/2} \end{aligned} \tag{3.4}$$

III:split-norms

which are the usual Sobolev norms in the case of $T = -\Delta$, and can be viewed as discrete Sobolev norms in the spatially discrete case.

Our main assumptions concern the commutator $[T,V] = TV - VT$ and the repeated commutator $[T,[T,V]] = T^2V - 2TVT + VT^2$. We assume that there are constants c_1 and c_2 such that the commutator bounds

$$\| [T,V]\varphi\| \;\leq\; c_1 \,\|\varphi\|_1 \tag{3.5}$$ III:split-comm1

$$\| [T,[T,V]]\varphi\| \;\leq\; c_2 \,\|\varphi\|_2 \tag{3.6}$$ III:split-comm2

are satisfied for all φ in a dense domain of $\mathcal{H}$. In the spatially continuous case with $T = -\Delta$ and a potential $V(x)$ that is bounded together with its first- to fourth-order derivatives, we see from the identities

$$\begin{aligned} [\Delta,V]\varphi &= \Delta V\,\varphi + 2\nabla V\cdot\nabla\varphi \\ [\Delta,[\Delta,V]]\varphi &= \Delta^2 V\,\varphi + 4\nabla\Delta V\cdot\nabla\varphi + 4\sum_{j,l}\partial_j\partial_l V\,\partial_j\partial_l\varphi \end{aligned}$$

that the commutator bounds (3.5)–(3.6) are indeed valid. For spatial discretization by the Fourier method, it is shown by Jahnke & Lubich (2000) that these commutator bounds hold with constants c_1 and c_2 that are independent of the discretization parameter. We then have the following second-order error bound.

thm:split-error **Theorem 3.1 (Error Bound for the Strang Splitting).** *Under the above conditions, the error of the splitting method (3.2) at $t = t^n$ is bounded by*

$$\|\psi^n - \psi(t)\| \leq C\,\Delta t^2\; t \max_{0\leq\tau\leq t} \|\psi(\tau)\|_2\,,$$

where C depends only on the bound B of (3.3) and on c_1, c_2 of (3.5)–(3.6).

It is a noteworthy fact that the time discretization error of the splitting method depends on the *spatial* regularity of the wave function, not on its temporal regularity. The proof is done in the usual way by studying the local error of the method (that is, the error after one step) and the error propagation. For the local error we have the following bounds.

lem:split-local **Lemma 3.2 (Local Error).** *(a) Under conditions (3.3) and (3.5),*

$$\|e^{-i\frac{\Delta t}{2}V}e^{-i\Delta t T}e^{-i\frac{\Delta t}{2}V}\varphi - e^{-i\Delta t(T+V)}\varphi\| \leq C_1\,\Delta t^2\,\|\varphi\|_1\,, \tag{3.7}$$ le1

where C_1 depends only on c_1 and B.

(b) Under conditions (3.3) and (3.5)–(3.6),

$$\|e^{-i\frac{\Delta t}{2}V}e^{-i\Delta t T}e^{-i\frac{\Delta t}{2}V}\varphi - e^{-i\Delta t(T+V)}\varphi\| \leq C_2\,\Delta t^3\,\|\varphi\|_2\,, \tag{3.8}$$ le2

where C_2 depends only on c_1, c_2 and B.

The local error bound (3.8) together with the telescoping formula

$$\psi^n - \psi(t^n) = S^n\psi^0 - E^n\psi^0 = \sum_{j=0}^{n-1} S^{n-j-1}(S-E)\,E^j\psi^0\,, \tag{3.9}$$ tele

with $S = e^{-i\frac{\Delta t}{2}V}e^{-i\Delta t T}e^{-i\frac{\Delta t}{2}V}$ and $E = e^{-i\Delta t(T+V)}$, immediately yields the error bound of Theorem 3.1. It thus remains to prove the lemma. The basic idea of the following proof is the reduction of the local error to quadrature errors.

