Various lecture notes for 18311.

R. R. Rosales

April 28, 2011 version 01.

Abstract

Notes, both complete and/or incomplete, for MIT's 18.311 (Principles of Applied Mathematics).
These notes will be updated from time to time. Check the date and version.

Contents

1 Convergence of numerical schemes.

In this section we introduce some theory dealing with the question of the convergence of numerical schemes for partial differential equations. For simplicity we will consider only the case of an **homogeneous, one step in time, linear scheme for an initial value problem (IVP) for an homogeneous linear system of first order in time partial differential equations (pde) in 1-D.**

1.1 The initial value problem.

The IVP to be solved has the form

$$\vec{u}_t = \mathcal{L}\,\vec{u} \quad \text{for} \;\; x_L < x < x_R \quad \text{and} \;\; t > 0, \tag{1.1.1}$$

with $\vec{u}(x,\,0) = \vec{u}_0(x)$, where $\vec{u}(x,\,t)$ is vector[1] valued, $\mathcal{L}$ is a linear differential operator, and appropriate homogeneous[2] boundary conditions (BC) apply.

Remark 1.1.1 *We will assume that (1.1.1) is a well posed problem, with a solution that is as smooth as needed (this is specified later).*

Example 1.1.1 Linear scalar equation. $u_t + a\,u_x = b\,u$, *where* $u = u(x,\,t)$ *is scalar valued,* $(a,\,b)$ *are given functions of* $(x,\,t)$*, and periodic BC apply:* $u(x_L,\,t) = u(x_R,\,t)$.

Example 1.1.2 Heat equation. $u_t = (\nu\,u_x)_x$, *where* $u = u(x,\,t)$ *is scalar valued,* $\nu > 0$ *is a given function of* $(x,\,t)$*, and either of the following BC apply (this list of BC is not exhaustive)*

(i) $u(x_L,\,t) = u(x_R,\,t)$ *and* $u_x(x_L,\,t) = u_x(x_R,\,t)$ *(periodic).*

(ii) $u(x_L,\,t) = u(x_R,\,t) = 0$.

(iii) $u_x(x_L,\,t) = u_x(x_R,\,t) = 0$.

(iv) $u(x_L,\,t) = u_x(x_R,\,t) = 0$.

Example 1.1.3 Wave equation: $(u_1)_t = u_2$ *and* $(u_2)_t = \left(c^2\,(u_1)_x\right)_x$, *where* $\vec{u} = (u_1,\,u_2)$, c *is a given function of* $(x,\,t)$*, and the same BC as in example 1.1.2 apply.*

Example 1.1.4 Klein-Gordon equation. $(u_1)_t = u_2$ *and* $(u_2)_t = \left(c^2\,(u_1)_x\right)_x - m^2\,u_1$, *where* $\vec{u} = (u_1,\,u_2)$, $(c,\,m)$ *are given functions of* $(x,\,t)$*, and the same BC as in example 1.1.2 apply.*

Example 1.1.5 Korteweg-de Vries equation. $u_t + a\,u_x + b\,u_{xxx} = 0$, *where* $u = u(x,\,t)$ *is scalar valued,* $(a,\,b)$ *are given functions of* $(x,\,t)$*, and periodic BC apply:* $u(x_L,\,t) = u(x_R,\,t)$, $u_x(x_L,\,t) = u_x(x_R,\,t)$, *and* $u_{xx}(x_L,\,t) = u_{xx}(x_R,\,t)$.

[1] $\vec{u} \in \boldsymbol{R^d}$**, is a** $\boldsymbol{d}$**-column vector, for some** $\boldsymbol{d = 1,\,2,\,3,\,\ldots}$**.**

[2] Homogeneous BC means BC that yield: If $\vec{u}_j$ solves (1.1.1) for the initial values $\vec{u}_0 = \vec{U}_j$ $(j = 1$ or $j = 2)$, then $\vec{u} = \alpha_1\,\vec{u}_1 + \alpha_2\,\vec{u}_2$ solves (1.1.1) for the initial value $\vec{u}_0 = \alpha_1\,\vec{U}_1 + \alpha_2\,\vec{U}_2$, where α_1 and α_2 are arbitrary constants.

1.2 The numerical scheme.

Assume an appropriate[3] grid in space-time, with constant grid separations $\Delta x > 0$ and $\Delta t > 0$. For example

$$\Delta x = \frac{1}{N+1}\,(x_R - x_L)\,, \quad x_n = x_L + n\,\Delta x \quad \text{for } \ 1 \leq n \leq N, \quad \text{and} \quad t_j = j\,\Delta t \quad \text{for } \ j \geq 0, \qquad (1.2.1)$$

where $N > 1$ is any sufficiently "large" integer.[4] We assume that the numerical scheme to solve the (1.1.1) IVP has the form

$$\mathbf{u}^{j+1} = \mathcal{S}_j\,\mathbf{u}^j \quad \text{for } \ j \geq 0, \qquad (1.2.2)$$

where $\mathbf{u}^j \in R^{d\times N}$ is a $d \times N$ matrix for each $j > 0$, and $\{\mathcal{S}_j\}_{j\geq 0}$ is a sequence of linear operators in $R^{d\times N}$ — which could be represented as $d\,N \times d\,N$ matrices if needed. The operators $\mathcal{S}_j$ generally depend on: Δt and Δx, possibly some numerical parameters (e.g.: artificial viscosity), and the details of the pde to be solved. Furthermore, each of the columns of $\mathbf{u}^j$ is a d-vector, which we denote by $\vec{u}_n^{\,j}$ — for $1 \leq n \leq N$. These vectors are interpreted as the approximations to the values of the solution at the grid points. Namely

$$\vec{u}_n^{\,j} \approx \vec{u}(x_n, t_j). \qquad (1.2.3)$$

In particular

$$\vec{u}_n^{\,0} = \vec{u}_0(x_n) \qquad (1.2.4)$$

should be used to initialize the numerical scheme.

Note that:

1.2a *The interpretation in (1.2.3–1.2.3) is not unique. For example, in many schemes $\vec{u}_n^{\,j}$ is taken as the average value of the solution over the n^{th} cell: $|x - x_n| \leq \frac{1}{2}\,\Delta x$.*

1.2b *Schemes with meshes where $t_{j+1} - t_j = \Delta t$ is not a constant are frequently used. Similarly, $x_{n+1} - x_n = \Delta x$ need not be a constant.*

1.2c *In (1.2.1) the end points x_L and x_R are not included in the numerical grid. This would be appropriate if the solution is prescribed there — e.g.: $\vec{u}(x_L,\,t) = \vec{u}(x_R,\,t) = 0$. For periodic BC, an appropriate choice is to include one of the end points but not the other, as in: $x_n = x_L + (n-1)\,\Delta x$ for $1 \leq n \leq N$, with $\Delta x = (x_R - x_L)/N$. Many other choices are possible.*

Example 1.2.1 *Consider the heat equation, as in example 1.1.2 – with the BC in (ii). Then, with the choice of grid in (1.2.1), a scheme of the form in (1.2.2) is given by*

$$u_n^{j+1} = u_n^j + \frac{\Delta t}{(\Delta x)^2}\,\left(\nu_{n+\frac{1}{2}}^j\,\left(u_{n+1}^j - u_n^j\right) - \nu_{n-\frac{1}{2}}^j\,\left(u_n^j - u_{n-1}^j\right)\right), \qquad (1.2.5)$$

where $\nu_{n\pm\frac{1}{2}}^j = \nu\left(x_n \pm \frac{1}{2}\,\Delta x,\,t_j\right)$, and $u_0^j = u_{N+1}^j = 0$ is used when evaluating the right hand side of equation (1.2.5) for $n = 1$ and $n = N$. This scheme makes sense for any $N \geq 1$.

[3] See item 1.2c below.

[4] The numerical scheme should be defined for any N large enough — where large enough is usually not very large — see example 1.2.1. We are, however, interested in the limit $N \to \infty$ here.

