1 The singular value decomposition

The singular value decomposition is one of the most important tools for the analysis and computation of the linear operators. It has two standard forms, one for the compact linear operators mapping from a Hilbert spaces to another both of infinite dimensions, the other for matrices in $\mathbb{C}^{m\times n}$ of finite dimensions. While both versions are equally important for conceptual purposes, the latter is readily implementable and will be the main subject of Section 1.1. The connections of the SVD to the least squares problems and ill-posed problems will be given in Section 1.2. Many other linear problems, continuous or discrete, finite or infinite dimensions, can be better understood if we first ask the question: what if we take the SVD?

1.1 SVD and compression of linear operators

Theorem 1.1 *(***The singular value decomposition, reduced form***) Suppose that $A \in \mathbb{C}^{m\times n}$ is of rank k. Then there exist three matrices $U \in \mathbb{C}^{m\times k}$, $\Sigma \in \mathbb{C}^{k\times k}$, $V \in \mathbb{C}^{n\times k}$, such that*

$$A = U \cdot \Sigma \cdot V^* = \sum_{j=1}^{k} \sigma_j u_j v_j^* \tag{1}$$

where $\Sigma = diag\{\sigma_1, \ldots \sigma_k\}$ is a diagonal matrix with positive and decreasing diagonal entries, $U = [u_1, \ldots, u_k]$, $V = [v_1, \ldots, v_k]$ have orthonormal columns; namely,

$$(u_i, u_j) = \delta_{ij}, \quad or \;\; U^*U = I_m, \tag{2}$$

$$(v_i, v_j) = \delta_{ij}, \quad or \;\; V^*V = I_n. \tag{3}$$

Furthermore, $\{\sigma_j\}$ are referred to as the singular values, U, V as the left and right singular vectors.

Among the most striking powers of the SVD is the so-called low rank approximation to, or compression of, a linear operator embodied in the next theorem.

Theorem 1.2 *(***Best low rank approximation***) Under the assumptions of the preceding theorem, let*

$$A_l = \sum_{j=1}^{l} \sigma_j u_j v_j^*, \quad 1 \le l \le k. \tag{4}$$

Then

$$\|A - A_l\|_2 = \min_{rank(B)\le l} \|A - B\|_2 = \sigma_{l+1}, \tag{5}$$

where if $l = k$, σ_{k+1} is understood as zero.

Remark 1.1 The above low-rank approximation also holds for the Frobenius matrix norm, which is the standard Euclidean norm for a matrix viewed as an $m \cdot n$-vector; in other words,

$$\|A - A_l\|_F = \min_{\mathrm{rank}(B)\le l} \|A - B\|_F = \left(\sum_{j=l+1}^{k} \sigma_j^2 \right)^{\frac{1}{2}}. \tag{6}$$

Remark 1.2 To compress a matrix A of low rank k via the SVD means to construct the SVD for A and then keep the first k left and right singular vectors, as well as the first k singular values. Obviously, if k is much smaller than $\min(m,n)$ there are worth while savings. According to Theorem 1.2, however, substantial savings can also be realized when k is not small but a proportion of the first k nonzero singular values are small so that they can be dropped. The error of the approximation thus incurred is given by (5) if A is to be used as a linear operator, or given by (6) if A is to be used as a vector.

The matrix 2-norm and the orthogonality conditions on the singular vectors are associated with the standard inner product. In fact, the SVD is generically attached to the 2-norm as opposed to the 1-norm or maximum norm which are not induced by a inner product of a pair of elements from the same space. However, the inner product may be weighted. Indeed, in many applications, such as representing a function $f(x)$ at unequally spaced Chebyshev points for high precision, or evaluating a function at Gaussian points for accurate integration, the natural inner product which induces the standard Euclidean normal or the 2-norm is a weighted one. We therefore define the weighted inner product for $\mathbb{C}^n$ and $\mathbb{C}^m$ by the formulae

$$\prec x, y \succ =: \sum_{j=1}^{n} w_j \overline{x_j} y_j = x^* W y, \tag{7}$$

$$\prec \xi, \eta \succ =: \sum_{j=1}^{m} \lambda_j \overline{\xi_j} \eta_j = \xi^* \Lambda \eta, \tag{8}$$

where $W = \mathrm{diag}\{w_j\}$ and $\Lambda = \mathrm{diag}\{\lambda_j\}$ are positive diagonal matrices. There is a SVD for the weighted inner products.