Proof. (a) We start from the variation-of-constants formula

$$e^{-i\Delta t(T+V)}\varphi = e^{-i\Delta t T}\varphi - i\int_0^{\Delta t} e^{-isT} V e^{-i(\Delta t-s)(T+V)}\varphi\, ds\,.$$

Expressing the last term under the integral once more by the same formula yields

$$e^{-i\Delta t(T+V)}\varphi = e^{-i\Delta t T}\varphi - i\int_0^{\Delta t} e^{-isT} V e^{-i(\Delta t-s)T}\varphi\, ds + R_1\varphi\,,$$

where the remainder

$$R_1 = -\int_0^{\Delta t} e^{sT} V \int_0^{\Delta t-s} e^{-i\sigma T} V e^{-i(\Delta t-s-\sigma)(T+V)}\, d\sigma\, ds$$

is bounded in the operator norm by $\|R_1\| \le \frac{1}{2}\Delta t^2 B^2$. On the other hand, using the exponential series for $e^{-i\frac{\Delta t}{2}V}$ leads to

$$e^{-i\frac{\Delta t}{2}V} e^{-i\Delta t T} e^{-i\frac{\Delta t}{2}V}\varphi = e^{-i\Delta t T}\varphi - \frac{i}{2}\Delta t\big(V e^{-i\Delta t T} + e^{-i\Delta t T} V\big)\varphi + R_2\varphi\,,$$

where $\|R_2\| \le \frac{1}{2}\Delta t^2 B^2$. Consequently, the error is of the form

$$e^{-i\frac{\Delta t}{2}V} e^{-i\Delta t T} e^{-i\frac{\Delta t}{2}V}\varphi - e^{-i\Delta t(T+V)}\varphi = d + r\,, \tag{3.10}$$ e

where $r = R_2\varphi - R_1\varphi$ and, with $f(s) = -i\, e^{-isT} V e^{-i(\Delta t-s)T}\varphi$,

$$\begin{aligned} d &= \frac{1}{2}\,\Delta t\,\big(f(0)+f(\Delta t)\big) - \int_0^{\Delta t} f(s)\, ds \qquad (3.11)\ \boxed{\text{d}} \\ &= -\Delta t^2 \int_0^1 (\tfrac{1}{2}-\theta)\, f'(\theta\Delta t)\, d\theta = \frac{1}{2}\,\Delta t^3 \int_0^1 \theta(1-\theta) f''(\theta\Delta t)\, d\theta \end{aligned}$$

is the error of the trapezoidal rule, written in first- and second-order Peano form. Since $f'(s) = -e^{-isT}[T,V]e^{-i(\Delta t-s)T}\varphi$, condition (3.5) yields the error bound (3.7).

(b) For the error bound (3.8), we use $f''(s) = i\, e^{-isT}[T,[T,V]]e^{-i(\Delta t-s)T}\varphi$ and condition (3.6) to bound

$$\|d\| \le \frac{1}{12}\, c_2\, \Delta t^3\, \|\varphi\|_2\,. \tag{3.12}$$ d2

It remains to study $r = R_2 v - R_1 v$. We have

$$R_1 = -\int_0^{\Delta t} e^{-isT} V \int_0^{\Delta t-s} e^{-i\sigma T} V e^{-i(\Delta t-s-\sigma)T}\, d\sigma\, ds + \widetilde{R}_1$$

with $\|\widetilde{R}_1\| \le C\Delta t^3 B^3$, and

$$R_2 = -\frac{1}{8}\,\Delta t^2\,\big(V^2 e^{-i\Delta t T} + 2V e^{-i\Delta t T} V + e^{-i\Delta t T} V^2\big) + \widetilde{R}_2$$

with $\|\widetilde{R}_2\| \le C\Delta t^3 B^3$. We thus obtain

$$r = \widetilde{d} + \widetilde{r}\,, \tag{3.13}$$ r

where $\widetilde{r} = \widetilde{R}_2\varphi - \widetilde{R}_1\varphi$ is bounded by $\|\widetilde{r}\| \le C\Delta t^3\, B^3\, \|\varphi\|$ and, with $g(s,\sigma) = -e^{-isT}Ve^{-i\sigma T}Ve^{-i(\Delta t-s-\sigma)T}\varphi$,

$$\widetilde{d} = \frac{1}{8}\,\Delta t^2\,\Big(g(0,0) + 2g(0,\Delta t) + g(\Delta t,0)\Big) - \int_0^{\Delta t}\int_0^{\Delta t-s} g(s,\sigma)\,d\sigma\,ds$$

is the error of a quadrature formula that integrates constant functions exactly. Hence,

$$\|\widetilde{d}\| \le \widetilde{c}\,\Delta t^3\,\left(\max\left\|\frac{\partial g}{\partial s}\right\| + \max\left\|\frac{\partial g}{\partial \sigma}\right\|\right)\,,$$