1.3 Consistency and stability.

Let $\|\cdot\|_N$ be a norm in $R^{d\times N}$, where the $\mathbf{u}^j$ belong — see equation (1.2.2). We assume the following:

$$\left.\begin{array}{l}\text{Let } \vec{u}=\vec{u}(x) \text{ be a suficiently nice } d\text{-vector valued function. Define } \mathbf{u}\in R^{d\times N} \text{ by}\\ \vec{u}_n=\vec{u}(x_n)\text{, where } \vec{u}_n \text{ is the } n^{th} \text{ column of } \mathbf{u}\text{, for } 1\leq n\leq N\text{. Then } \|\mathbf{u}\|_N\to\|\vec{u}\|_*\\ \text{as } N\to\infty\text{, where } \|\cdot\|_* \text{ is a norm defined for } d\text{-vector valued functions.}\end{array}\right\} \qquad (1.3.1)$$

Example 1.3.1 *For $d=1$, let $\|\mathbf{u}\|_N=\sqrt{\sum_1^N |u_n|^2\,\Delta x}$. Then $\|u\|_*=\sqrt{\int_{x_L}^{x_R}|u(x)|^2\,dx}$, and in (1.3.1) "sufficiently nice" means continuous.*

Example 1.3.2 *For $d=1$, let $\|\mathbf{u}\|_N=\sqrt{\sum_1^N |u_n|^2\,\Delta x+\sum_2^{N-1}|u_{n+1}-u_n|^2\,\Delta x}$. Then $\|u\|_*=\sqrt{\int_{x_L}^{x_R}|u(x)|^2\,dx+\int_{x_L}^{x_R}|u'(x)|^2\,dx}$, and in (1.3.1) "sufficiently nice" means C^1 — i.e.: $\vec{u}$ has a continuous derivative.*[5]

Remark 1.3.1 *Recall that a norm is a real valued function defined on a vector space $\mathcal{V}$ such that, for any $v\in\mathcal{V}$ and $w\in\mathcal{V}$, and scalars a and b, the following applies: (i) $\|v\|\geq 0$, (ii) $\|v\|=0$ if and only if $v=0$, and (iii) $\|v+w\|\leq\|v\|+\|w\|$.*

Definition 1.3.1 *The numerical scheme in § 1.2 is* **consistent** *if and only if the following applies:*

Let $\vec{u}=\vec{U}(x,\,t)$ be the solution to the IVP (1.3.1) for some arbitrary initial condition $\vec{u}_0$. Assume that $\vec{U}$ is is sufficiently smooth and define $\mathbf{U}^j\in R^{d\times N}$ by $\vec{U}_n^j=\vec{U}(x_n,\,t_j)$, where $1\leq n\leq N$, $j\geq 0$, and $\vec{U}_n^j$ is the n^{th} column of $\mathbf{U}^j$. Then

$$\|\mathbf{U}^{j+1}-\mathcal{S}_j\,\mathbf{U}^j\|_N\leq f_c(t_{j+1})\,\Delta t\,\left((\Delta t)^p+(\Delta x)^q\right), \qquad (1.3.2)$$

where $p>0$ is the order of the method in time, $q>0$ is the order of the method in space, and $0<f_c(t)<\infty$ is some grid independent bounded function — determined by the solution $\vec{U}$ and its partial derivatives up to some order,^{a} as well as the coefficients^b of the equation in the IVP (1.1.1).

(a) This is why $\vec{U}$ needs to be sufficiently smooth.

(b) These coefficients must also be sufficiently smooth.

Example 1.3.3 *Consider the numerical scheme in example 1.2.1. In this case (1.3.2) applies with $p=1$, $q=2$, and*

$$f_c(t)=\max_N\left(\|1\|_N\right)\,\max(M_1,M_2), \qquad (1.3.3)$$

where

(i) $\|1\|_N$ indicates the norm of the vector all whose entries are one — since $\|1\|_N\to\|1\|_*$ as $N\to\infty$, $\{\|1\|_N\}_N$ is a bounded sequence with a maximum.

[5] Actually, less is needed — e.g.: an integrable bounded derivative will do (dominated convergence theorem).

(ii) $M_1 = M_1(t)$ *is the maximum of* $\frac{1}{2}\,|U_{tt}(x,\, s)|$*, for* $x_L \le x \le x_R$*, and* $0 \le s \le t$*.*

(iii) $M_2 = M_2(t)$ *is the maximum of* $\frac{1}{24}\,|G_{hhhh}(x,\, h, s)|$*, for* $x_L + h \le x \le x_R - h$*, and* $0 \le s \le t$ *— where* G *is defined in (1.3.5).*

Proof. *Using in (1.2.5) Taylor expansions with remainder it follows that*

$$\begin{aligned} &U_n^{j+1} - U_n^j - \frac{\Delta t}{(\Delta x)^2}\left(\nu_{n+\frac{1}{2}}^j\left(U_{n+1}^j - U_n^j\right) - \nu_{n-\frac{1}{2}}^j\left(U_n^j - U_{n-1}^j\right)\right) \\ &= \frac{1}{2}\,U_{tt}(x_n, \tau_n^j)\,(\Delta t)^2 - \frac{1}{24}\,G_{hhhh}(x_n, h_n^j, t_j)\,(\Delta x)^2\,\Delta t, \end{aligned} \tag{1.3.4}$$

for some $t_j \le \tau_n^j \le t_{j+1}$ *and* $0 \le h_n^j \le \Delta x$*, where*

$$G(x,\, h,\, t) = \nu(x + \frac{1}{2}\,h)\ (U(x+h,t) - U(x,t)) - \nu(x - \frac{1}{2}\,h)\ (U(x,t) - U(x-h,t))\,. \tag{1.3.5}$$

Hence (1.3.3) follows. ♣

Remark 1.3.2 *Methods exist for which (1.3.2) does not strictly apply. For example, one may have*

$$\|\mathbf{U}^{j+1} - \mathcal{S}_j\,\mathbf{U}^j\|_N \le f_c(t_{j+1})\ \left((\Delta t)^{p+1} + (\Delta x)^{q'}\right). \tag{1.3.6}$$

However, in a numerical method one is interested in the situations where both Δt *and* Δx *are small,*[6] *and generally* Δt *and* Δx *are related to each other — e.g.* $\Delta t = \text{constant}\,\Delta x$*, in which case (1.3.2) and (1.3.6) are equivalent.*

Definition 1.3.2 *The numerical scheme in § 1.2 is* **stable** *if and only if*

$$\|\mathbf{u}^j\|_N \le f_s(t_j)\,\|\mathbf{u}^k\|_N, \quad \textit{for any} \quad 0 \le k \le j, \tag{1.3.7}$$

where $0 < f_s(t) < \infty$ *is some grid (and solution) independent*^{a} *bounded function. Note that, for equation (1.3.7) to apply, restrictions might be needed on* Δt *and* Δx *— such as:* $\Delta t \le \text{constant}\,\Delta x$ *or* $\Delta t \le \text{constant}\,(\Delta x)^2$*. These restrictions must allow* Δt *and* Δx *to vanish simultaneously.*

(a) Of course, f_s will depend on the coefficients of the equation in the IVP (1.1.1).

1.4 Lax convergence theorem.

Theorem 1.4.1 *If the scheme in § 1.2 is consistent and stable, then it converges — in any fixed time interval* $0 \le t \le T$ *— as* $\Delta t \to 0$ *and* $\Delta x \to 0$ *(provided that any restrictions on* Δt *and* Δx *required by (1.3.7) apply). By converges we mean that*

$$\|\mathbf{u}^j - \mathbf{U}^j\|_N \to 0, \ \ \textit{for} \ \ 0 \le t_j \le T, \quad \textit{as} \quad \Delta t + \Delta x \to 0, \tag{1.4.1}$$

where $\mathbf{U}^j$ *is as in definition 1.3.1,* $\mathbf{u}^j$ *is the numerical solution (1.2.2) — initialized as in (1.2.4), and the convergence is uniform in* $0 \le t_j \le T$*.*

[6] Formally, $\Delta t \to 0$ and $\Delta x \to 0$.