Theorem 1.3 *(***The singular value decomposition, reduced and weighted form***) Suppose that $A \in \mathbb{C}^{m\times n}$ is of rank k. Suppose further that $W = diag\{w_j\}$ and $\Lambda = diag\{\lambda_j\}$ are positive diagonal matrices. Then there exist three matrices $U \in \mathbb{C}^{m\times k}$, $\Sigma \in \mathbb{C}^{k\times k}$, $V \in \mathbb{C}^{n\times k}$, such that*

$$A = U \cdot \Sigma \cdot V^* = \sum_{j=1}^{k} \sigma_j u_j v_j^* \tag{9}$$

where $\Sigma = diag\{\sigma_j\}$ with positive and decreasing diagonal; $U = [u_j]$, $V = [v_j]$ have orthonormal columns; namely,

$$\begin{aligned} \prec u_i, u_j \succ &= \delta_{ij}, \quad or \;\; U^* \Lambda U = I_m, & (10)\\ \prec v_i, v_j \succ &= \delta_{ij}, \quad or \;\; V^* W V = I_n. & (11)\end{aligned}$$

Furthermore, this factorization can be accomplished by the regular SVD on the matrix

$$\tilde{A} = \Lambda^{\frac{1}{2}} A \, W^{\frac{1}{2}}. \tag{12}$$

More precisely, let

$$\tilde{A} = \tilde{U} \cdot \tilde{\Sigma} \cdot \tilde{V}^* \tag{13}$$

be the regular SVD of $\tilde{A}$ with $\tilde{U}^* \tilde{U} = I_m$ and $\tilde{V}^* \tilde{V} = I_n$. Then

$$U = \Lambda^{-\frac{1}{2}} \tilde{U}, \;\; V = W^{-\frac{1}{2}} \tilde{V}, \;\; \Sigma = \tilde{\Sigma}. \tag{14}$$

Remark 1.3 Obviously, the above theorem holds when the positive diagonal weight matrices W, Λ are replaced by positive definite Hermitian matrices.

Remark 1.4 According to Theorem 1.3, the least-squares solution of the linear system $AWx = b$ is

$$x = V \Sigma^{-1} U^* \Lambda b \tag{15}$$

where x assumes the mininum W-norm among those vectors in $\mathbb{C}^n$ which minimize the residual $b - AWx$ in the Λ-norm.

Remark 1.5 Similarly, the mininum W-norm solution x, to the problem of minimizing the residual $b - Ax$ in the Λ-norm, is given by the formula

$$x = W^{-\frac{1}{2}} V \Sigma^{-1} U^* \Lambda b \tag{16}$$

where $A = U\Sigma V^*$ is the regular SVD for A.

Now we denote by $A : X \mapsto Y$ a compact linear operator mapping form a Hilbert space X to another Hilbert space Y. Then the adjoint operator $A^* : Y \mapsto X$ such that $(x, A^*y) = (Ax, y)$ is also compact. The square root of the eigenvalues of the A^*A, which are nonnegative, are referred to as the singular values of the map A. We will denote by $\{\sigma_j\}$ the nonincreasing sequence of all nonzero singular values of $A \neq 0$, and by $P : X \mapsto \mathcal{N}(A)$ the orthogonal projector to the null space. Theorem 1.1 is a special case of the next theorem.

Theorem 1.4 *(***The singular value decomposition for compact operator***) Suppose that A is a compact linear operator mapping form a Hilbert space X to another Hilbert space Y. Then there exist orthonormal sequences $\{v_j\}$ in X and $\{u_j\}$ in Y such that*

$$Av_j = \sigma_j u_j, \quad Au_j = \sigma_j v_j. \tag{17}$$

Give an arbitrary $x \in X$, it has the decomposition

$$x = \sum_j (v_j, x) v_j + Qx \tag{18}$$

and

$$Ax = \sum_j \sigma_j (v_j, x) u_j. \tag{19}$$

1.2 Least squares problems

The least squares problem divides itself into three groups, each of which could be regarded as an ill-posed (linear) problem whose unique solution is specified by minimizing the 2-norm of either the solution or its residual or both. We summarize their solution method in connection with the our analysis of the ill-posed problems.

Remark 1.6 The linear ill-posed problems which arise from the inverse problems, however, are remarkably different in their treatment from the least squares problems. The conditions for unique determination of the "best" solution are usually not to minimize the 2-norm. They are essentially not a part of linear algebra. Rather, these conditions are based on certain considerations, or maybe aesthetics, derived from the underlying physics, or from signal processing—after all, a large portion of the practical inverse scattering activities is aimed at reconstructing an image.