where the maxima are taken over the triangle $0 \le s \le \Delta t$, $0 \le \sigma \le \Delta t - s$. Since

$$\frac{\partial g}{\partial s}(s,\sigma) = i\,e^{-isT}[T,V]e^{-i\sigma T}Ve^{-i(\Delta t-s-\sigma)T}\varphi + i\,e^{-isT}Ve^{-i\sigma T}[T,V]e^{-i(\Delta t-s-\sigma)T}\varphi\,,$$

we obtain, using (3.5),

$$\left\|\frac{\partial g}{\partial s}\right\| \le c_1\,(c_1+B)\,\|\varphi\|_1 + B\,c_1\,\|\varphi\|_1\,.$$

Similarly, $\|\partial g/\partial\sigma\| \le B\,c_1\,\|\varphi\|_1$, so that finally

$$\|\widetilde{d}\| \le C\Delta t^3\,\|\varphi\|_1\,.$$

Together with the above bounds for $\widetilde{r}$ and d this yields the error bound (3.8). □

III.3.3 Higher-Order Compositions

III:higher-order

The Strang splitting $S(\Delta t) = e^{-i\frac{\Delta t}{2}V}e^{-i\Delta t T}e^{-i\frac{\Delta t}{2}V}$ yields a second-order method. Higher-order methods can be obtained by a suitable composition of steps of different size of the basic method:

$$\psi^{n+1} = S(\gamma_s\Delta t)\ldots S(\gamma_1\Delta t)\psi^n$$

with symmetrically arranged coefficients $\gamma_j = \gamma_{s+1-j}$ determined such that

$$S(\gamma_s\Delta t)\ldots S(\gamma_1\Delta t) = e^{-i\Delta t(T+V)} + \mathcal{O}\big(\Delta t^{p+1}(\|T\|+\|V\|)^{p+1}\big)$$

with an order $p > 2$. Composition methods of this or similar type have been devised by Suzuki (1990) and Yoshida (1990), and improved methods have since been constructed, e.g., by McLachlan (1995), Kahan & Li (1997), Blanes & Moan (2002), Sofroniou & Spaletta (2005). We refer to Hairer, Lubich & Wanner (2006), Sect. V.3, and McLachlan

& Quispel (2002) for reviews of composition methods, for their order theory, for their coefficients, and for further references. For example, an excellent method of order $p = 8$ with $s = 17$ by Kahan & Li (1997) has the coefficients

$$\begin{aligned}
\gamma_1 = \gamma_{17} &= 0.13020248308889008087881763\\
\gamma_2 = \gamma_{16} &= 0.56116298177510838456196441\\
\gamma_3 = \gamma_{15} &= -0.38947496264484728640807860\\
\gamma_4 = \gamma_{14} &= 0.15884190655515560089621075\\
\gamma_5 = \gamma_{13} &= -0.39590389413323757733623154\\
\gamma_6 = \gamma_{12} &= 0.18453964097831570709183254\\
\gamma_7 = \gamma_{11} &= 0.25837438768632204729397911\\
\gamma_8 = \gamma_{10} &= 0.29501172360931029887096624\\
\gamma_9 &= -0.60550853383003451169892108
\end{aligned} \tag{3.14}$$

eq:comp_order8a

As with the basic Strang splitting method, the presence of powers of $\|T\|$ in the error bound would seem to make a step-size restriction $\Delta t \ll \Delta x^2$ necessary, but indeed this is not the case. Thalhammer (2008) proves high-order error bounds for such methods that require no bound of T. By a formidable extension of the approach in the proof of Theorem 3.1, using p-fold repeated commutator bounds and achieving a reduction to quadrature errors, it is shown that in the spatially continuous case with $T = -\Delta$ and a smooth bounded potential, there is a pth-order error bound at $t = t^n$

$$\|\psi^n - \psi(t)\| \le C\,\Delta t^p\, t \max_{0\le\tau\le t} \|\psi(\tau)\|_p$$

with the pth-order Sobolev norm. It is to be expected that in the spatially discretized case, the required commutator bounds hold uniformly in Δx so that the error bound becomes uniform in the spatial discretization parameter.

III.4 Integrators for Time-Dependent Potentials

III.4.1 Magnus Methods

III.4.2 Adiabatic Integrators