Proof: we have $\mathbf{u}^{j+1} - \mathbf{U}^{j+1} = (\mathbf{u}^{j+1} - \mathcal{S}_j\,\mathbf{u}^j) + (\mathcal{S}_j\,\mathbf{u}^j - \mathcal{S}_j\,\mathbf{U}^j) + (\mathcal{S}_j\,\mathbf{U}^j - \mathbf{U}^{j+1})$. Hence, since $\mathbf{u}^{j+1} - \mathcal{S}_j\,\mathbf{u}^j = 0$, we can write $\mathbf{u}^{j+1} - \mathbf{U}^{j+1} = \mathcal{S}_j\,(\mathbf{u}^j - \mathbf{U}^j) + (\mathcal{S}_j\,\mathbf{U}^j - \mathbf{U}^{j+1})$. Recursive application of this then yields

$$\begin{aligned}\mathbf{u}^j - \mathbf{U}^j \;=\; & \underbrace{\mathcal{S}_{j-1}\,\mathcal{S}_{j-2}\,\ldots\,\mathcal{S}_1\,\mathcal{S}_0\,(\mathbf{u}^0 - \mathbf{U}^0)}_{A} + (\mathcal{S}_{j-1}\,\mathbf{U}^{j-1} - \mathbf{U}^j) + S_{j-1}\,(\mathcal{S}_{j-2}\,\mathbf{U}^{j-2} - \mathbf{U}^{j-1}) + \\ & S_{j-1}\,S_{j-2}\,(\mathcal{S}_{j-3}\,\mathbf{U}^{j-3} - \mathbf{U}^{j-2}) + \ldots + \mathcal{S}_{j-1}\,\mathcal{S}_{j-2}\,\ldots\,\mathcal{S}_1\,(\mathcal{S}_0\,\mathbf{U}^0 - \mathbf{U}^1),\end{aligned} \qquad (1.4.2)$$

where $A = 0$ — since $\mathbf{u}^0 = \mathbf{U}^0$. Let $0 < K < \infty$ be a bound on f_c and f_s for $0 \le t \le T$. Then, from stability and consistency

$$\|\mathcal{S}_{j-1}\,\mathcal{S}_{j-2}\,\ldots\,\mathcal{S}_{\ell+1}\,(\mathcal{S}_\ell\,\mathbf{U}^\ell - \mathbf{U}^{\ell+1})\|_N \le K\,\|\mathcal{S}_\ell\,\mathbf{U}^\ell - \mathbf{U}^{\ell+1}\|_N \le K^2\,\left((\Delta t)^p + (\Delta x)^q\right)\,\Delta t, \qquad (1.4.3)$$

for any $0 \le \ell < j$. Using this in (1.4.2) then yields

$$\|\mathbf{u}^j - \mathbf{U}^j\|_N \le K^2\,\left((\Delta t)^p + (\Delta x)^q\right)\,t_j \le K^2\,\left((\Delta t)^p + (\Delta x)^q\right)\,T, \qquad (1.4.4)$$

from which (1.4.1) follows. ♣

1.5 Example: von Neumann stability.

Consider now the situation when the equation in (1.1.1) has constant coefficients, and periodic BC apply. This is the case where stability can be ascertained using a von Neumann analysis, as we show next. For simplicity we will assume a scalar equation (i.e. $d = 1$), and a normalized period $= 2\,\pi$, with $x_L = 0$ and $x_R = 2\,\pi$. We then use the space grid $x_n = n\,\Delta x$, $1 \le n \le N$, with $\Delta x = 2\,\pi/N$. The initial data for the scheme $\mathbf{u}^0 = \{u_n^0\}_{n=1}^N$ can be written in the form (see § 2)

$$u_n^0 = \sum_{\ell=1}^N a_\ell\,e^{i\,\ell\,x_n} = \sum_{\ell=1}^N a_\ell\,e^{i\,k_\ell\,n} = \sum_{\ell=1}^N a_\ell\,e^{i\,2\,\pi\,\ell\,n/N} = \sum_{\ell=1}^N a_\ell\,w^{\ell\,n} \quad \text{for } \; 1 \le n \le N, \qquad (1.5.1)$$

where $w = e^{i\,2\,\pi/N}$ is the N^{th} fundamental root of unity, $k_\ell = \ell\,\Delta x = 2\,\pi\,\ell/N$, and the $\{a_\ell\}_\ell^N$ are some (complex) constants. Then a von Neumann analysis shows that the solution $\mathbf{u}^j = \{u_n^j\}_{n=1}^N$ to the numerical scheme has the form

$$u_n^j = \sum_{\ell=1}^N a_\ell\,(G_\ell)^j\,e^{i\,\ell\,x_n} \quad \text{for } \; 1 \le n \le N, \;\text{ and } \; j \ge 0, \qquad (1.5.2)$$

where the $\{G_\ell\}_\ell^N$ are constants that depend on Δt, Δx, the coefficients of the equation, and any numerical parameters.

Remark 1.5.1 *Finite differences approximations to constant coefficients IVP with periodic boundary conditions generally yield numerical schemes for which a von Neumann analysis works. Namely: in (1.2.2) $\mathcal{S}_j = \mathcal{S}$ is independent of j, and S applied to a $\mathbf{u}^j$ whose components are proportional to exponentials of the form $e^{i\,k\,n}$, for some k, yields an answer of the same type.*

We point out that it is possible to produce numerical schemes (not necessarily using finite differences) for constant coefficients IVP with periodic boundary conditions, for which a von Neumann analysis does not work. Here we assume that this is not the case.

Apply now the norm introduced in example 1.3.1 to the expression in equation (1.5.2). Using the fact that $\boldsymbol{\sum_{n=1}^{N} w^{n\,(\ell_1-\ell_2)} = N\,\delta_{\ell_1\,\ell_2}}$ for $1 \leq \ell_1,\, \ell_2 \leq N$, we obtain

$$\|\mathbf{u}^j\|_N = \sqrt{2\,\pi \sum_{\ell=1}^{N} |a_\ell|^2\,(|G_\ell|^2)^j} \quad \text{for } \; j \geq 0. \tag{1.5.3}$$

Define now $\boldsymbol{G = \max_{1\leq\ell\leq N} |G_\ell|}$. Then

$$\|\mathbf{u}^j\|_N \leq G^{j-k}\,\|\mathbf{u}^k\|_N \quad \text{for } \; 0 \leq k \leq j. \tag{1.5.4}$$

Comparing this with (1.3.7), we see that stability can be ascertained by studying the behavior of G^j as $j \to \infty$ with $t_j = j\,\Delta t$ bounded. In particular:

$$\textbf{If } \boldsymbol{G \leq 1}\textbf{, the scheme is stable.} \tag{1.5.5}$$

Problem 1.5.1 **Complete the details for the von Neumann stability example.**
Show that[7] $\sum_{n=1}^{N} w^{n\,(\ell_1-\ell_2)} = N\,\delta_{\ell_1\,\ell_2}$ *for* $1 \leq \ell_1,\, \ell_2 \leq N$*, where* $N > 1$ *is an integer, and* $w = e^{i\,2\,\pi/N}$ *is the* N^{th} *fundamental root of unit. Then use this to derive (1.5.3).*
Hint: $w^M = 1$ if and only if M is a multiple of N.

2 DFT, FFT, and Fourier series.

This section deals with the Discrete Fourier Transform (DFT), it fast implementation using the Fast Fourier Transform (FFT), and the relationship of the DFT with Fourier series for periodic functions.

2.1 Introduction and motivation.

In von Neumann stability analysis — see § 1.5, we conclude that a numerical scheme for a situation with periodic boundary conditions[8] is stable if and only if the solutions to the scheme of the form

$$u_N^j = G^j\,e^{i\,k\,n} \tag{2.1.1}$$

[7] The notation $\delta_{\ell\,j}$ is used for the Kronecker delta: $\delta_{\ell\,j} = 0$ if $\ell \neq j$, and $\delta_{j\,j} = 1$.

[8] Example: finite differences for a 1-D linear, constant coefficients, equation for wave propagation.

remain bounded as $M \to \infty$, for $0 \leq t_j = j\,\Delta t \leq T$, $\Delta t = T/M$, and T fixed — perhaps with a constraint[9] relating Δx to Δt. In particular, this happens if $|G| \leq 1$ for all solutions of this form.

However, for the result in § 1.5 to be true, it must be that all the solutions are of the form in (2.1.1), or linear combinations of them. This motivates the **question:** *Given any sequence $\{u_n\}_{-\infty}^{\infty}$, with $u_{n+N} = u_n$ for some integer $N > 0$, can we write*

$$u_n = \sum_{\ell} a_\ell \, e^{i\,k_\ell\,n} \tag{2.1.2}$$

for some finite set of coefficients a_ℓ and wave numbers k_ℓ?

The answer to this question is **yes**, and is given in detail by theorem 2.1.1 below.