1.2.1 Tall matrix for over-determined linear system

Suppose that $A \in \mathbb{C}^{m\times n}$ is of rank k with $k = n < m$. Then given $b \in \mathbb{C}^m$ there exists a unique $x \in \mathbb{C}^n$ such that the residual $Ax - b$ is minimized. This is the classical least squares problem used for curve fitting. The standard reformulation of the linear system $Ax = b$ is the normal equation

$$A^*Ax = A^*b \tag{20}$$

which gives rise to the standard formula of least squares solution

$$x = (A^*A)^{-1}A^*b = A^+b \tag{21}$$

where A^+ is referred to as the psudo-inverse of A.

Remark 1.7 The normal equation is also widely known for its numerical instability, for the condition number of A^*A is the square of that of A. A more computational efficient and stable method for the least squares solution is the Householder QR. Column pivoting will make the QR method more stable with minimum increase of flops. The SVD is equally as stable but is much more expensive.

There is actually a twist to the story of normal equation in the case of ill-posed linear problems; see Remark 1.13 for details.

Remark 1.8 Again, suppose that $A \in \mathbb{C}^{m\times n}$ is full rank with $n < m$. The psudo-inverse of A can be expressed via its QR or SVD form. Let $A = QR$, and $A = U\Sigma V^*$ be the two factorizations where $Q,\ U \in \mathbb{C}^{m\times n}$, $V \in \mathbb{C}^{n\times n}$ have orthonormal columns; $R \in \mathbb{C}^{n\times n}$ is upper triangular, $\Sigma \in \mathbb{C}^{n\times n}$ is diagonal, and neither is singular. Then

$$A^+ = V\Sigma^{-1}U^* = R^{-1}Q^* \tag{22}$$

1.2.2 Flat matrix for under-determined linear system

Suppose that $A \in \mathbb{C}^{m\times n}$ is full rank with $m < n$. Then given $b \in \mathbb{C}^m$, linear system $Ax = b$ has solution but not unique. When the inverse scattering problem is discretized and solved iteratively, each linearization gives rise to a linear system of this type or one similar to it, as opposed to the type of the full-rank tall matrix. To pick a physically or "aesthetically" meaningful solution out of this linear (affine) subspace of solutions, a condition of minimum norm is usually place on the solutions to uniquely determine a solution provided that the norm is convex. If we further restrict our attention to norms induced by inner product, we see that it is generally of the form $(x, y) = x^*Wy$ for a positive definite Hermitian matrix W. Therefore, without loss of generality, we assume that $W = I$ in the sequel, well knowing that we are able to construct the Householder QR and the SVD with the weighted inner product. The unique least-norm, or least squares, solution can again be expressed via the psudo-inverse,

$$x = A^+b \tag{23}$$

where $A^+ = A^*(A\ A^*)^{-1}$ assumes the form

$$A^+ = V\Sigma^{-1}U^* = Q(R^*)^{-1}, \tag{24}$$

and $Q \cdot R = A^*$ is the standard QR decomposition for A^*.

Remark 1.9 When dealing with least squares problem $Ax = b$ in the setting of a Hilbert space $\mathcal{H}$ of (smooth) functions f and the Hilbert space $\mathcal{C}$ of their linear combination coefficients α, the vector x usually represents either f or α, namely either $x \in \mathcal{H}$ or $x \in \mathcal{C}$. For a flat, full rank matrix A, usually $x = f$; in this case, the columns of A and therefore of Q and U represent basis functions in $\mathcal{H}$. On the other hand, for a tall, full rank matrix A, usually $x = \alpha$; in this case, the rows of A and therefore the columns of Q and V represent basis functions in $\mathcal{C}$.

1.2.3 Rectangular matrix for rank deficient linear system

Suppose that $A \in \mathbb{C}^{m\times n}$ is full rank with $k < \min(m, n)$. Then the least squares (or least-norm) solution is defined as

$$x = A^+b = V\Sigma^+U^*,\ \ \Sigma^+ = \mathrm{diag}(\sigma_1^{-1}, \sigma_2^{-1}, \ldots, \sigma_k^{-1}). \tag{25}$$

It is easy to verify that this is the only vector in $\mathbb{C}^n$ with the least-norm that minimizes the norm of residual $Ax - b$ in $\mathbb{C}^m$. The norms of solution and residual may be chosen other than the standard Euclidean norms if the SVD used in (25) is constructed with weights (see Theorem 1.3).

Remark 1.10 The rank deficient linear system is better suited, than the full-rank flat linear system, to describe, and for the solution of, the ill-posed linear problems.