Before going into details, notice that **periodicity** $u_{n+N} = u_n$ **constraints the possible wave numbers that can occur in (2.1.2),** since it requires that $k_\ell\,N$ be a multiple of 2π. Thus

> **the wave numbers are restricted to the set $\boldsymbol{k_\ell = \frac{2\pi}{N}\,\ell}$, where $\boldsymbol{\ell}$ is an integer.**
> Furthermore, $e^{i\,k_\ell\,n} = e^{i\,k_{\ell+N}\,n}$ for any integer n, hence
> **the wave numbers can be selected in any range $\boldsymbol{\ell_* \leq \ell \leq \ell_* + N - 1}$,** where $\boldsymbol{\ell_*}$ is some arbitrary integer. (2.1.3)

Theorem 2.1.1 *Given any periodic sequence $\{u_n\}_{n=-\infty}^{n=+\infty}$ of complex numbers, with period N — $u_{n+N} = u_n$, one can write*

$$u_n = \sum_{\ell=\ell_*}^{\ell=\ell_*+N-1} a_\ell \, e^{i\,k_\ell\,n}, \quad \textit{where } \ell_* \textit{ is any integer, } k_\ell = \frac{2\,\pi}{N}\,\ell, \tag{2.1.4}$$

and $\{a_\ell\}_{\ell=-\infty}^{\ell=+\infty}$ is the periodic, with period N, sequence of complex numbers given by

$$a_\ell = \frac{1}{N} \sum_{n=n_*}^{n=n_*+N-1} u_n \, e^{-i\,k_n\,\ell}, \quad \textit{where } n_* \textit{ is any integer.} \tag{2.1.5}$$

The transformation between periodic sequences $\{u_n\} \to \{a_\ell\}$ in (2.1.5), giving the coefficients a_ℓ in (2.1.4), is the **Discrete Fourier Transform (DFT).** *It's inverse in (2.1.4) is the* **Inverse Discrete Fourier Transform (iDFT).** *The names follow from the connection with Fourier series — see § 2.3.*

Remark 2.1.1 *Consider the case when $\ell_* = n_* = 1$. Then (2.1.4) and (2.1.5) become*

$$u_n = \sum_{\ell=1}^{\ell=N} a_\ell \, e^{i\,k_\ell\,n} \quad \textit{and} \quad a_\ell = \frac{1}{N} \sum_{n=1}^{n=N} u_n \, e^{-i\,k_n\,\ell}. \tag{2.1.6}$$

Because of the periodicity, we need only consider $\{u_n\}$ and $\{a_\ell\}$ for $1 \leq n,\ \ell \leq N$. *Hence, in terms of*

(a) the N-vectors $\vec{\boldsymbol{u}}$ and $\vec{\boldsymbol{a}}$ whose components are $\{u_n\}$ and $\{a_\ell\}$, respectively,

(b) the $N \times N$ matrix $\boldsymbol{\mathcal{D}}$ whose entries are $\boldsymbol{D_{\ell n} = e^{i\,k_\ell\,n} = w^{\ell n}}$, where $\boldsymbol{w = e^{i\,2\,\pi/N}}$,

we can write

$$\vec{u} = \mathcal{D}\,\vec{a} \quad \textit{and} \quad \vec{a} = \frac{1}{N}\,\mathcal{D}^\dagger\,\vec{u}, \quad \Longleftrightarrow \quad \mathcal{D}^{-1} = \frac{1}{N}\,\mathcal{D}^\dagger, \tag{2.1.7}$$

where † denotes the adjoint of a matrix.

[9] Examples: $\Delta t \leq \text{constant}\Delta x$ or $\Delta t \leq \text{constant}(\Delta x)^2$. The constraint must allow Δx to vanish as $\Delta t \to 0$, so that convergence can occur.

Problem 2.1.1 **Verify the DFT and iDFT formulas.**
Show that

(i) If $\{a_\ell\}_{\ell=-\infty}^{\ell=+\infty}$ *is a periodic sequence of period* N, **the value of the right hand side in (2.1.4) is independent of the choice of** ℓ_***.** *Similarly, if* $\{u_n\}_{n=-\infty}^{n=+\infty}$ *is a periodic sequence of period* N, **the value of the right hand side in (2.1.5) is independent of the choice of** n_***.**

(ii) For any set of constants[10] a_ℓ, $\ell_* \leq \ell < \ell_* + N$, $\{u_n\}$ *— as given by (2.1.4), is periodic of period* N*. Similarly, for any set of constants* u_n, $n_* \leq n < n_* + N$, $\{a_\ell\}$ *— as given by (2.1.5), is periodic of period* N*.*

(iii) Substituting (2.1.5) into the right hand side of (2.1.4) yields the left hand side.

(iv) Substituting (2.1.4) into the right hand side of (2.1.5) yields the left hand side.

Hints:

1) For part (i), denote by $S_a = S_a(\ell_*)$ the value of the right hand side of (2.1.4) as a function of ℓ_* — for some given, fixed, periodic sequence $\{a_\ell\}_{-\infty}^{+\infty}$. Then show that $S(\ell_* + 1) = S(\ell_*)$ for any ℓ_*, from which $S = \text{constant}$ follows. The same idea works for (2.1.5).

2) You will need the following result, obtained in problem 1.5.1: $\boldsymbol{\sum_{n=1}^{N} w^{n\,(\ell_1-\ell_2)} = N\,\delta_{\ell_1\,\ell_2}}$ for $1 \leq \ell_1,\, \ell_2 \leq N$, where $N > 1$ is an integer, $\boldsymbol{w = e^{i\,2\,\pi/N}}$ is the N^{th} fundamental root of unit, and $\boldsymbol{\delta_{i\,j}}$ denotes the Kronecker delta: $\delta_{i\,j} = 0$ if $i \neq j$, and $\delta_{j\,j} = 1$.

3) Note that $\boldsymbol{e^{i\,k_\ell\,n} = w^{\ell\,n}}$, and $\boldsymbol{e^{-i\,k_n\,\ell} = w^{-\ell\,n}}$.

2.2 Alternative formulations for the DFT.

If we define $\boldsymbol{v_n = u_{n-1}}$ and $\boldsymbol{b_\ell = a_{\ell-1}}$, then the equations in (2.1.4–2.1.5) take the form[11]

$$\underbrace{v_n = \sum_{\ell=\ell_*}^{\ell_*+N-1} b_\ell\, w^{(\ell-1)\,(n-1)}}_{\textbf{iDFT}} \quad \text{and} \quad \underbrace{b_\ell = \frac{1}{N}\sum_{n=n_*}^{n_*+N-1} v_n\, w^{-(n-1)\,(\ell-1)}}_{\textbf{DFT}}. \tag{2.2.1}$$

where ℓ_*, n_* are arbitrary integers, and $\boldsymbol{w = e^{i\,2\,\pi/N}}$. These two formulas constitute an alternative formulation of the DFT:$\boldsymbol{\{v_n\} \longrightarrow \{b_\ell\}}$, and the iDFT:$\boldsymbol{\{b_\ell\} \longrightarrow \{v_n\}}$, relating $\boldsymbol{N}$-periodic sequences.

In particular, if we select $\ell_* = n_* = 1$, we obtain

$$v_n = \sum_{\ell=1}^{N} b_\ell\, w^{(\ell-1)\,(n-1)} \quad \text{and} \quad b_\ell = \frac{1}{N}\sum_{n=1}^{N} v_n\, w^{-(n-1)\,(\ell-1)}. \tag{2.2.2}$$

[10] For this part of the problem there is no sequence $\{a_\ell\}$, just N constants.

[11] To get the equations here, replace $\ell_* \to \ell_* - 1$ in (2.1.4) and $n_* \to n_* - 1$ in (2.1.4).

If v is the N-vector array whose entries are the $\{v_n\}$, and b is the N-vector array whose entries are the $\{b_\ell\}$, then the transformations in (2.2.2) are executed by the **MATLAB commands**

$$\mathsf{b} = \frac{1}{N}\,\mathsf{fft(v)} \quad \text{and} \quad \mathsf{v} = N\,\mathsf{ifft(b)}. \tag{2.2.3}$$

Remark 2.2.1 *A numerical implementation of the DFT and iDFT, as written in (2.2.2), has an $O(N^2)$ operation, and it is thus rather costly. Fortunately, there is a way (algorithm) to organize the calculations that leads to computation whose operation count is $O(N\ln(N))$. This algorithm is known by the name of the* **Fast Fourier Transform (FFT)**, *with inverse given by the* **inverse Fast Fourier Transform (iFFT)**. *This algorithm, of course, is the one MATLAB implements.*[12]

The FFT is important because it allows the fast/efficient implementation of the **DFT** *and* **iDFT**, *which* **allow the approximate, efficient, and accurate, calculation of Fourier series and Fourier coefficients.**[13] *Since Fourier series appear in very many applications, the FFT and iFFT are widely used.* A brief description of the main idea behind the FFT algorithm is included in § 2.5.

2.3 Relationship between Fourier series and the DFT/iDFT.

Let $f = f(x)$ be a sufficiently nice[14] periodic function — assume (for simplicity) that the period is $2\,\pi$, so that $f(x + 2\,\pi) = f(x)$. Then f has a Fourier series

$$f(x) = \sum_{-\infty}^{\infty} \mathcal{C}_n(f)\, e^{i\,n\,x}, \tag{2.3.1}$$

where $$\mathcal{C}_n = \frac{1}{2\,\pi}\int_{x_0}^{x_0+2\,\pi} f(x)\, e^{-i\,n\,x}\, dx \quad \text{for} \quad n = 0,\, \pm1,\, \pm2,\, \pm3,\, \dots \tag{2.3.2}$$

Here x_0 is any real number — the Fourier coefficients $\mathcal{C}_n$ do not depend on the choice of x_0.

There are many notions (types of convergence) in which the right hand side in (2.3.1) can represent f. We will explore this question (lightly) in § 2.4. The important point here is that, the smoother the function f is, the better and faster (less terms needed to get a good approximation) the convergence of the series in (2.3.1), in particular: for smooth functions the convergence is very fast:

$$f(x) = \sum_{-N}^{N} \mathcal{C}_n(f)\, e^{i\,n\,x} + \mathcal{E}_N, \tag{2.3.3}$$

where $\mathcal{E}_N \to 0$ vanishes faster than any power N^{-p} as $N \to \infty$.

[12] Hence the command names fft and ifft in (2.2.3).

[13] We study the relationship between the DFT/iDFT and Fourier series in § 2.3.

[14] Example: f has two derivatives, with f'' integrable — see § 2.4. Note that far less is needed for a Fourier series to exist and converge — e.g.: see § 2.4.1.

Here we will assume that f is "nice" enough to justify the calculations below (two continuous derivatives is enough, but not necessary). To be precise, here **we assume that**

$$\left.\begin{array}{l}\text{The series in (2.3.1) converges absolutely and uniformly. In fact, we further}\\ \text{assume that}\\ \qquad\qquad |\mathcal{C}_n(f)| \leq \dfrac{\mathcal{C}}{|n|^p} \quad \text{for} \quad n \neq 0,\\ \text{where } \mathcal{C} > 0 \text{ and } p > 1 \text{ are constants — e.g. see remark 2.4.1.}\end{array}\right\} \tag{2.3.4}$$

Introduce a numerical grid on the line by

$$\boldsymbol{x_n = (n-1)\,\Delta x}, \quad \text{where} \;\; \boldsymbol{\Delta x = \frac{2\,\pi}{N}} \tag{2.3.5}$$

$\boldsymbol{n}$ runs over the integers, and $\boldsymbol{N}$ is some "large" natural number. Then use the trapezoidal rule to approximate the integrals in (2.3.2), thus obtaining discrete approximations for the Fourier coefficients:

$$\mathcal{C}_\ell = \frac{1}{2\,\pi}\int_0^{2\,\pi} f(x)\,e^{-i\,\ell\,x}\,dx \approx \frac{1}{2\,\pi}\sum_{n=1}^{N} f_n\,e^{-i\,\ell\,x_n}\,\Delta x = \frac{1}{N}\sum_{n=1}^{N} f_n\,w^{-(n-1)\,\ell} \tag{2.3.6}$$

where $\boldsymbol{f_n = f(x_n)}$ and (as usual) $\boldsymbol{w = e^{i\,2\,\pi/N}}$. Thus we write

$$\mathcal{C}_\ell \approx c_\ell \quad \text{where} \quad c_\ell = \frac{1}{N}\sum_{n=1}^{N} f_n\,w^{-(n-1)\,\ell}. \tag{2.3.7}$$

Now, question:

$$\textbf{How good an approximation to } \boldsymbol{\mathcal{C}_\ell} \textbf{ is } \boldsymbol{c_\ell}\textbf{?} \tag{2.3.8}$$

Remark 2.3.1 *From the results in § 2.6, we should expect the approximation to be quite good, but some* caution is needed: *the results in § 2.6 indicate that, for a* <u>*fixed*</u> *"nice enough" function, the trapezoidal rule provides a very good approximation to the integral of the function as $N \to \infty$. However, here we have a whole sequence of functions that we are integrating (one for each ℓ), so the results in § 2.6 have to be taken with a grain of salt. At best they can be used to state that,* for a fixed set of $\mathcal{C}_\ell$, *the approximation in (2.3.7) gets better very fast as $N \to \infty$ (at least for functions with many derivatives).*

Remark 2.3.2 *A* second note of caution *comes from the observation that, when ℓ is comparable with N in size, the integrand in (2.3.6) has $O(1)$ oscillations that occur on the same scale as Δx. Clearly, when this happens, (2.3.7) can be a good approximation by accident only — definitely not for generic functions f, no matter how nice they might be!*

Remark 2.3.3 *A* final note of caution *comes from the observation that (2.3.7) defines c_ℓ for all integer values of ℓ, in such a way that $\boldsymbol{c_{\ell+N} = c_\ell}$. This makes $\mathcal{C}_\ell \approx c_\ell$ compatible with (2.3.4) only for the trivial case $\mathcal{C}_\ell \equiv 0$. Clearly, in general $\boldsymbol{\mathcal{C}_\ell \approx c_\ell}$* **must fail for $|\boldsymbol{\ell}| = \boldsymbol{O(N)}$ or larger.**

When (2.3.4) applies, an explicit formula relating the c_ℓ to the $\mathcal{C}_\ell$ is possible, which can be used to answer (2.3.8). The idea is to plug in the Fourier series for f into the definition of c_ℓ. This yields[15]

$$c_\ell = \frac{1}{N}\sum_{n=1}^{N}\left(\sum_{j=-\infty}^{\infty}\mathcal{C}_j(f)\,e^{i\,j\,x_n}\right)\,w^{-(n-1)\,\ell} = \frac{1}{N}\sum_{j=-\infty}^{\infty}\mathcal{C}_j(f)\sum_{n=1}^{N}w^{(n-1)(j-\ell)} = \sum_{k=-\infty}^{\infty}\mathcal{C}_{\ell+k\,N}, \qquad (2.3.9)$$

where we used that $J=\sum_{n=1}^{N}w^{(n-1)\,q}=0$, unless $q=$ multiple of N, when $J=N$. Note that this formula also shows that $\{c_\ell\}$ is periodic of period N. This last equation can also be rewritten as

$$c_\ell = \mathcal{C}_\ell + (\mathcal{C}_{\ell+N}+\mathcal{C}_{\ell-N}) + (\mathcal{C}_{\ell+2N}+\mathcal{C}_{\ell-2N}) + \dots \qquad (2.3.10)$$

This makes it clear that the approximation in (2.3.7–2.3.8) is not very good outside the range $|\ell| \ll N$. On the other hand, it also shows that we can write

$$\boldsymbol{c_\ell = \mathcal{C}_\ell + O(N^{-p})} \quad \text{for} \quad \boldsymbol{|\ell| \leq \frac{N}{2}}, \qquad (2.3.11)$$

where we have used (2.3.4). Hence, **when p is large (the function has many derivatives), the approximation in (2.3.7) is very good, at least for $|\ell| \leq \frac{N}{2}$.**

Remark 2.3.4 *Note that, for $\ell \approx \pm\frac{1}{2}N$ the relative error in (2.3.11) is quite large, since then all the terms have (roughly) the same size. However, in this case both c_ℓ and $\mathcal{C}_\ell$ are small, and then it does not matter.*

Guided by the results above, we now complete the discretization of the Fourier series formulas by adding to the approximation in (2.3.7) of the Fourier coefficients, the following approximations to the Fourier series

$$f(x) \approx \sum_{\ell=-M}^{M} c_\ell\,e^{i\,\ell\,x} \quad \text{for } N = 2\,M+1 \text{ odd.}$$

$$f(x) \approx \frac{1}{2}\,c_{-M}\,e^{-i\,M\,x} + \sum_{\ell=1-M}^{M-1} c_\ell\,e^{i\,\ell\,x} + \frac{1}{2}\,c_M\,e^{i\,M\,x} \quad \text{for } N = 2\,M \text{ even.}$$

When used on the grid points x_n these yield

$$\begin{aligned} f_n &= \sum_{\ell=-M}^{M} c_\ell\,w^{\ell\,(n-1)} &&= \sum_{0}^{N-1} c_\ell\,w^{\ell\,(n-1)}, \\ f_n &= \frac{1}{2}\,c_{-M}\,w^{-M\,(n-1)} + \sum_{\ell=1-M}^{M-1} c_\ell\,w^{\ell\,(n-1)} + \frac{1}{2}\,c_M\,w^{M\,(n-1)} &&= \sum_{0}^{N-1} c_\ell\,w^{\ell\,(n-1)}, \end{aligned}$$

where in each case we use the periodicity properties to shift the summations to the range $0 \leq \ell < N$.

[15] Absolute convergence justifies all the calculations here.

Putting this all together yields the following **discrete approximation to the Fourier series:**

$$\boldsymbol{f_n = \sum_{0}^{N-1} c_\ell\, w^{\ell\,(n-1)}} \quad \text{and} \quad \boldsymbol{c_\ell = \frac{1}{N} \sum_{n=1}^{N} f_n\, w^{-(n-1)\,\ell},} \tag{2.3.12}$$

where the $\{f_n\}$ correspond to the function values at the grid points — $f_n \approx f(x_n)$, and the $\{c_ell\}$ are approximations to the Fourier coefficients — see (2.3.11).

$$\left.\begin{array}{l}\textbf{Equation (2.3.12) should be compared with the DFT and iDFT} \\ \textbf{formulas in (2.2.2) and (2.2.3). It should be clear that, with} \\ \textbf{the identification } \boldsymbol{v_n = f_n} \textbf{ and } \boldsymbol{b_\ell = c_{\ell-1}}\textbf{, they are the same.}\end{array}\right\} \tag{2.3.13}$$

2.3.1 Fourier series, DFT, and derivatives.

This subsection is yet to be written.

2.4 Simple convergence results for Fourier series.

In this subsection we present some simple convergence results for Fourier series

Theorem 2.4.1 *Let $f = f(x)$ be a periodic function of period $2\,\pi$ — $f(x + 2\,\pi) = f(x)$. Assume that f has at least two derivatives, with f'' integrable. Then* **the Fourier series in (2.3.1) converges to f absolutely and uniformly.** *In fact*

$$\left| f(x) - \sum_{-N}^{N} \mathcal{C}_n(f)\, e^{i\,n\,x} \right| \leq 2\,\|f''\|_1 \sum_{n>N} \frac{1}{n^2} = O\left(\frac{1}{N^2}\right), \quad \textit{where} \quad \|f''\|_1 = \frac{1}{2\,\pi} \int_0^{2\,\pi} |f''(x)|\; dx \tag{2.4.1}$$

and $N \geq 0$ is any natural number.

Proof. Integrate by parts twice, and use the periodicity, to obtain

$$\mathcal{C}_n(f) = \frac{1}{2\,\pi\,i\,n} \int_0^{2\,\pi} f'(x)\, e^{-i\,n\,x}\, dx = -\frac{1}{2\,\pi\,n^2} \int_0^{2\,\pi} f''(x)\, e^{-i\,n\,x}\, dx, \tag{2.4.2}$$

from which

$$|\mathcal{C}_n(f)| \leq \frac{1}{2\,\pi\,n^2} \int_0^{2\,\pi} |f''(x)|\; dx = \frac{1}{n^2} \|f''\|_1. \tag{2.4.3}$$

It follows that the series on the right hand side of (2.3.1) converges absolutely and uniformly. Hence it defines a continuous function

$$g(x) = \sum_{-\infty}^{\infty} \mathcal{C}_n(f)\, e^{i\,n\,x}. \tag{2.4.4}$$

Clearly, it is also the case that $\mathcal{C}_n(g) = \mathcal{C}_n(f)$ for all n. Hence, from theorem 2.4.2, $g = f$. The result in (2.4.1) then follows from (2.4.3). ♣

Remark 2.4.1 *It should be clear that, if f is smoother than in the statement of the theorem above, then the convergence properties of the Fourier series are even better. For example, if f has $p \geq 2$ derivatives, with $f^{(p)}$ integrable, then (2.4.1) generalizes to*

$$\left| f(x) - \sum_{-N}^{N} \mathcal{C}_n(f)\, e^{i\,n\,x} \right| \leq 2\,\|f^{(p)}\|_1 \sum_{n>N} \frac{1}{n^p} = O\left(\frac{1}{N^p}\right), \quad \textit{where } \|f^{(p)}\|_1 = \frac{1}{2\,\pi}\int_0^{2\,\pi} \left|f^{(p)}(x)\right|\, dx \tag{2.4.5}$$

and $N \geq 0$ is any natural number.[16] *In particular:* **for smooth functions, the Fourier series converges faster than any negative power of N = number of terms in the partial sums.**

On the other hand, Fourier series also converge for periodic functions that are less smooth than in the statement of the theorem above — although the convergence is then, generally, in weaker senses than absolutely and uniformly. For example:

(1) If f is square integrable, then the Fourier series for f converges to f in the square norm. This means that

$$\left\| f(x) - \sum_{-N}^{N} \mathcal{C}_n(f)\, e^{i\,n\,x} \right\|_2 \to 0 \;\; \textit{as } N \to \infty, \quad \textit{where } \|g\|_2 = \sqrt{\frac{1}{2\,\pi}\int_0^{2\,\pi} |g^2(x)|\, dx} \tag{2.4.6}$$

for any function square integrable function g.

(2) Let f be such that: (i) f is discontinuous at a finite set of points per period. (ii) Let a and b be two successive points of discontinuity for f. Then f is twice continuously differentiable for $a \leq x \leq b$. In this case the Fourier series for f converges as follows:

— To $\boldsymbol{f(x)}$ at every point $\boldsymbol{x}$ at which $\boldsymbol{f}$ is continuous.

— To $\frac{1}{2}\left(\boldsymbol{f(x+0)} + \boldsymbol{f(x-0)}\right)$ at every point $\boldsymbol{x}$ at which $\boldsymbol{f}$ is discontinuous.

It is important to note that, in this case the convergence is not uniform. In fact, near each discontinuity the partial Fourier sums over-shoot and under-shoot, by amounts that do not vanish as $N \to \infty$, and exhibit a oscillations in a small[17] interval near each discontinuity. *This is the* **Gibbs' phenomenon,** *of which we will see more later.*

There are many, many, more known results characterizing how Fourier series converge under various conditions. Another example is given in § 2.4.1.

Theorem 2.4.2 *Let $f(x)$ and $g(x)$ be two continuous, 2π-periodic functions with the same Fourier coefficients. Then $f = g$.*

Proof. Clearly, $\mathcal{C}_n(h) = 0$ for all n, where $h = g - f$. Hence

$$\int_{a-\pi}^{a+\pi} h(x) \underbrace{\left(\sum c_n\, e^{-i\,n\,x}\right)}_{p=p(x)} dx = 0, \tag{2.4.7}$$

[16] The proof of this follows by performing p integrations by parts, instead of only two, in (2.4.2).

[17] Vanishes as $N \to \infty$.

where a is arbitrary and the summation is over a finite number of exponentials, with coefficients c_n — we call p a trigonometric polynomial. We show below that this implies that $h \equiv 0$. (2.4.8)

Define the following trigonometric polynomials:

$$p_m(x) = \gamma_m^{-1} \left(\frac{e^{i\,x} + 2 + e^{-i\,x}}{4} \right)^m = \gamma_m^{-1} \left(\cos^2 \left(\frac{x}{2} \right) \right)^m , \tag{2.4.9}$$

where $m = 1,\, 2,\, 3,\, \ldots$ and

$$\gamma_m = \int_{-\pi}^{+\pi} \left(\cos \frac{x}{2} \right)^{2\,m} dx > 0. \tag{2.4.10}$$

It should be clear that the p_m have the properties below — see figure 2.4.1.

(i) $p_m(-\pi) = p_m(\pi) = 0$ and $p_m(x) > 0$ for $-\pi < x < \pi$, with $\int_{-\pi}^{+\pi} p_m(x)\, dx = 1$.

(ii) For any ϵ, $\delta > 0$ there exists $0 < M < \infty$ such that $p_m(x) \leq \epsilon$ for $\delta \leq |x| \leq \pi$ and $m \geq M$.

These properties imply that $p_m(x)$ behaves like the Dirac "delta function" as $m \to \infty$. In particular, let $y = y(x)$ be a continuous function in $-\pi \leq x \leq \pi$, then

$$\lim_{m \to \infty} \int_{-\pi}^{+\pi} y(x)\, p_m(x)\, dx = y(0). \tag{2.4.11}$$

This is proved in lemma 2.4.1. Thus we can write, using (2.4.7)

$$0 = \lim_{m \to \infty} \int_{a-\pi}^{a+\pi} h(x) p_m(x-a)\, dx = \lim_{m \to \infty} \int_{-\pi}^{+\pi} h(x+a) p_m(x)\, dx = h(a). \tag{2.4.12}$$

Since a is arbitrary, this proves (2.4.8). ♣

Lemma 2.4.1 *Equation (2.4.11) applies.*

Proof. Let $Y = \max\limits_{-\pi \leq x \leq \pi} |y(x) - y(0)|$. Also, because y is continuous, for any $\epsilon > 0$, there is a $\delta > 0$ such that $|y(0) - y(x)| \leq \epsilon$ for $|x| \leq \delta$. Select now M as in item (ii) above (2.4.11). Then, for $m \geq M$

$$\begin{aligned} \underbrace{\int_{-\pi}^{+\pi} y(x)\, p_m(x)\, dx - y(0)}_{I} \quad &= \quad \underbrace{\int_{-\pi}^{-\delta} (y(x) - y(0))\, p_m(x)\, dx}_{I_1} + \underbrace{\int_{\delta}^{\pi} (y(x) - y(0))\, p_m(x)\, dx}_{I_2} \\ &+ \quad \underbrace{\int_{-\delta}^{+\delta} (y(x) - y(0))\, p_m(x)\, dx}_{I_3}, \end{aligned} \tag{2.4.13}$$

where we have used item (i) above (2.4.11). Clearly $|I_1| \leq \epsilon\, Y\, (\pi - \delta)$ and $|I_2| \leq \epsilon\, Y\, (\pi - \delta)$, while

$$|I_2| \leq \epsilon \int_{-\delta}^{+\delta} p_m(x)\, dx \leq \epsilon.$$

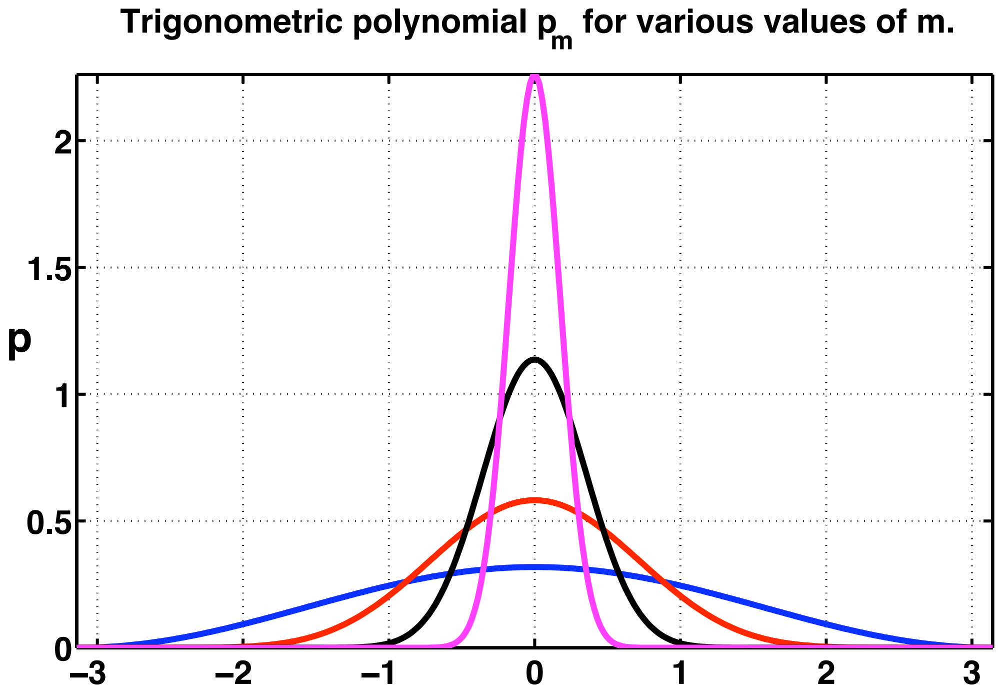

Figure 2.4.1: Trigonometric polynomials p_m for: $m = 1$ (blue), $m = 4$ (red), $m = 16$ (black), and $m = 64$ (magenta). As $m \to \infty$, $p_m \sim \delta(x)$ for $|x| \leq \pi$ — where $\delta(x)$ is the Dirac "delta function".

Hence

$$\left| \int_{-\pi}^{+\pi} y(x)\, p_m(x)\, dx - y(0) \right| \leq \epsilon + 2\,\epsilon\, Y\,(\pi - \delta),$$

which shows that I can be made as small as desired by taking m large enough. ♣

Problem 2.4.1 **Compute the asymptotic value of an integral.**

Find the leading order contribution to the integral defining γ_m in equation (2.4.10), when $m \gg 1$.

Hint. As items (i-ii) below equation (2.4.10) show — see also figure 2.4.1, as $m \to \infty$, the main contribution to the integral defining γ_m arises from a small neighborhood of the origin. To be precise

$$\gamma_m = 2 \int_0^{\delta} \left(\cos\frac{x}{2}\right)^{2\,m} dx + \underbrace{2 \int_{\delta}^{+\pi} \left(\cos\frac{x}{2}\right)^{2\,m} dx}_{I}, \quad \text{where} \;\; \delta = 2\,m^{-\frac{3}{8}}. \tag{2.4.14}$$

However, since $0 \leq \cos\left(\frac{x}{2}\right) \leq e^{-\frac{1}{8}x^2}$, it follows that

$$0 < I \leq 2 \int_{\delta}^{+\pi} e^{-\frac{m}{4}x^2}\, dx \leq 2\,(\pi - \delta)\, e^{-\frac{m}{4}\delta^2} \leq 2\,\pi\, e^{-m^{\frac{1}{4}}} = \text{exponentially small.} \tag{2.4.15}$$

Hence, for $m \gg 1$,

$$\gamma_m = 2 \int_0^{\delta} \left(\cos\frac{x}{2}\right)^{2\,m} dx + O\left(e^{-m^{\frac{1}{4}}}\right). \tag{2.4.16}$$

This can be exploited to find a simple approximation to the value of γ_m for $m \gg 1$, as follows:

(a) Write $\left(\cos\left(\frac{x}{2}\right)\right)^{2\,m} = \exp\left(2\,m\,\ln\cos\left(\frac{x}{2}\right)\right)$.

(b) Expand $\ln\cos\left(\frac{x}{2}\right)$ in powers of x, to obtain an approximation to $\left(\cos\left(\frac{x}{2}\right)\right)^{2\,m}$, valid for $0 \leq x \leq \delta$.

(c) Substitute the result of (b) into (2.4.16), and do the integral. This should give a very simple formula for the approximate value of γ_m for $m \gg 1$.

(d) You will need the fact that $\int_0^\infty e^{-x^2}\,dx = \frac{1}{2}\sqrt{\pi}$.

2.4.1 Convergence in the weak sense.

This subsubsection is yet to be written.

2.5 The main idea behind the FFT algorithm.

This subsection is yet to be written.

2.6 Trapezoidal rule and the Euler-Maclaurin formula.

The purpose of this subsection is to produce explicit formulas for the error in the trapezoidal rule of integration. In particular, we will show that: *when integrating a periodic function over a period, the order of the error is directly proportional to the degree of smoothness of the integrand.*

We begin by considering a function $f = f(x)$ with N continuous derivatives, defined for $0 \leq x \leq 1$. Then we use integration by parts, using the properties of the Bernoulli polynomials — see § 2.6.1 — to obtain a formula for $\int_0^1 f(x)\,dx$ in terms of the values of the function and its derivatives at the end points. Note that

$$\begin{aligned}
\int_0^1 f(x)\,dx &= \int_0^1 B_0(x)\,f(x)\,dx, \\
\int_0^1 B_0(x)\,f(x)\,dx &= \int_0^1 B_1'(x)\,f(x)\,dx &&= \frac{1}{2}\,(f(1)+f(0)) - \int_0^1 B_1(x)\,f'(x)\,dx, \\
\int_0^1 B_1(x)\,f'(x)\,dx &= \int_0^1 \frac{B_2'(x)}{2}\,f'(x)\,dx &&= \frac{\beta_2}{2}\,(f'(1)-f'(0)) - \int_0^1 \frac{B_2(x)}{2}\,f''(x)\,dx, \\
\int_0^1 \frac{B_2(x)}{2}\,f''(x)\,dx &= \int_0^1 \frac{B_3'(x)}{6}\,f''(x)\,dx &&= \frac{\beta_3}{6}\,(f''(1)-f''(0)) - \int_0^1 \frac{B_3(x)}{6}\,f^{(3)}(x)\,dx,
\end{aligned}$$

and so on, with the general formula for $n \geq 1$ being

$$\begin{aligned}
\int_0^1 \frac{B_n(x)}{n!}\,f^{(n)}(x)\,dx &= \int_0^1 \frac{B_{n+1}'(x)}{(n+1)!}\,f^{(n)}(x)\,dx \\
&= \frac{\beta_{n+1}}{(n+1)!}\left(f^{(n)}(1) - f^{(n)}(0)\right) - \int_0^1 \frac{B_{n+1}(x)}{(n+1)!}\,f^{(n+1)}(x)\,dx.
\end{aligned}$$

Putting this all together, we obtain

$$\begin{aligned}\int_0^1 f(x)\,dx &= \frac{1}{2}\left(f(1)+f(0)\right)-\frac{\beta_2}{2}\left(f'(1)-f'(0)\right)+\frac{\beta_3}{6}\left(f''(1)-f''(0)\right)+\ldots+\\ &\quad (-1)^{N+1}\,\frac{\beta_N}{N!}\left(f^{(N-1)}(1)-f^{(N-1)}(0)\right)+\underbrace{(-1)^N\int_0^1\frac{B_N(x)}{N!}\,f^{(N)}(x)\,dx}_{\mathbf{\mathcal{E}_N=\mathcal{E}_N(f,\,0,\,1)}}\\ &= \frac{1}{2}\left(f(1)+f(0)\right)+\sum_{n=1}^{N-1}(-1)^n\,\frac{\beta_{n+1}}{(n+1)!}\left(f^{(n)}(1)-f^{(n)}(0)\right)+\mathcal{E}_N. \qquad (2.6.1)\end{aligned}$$

We are now ready to write the **Euler-Maclaurin formula**, which provides an expression for the error in approximating the integral of a function using the trapezoidal rule. **Let $\boldsymbol{g=g(x)}$ be a function with $\boldsymbol{N}$ continuous derivatives, defined in some interval $\boldsymbol{a\le x\le b}$.**

Introduce a numerical grid in $a\le x\le b$ by defining $\boldsymbol{x_\ell=a+\ell\,h}$ for $\boldsymbol{0\le\ell\le M}$, where $\boldsymbol{M>0}$ is a natural number and $\boldsymbol{h=\Delta x=\frac{b-a}{M}}$. Then we can write

$$\int_a^b g(x)\,dx=\sum_{\ell=0}^{M-1}\int_{x_\ell}^{x_\ell+h}g(x)\,dx=h\sum_{\ell=0}^{M-1}\int_0^1 g(x_\ell+h\,x)\,dx. \qquad (2.6.2)$$

We now apply the equality in (2.6.1) to each one of the terms in the sum on the right hand side in (2.6.2), with $f(x)=f_\ell(x)=g(x_\ell+h\,x)$ in each case. It is then easy to see that this yields

$$\begin{aligned}\int_a^b g(x)\,dx &= \underbrace{\left(\frac{1}{2}\,g(a)+\sum_{\ell=1}^{M-1}g(x_\ell)+\frac{1}{2}\,g(b)\right)h}_{\text{Trapezoidal rule.}}-\sum_{n=1}^{N-1}(-h)^{n+1}\,\frac{\beta_{n+1}}{(n+1)!}\left(g^{(n)}(b)-g^{(n)}(a)\right)+\\ &\quad \underbrace{h\sum_{\ell=0}^{M-1}\mathcal{E}_N(f_\ell,\,0,\,1)}_{\mathbf{\mathcal{E}_N(g)=\mathcal{E}_N(g,\,a,\,b)}}, \qquad (2.6.3)\end{aligned}$$

which is the **Euler-Maclaurin summation formula.** In particular, note that

$$\mathcal{E}_N(f_\ell,\,0,\,1)=(-h)^N\int_0^1\frac{B_N(x)}{N!}\,g^{(N)}(x_\ell+h\,x)\,dx. \qquad (2.6.4)$$

Hence, using (2.6.12) we see that

$$|\mathcal{E}_N(f_\ell,\,0,\,1)|\le h^N\int_0^1\left|g^{(N)}(x_\ell+h\,x)\right|\,dx=h^{N-1}\int_{x_\ell}^{x_\ell+h}\left|g^{(N)}(x)\right|\,dx, \qquad (2.6.5)$$

from which it follows that

$$|\mathcal{E}_N(g,\,a,\,b)|\le h^N\int_a^b\left|g^{(N)}(x)\right|\,dx. \qquad (2.6.6)$$

Remark 2.6.1 **Trapezoidal rule for periodic functions.** *Suppose that g above in (2.6.3) is* **periodic of period $b-a$**. *Then*

$$\int_a^b g(x)\,dx = h\sum_{\ell=1}^{M} g(x_\ell) + \mathcal{E}_N(g,\,a,\,b), \quad \textit{where } \mathcal{E}_N(g,\,a,\,b) = O(h^N). \tag{2.6.7}$$

In particular: **for smooth periodic functions, the error the trapezoidal rule approximation to their integral over one period vanishes, as $h \longrightarrow 0$, faster than any power of h.**

By contrast, notice that Equations (2.6.3) and (2.6.6) show that, for **generic functions** *(with, at least, a second derivative that is integrable) the* **trapezoidal rule is second order** *only*

$$\int_a^b g(x)\,dx = \left(\frac{1}{2}\,g(a) + \sum_{\ell=1}^{M-1} g(x_\ell) + \frac{1}{2}\,g(b)\right) h - h^2\,\frac{1}{2}\,\beta_2\,(g'(b)-g'(a)) + \mathcal{E}_2(g,\,a,\,b), \tag{2.6.8}$$

where $\mathcal{E}_2(g,\,a,\,b) = O(h^2)$.

2.6.1 Bernoulli polynomials.

The **Bernoulli polynomials $B_n = B_n(x)$** and the **Bernoulli numbers β_n** are **defined** as follows

$$\left.\begin{array}{llll} \text{(a)} & B_0 & = & 1, \\ \text{(b)} & B_n' & = & n\,B_{n-1} \qquad \text{for } n>0, \\ \text{(c)} & 0 & = & \int_0^1 B_n(x)\,dx \quad \text{for } n>0, \\ \text{(d)} & \beta_n & = & B_n(0), \end{array}\right\} \tag{2.6.9}$$

where the prime denotes $\frac{d}{dx}$ and $n = 0,\,1,\,2,\,3,\,\ldots$ These equations determine the polynomials recursively — the arbitrary constant of integration in (b) is determined by the condition in (c). The first few Bernoulli polynomials are $\boldsymbol{B_0 = 1}$, $\boldsymbol{B_1 = x - \frac{1}{2}}$, $\boldsymbol{B_2 = \frac{1}{2}\,x^2 - \frac{1}{2}\,x + \frac{1}{12}}$, ...

Notice that

B_n is a degree n polynomial. (2.6.10)

$B_n(0) = B_n(1) = \beta_n$ for $n \geq 2$. (2.6.11)

The first statement here follows by induction from (a) and (b) in (2.6.9). The second follows from (b) and (c) in (2.6.9).

Lemma 2.6.1 *Define the constants* $\boldsymbol{M_n}$ *by* $\boldsymbol{M_n = \max_{0\leq x\leq 1} \|B_n(x)\|}$. *Then* $\boldsymbol{M_n \leq n!}$. (2.6.12)

Proof: Clearly true for $n=0$. For $n>0$, (c) in (2.6.9) shows that B_n has a zero at some $0 < x_n < 1$. Hence from (b) in (2.6.9) $B_n(x) = n\int_{x_n}^{x} B_{n-1}(s)\,ds$, thus $|B_n(x)| \leq n\,M_{n-1}\,|x-x_n| \leq n\,M_{n-1}$. ♣

Definition 2.6.1 *The* **generating function for the Bernoulli polynomials** *is defined by*

$$G(x,t) = \sum_{n=0}^{\infty} \frac{1}{n!}\, t^n\, B_n(x). \tag{2.6.13}$$

Note that, from (2.6.12), the series defining G converges for all $0 \leq x \leq 1$ and $|t| < 1$ — in fact, it can be show that it converges for all x and all $|t| < 2\,\pi$.

Problem 2.6.1 **Compute the Bernoulli polynomials' generating function.**
Find an explicit expression for G in (2.6.13).

Hint: Use (b) in (2.6.9) to compute G_x and find a simple ode in x that G solves, for each t. Write the general solution to this ode — this solution will have a free "constant" — which is actually a function of t, say $f = f(t)$. Determine $f = f(t)$ using (a) and (c) in (2.6.9), which can be used to obtain the value of $\int_0^1 G(x,\,t)\,dx$.

The End.