1.3 Ill-posed system and its regularization

An ill-posed linear problem, when properly discretized, results in a rectangular linear system with decaying singular values, and its regularization is a process of determining its numerical rank so as not to perform the unstable division by small singular values required in (25). For the continuous case $Au = f$ where $A : X \mapsto Y$ is compact, its regularization is understood as to approximate the unbounded A^{-1} by a family of bounded linear operator $R_\alpha : Y \mapsto X$ with the regularization parameter $\alpha \in (0, 1]$ such that for any u in the domain of A

$$u = \lim_{\alpha \to 0} R_\alpha \cdot Au. \tag{26}$$

Obviously, $\|R_\alpha\|$ blows up as α goes down to zero. Now suppose that A is injective (there is no loss of generality for otherwise X can be replaced by $X' =: X/\text{Null}(A))$, f is in $A(X)$ (the range of A) so that $Au = f$ has a unique solution. Suppose further that $f^\delta = f + \delta f$ is the contaminated right hand side for which we solve the equation

$$Au = f^\delta \tag{27}$$

with the regularization scheme R_α. The regularized solution $u_\alpha^\delta = R_\alpha f^\delta$ is expected to converge to the solution $u = A^{-1} f$ as $\alpha \to 0$ and as the level of contamination $\delta =: \|\delta f\|$ dies down. These two competing processes exhibit the following standard behavior

$$\|u_\alpha^\delta - u\| = \|R_\alpha f^\delta - R_\alpha f + R_\alpha(Au) - u\| \le \delta \|R_\alpha\| + \|R_\alpha Au - u\|. \tag{28}$$

Thus, on one hand, a shallow regularization (i.e., smaller α so that R_α better approximates A^{-1}) is required to reduce the second term of the estimate (28); on the other hand, this makes $\|R_\alpha\|$ in the first term grow. For a given level of noise δ, therefore, the level of regularization must not be so shallow that the first term is predominant; it should be such that the two terms are comparable. In short, a superior approximation of R_α to A^{-1} can be hazardous to the ill-posed problem. This phenomenon is frequently observed numerically.

Remark 1.11 This is not to advocate any lousy and cursory numerical treatment of the ill-posed problem; on the contrary, due to the intricate nature of ill-posedness and oscillatory nature of the scattering solution, high accuracies and superior approximations are needed to make the computation as reliable and stable as possible. The issue here is how to design regularization scheme R_α such that it approximates A^{-1} well for some of the modes in f^δ that are relevant, and it departs from A^{-1} for other modes that are irrelevant or noisy. This concept of separation of modes will be naturally realized and illustrated in the two of the following three examples of regularization scheme.

1.3.1 The spectrum cutoff

We arrange the singular values of the compact linear operator $A : X \mapsto Y$ (see Theorem 1.4)in a non-increasing order. They decay to zero because of the compactness. The spectrum cutoff is one of the most useful computational method of regularization. For a given threshold $\alpha \in (0, 1]$, designated as the relative error for the low-rank approximation (see Theorem 1.2), we set zero those singular values σ_j for which

$$\frac{\sigma_j}{\sigma_1} < \alpha. \tag{29}$$

This rank deficient problem is then solved by the standard least squares solver (25). We remark that the numerical rank k is necessarily finite, and that α is the level of approximation which should be chosen so as to balance the two types of error indicated in (28). Numerically, the linear operator A is approximated and its approximate singular values are processed. For a given level of noise in the right hand side, there are two reasons in favor of using a high accuracy approximation of A and its singular system. First, suppose that the noise resides mainly in the high-frequency modes, namely, in the left eigen-vectors u_j corresponding to small singular values. Then a good approximation of A will obviously produce an approximate solution which is accurate in the low-frequency modes, namely, in the right eigen-vectors v_j corresponding to singular values that are not small. The second reason is quite prevalent. Suppose in a computation our final answer is required to be accurate to three digits, a relatively low accuracy. Our question is that do we want to use functions (such as the sine and square root) from the Fortran or C library that are accurate to three or five digits to write our computer code. The answer is no because of reliability and stability concerns; such a code is difficult to debug, for example.

Remark 1.12 The spectrum cutoff is based on the SVD, and therefore the weights for SVD (see Theorem 1.3) specify the variety of the spectrum cutoff.

1.3.2 The Tikhonov regularization

Tikhonov regularization is useful for the purpose of analyzing the ill-posed problems because of its simplicity—we don't have to know the SVD in order to perform this regularization

$$R_\alpha =: (\alpha I + A^* A)^{-1} A^*, \quad \alpha \in (0, 1]. \tag{30}$$

In terms of the SVD, it is easy to verify that

$$R_\alpha = V(\alpha I + \Sigma^2)^{-1} \Sigma U^* \tag{31}$$

so that unlike the spectrum cutoff, Tikhonov regularization makes no discrimination—it perturbs every mode equally. Therefore, it is usually not recommended as a computational method. The Tikhonov regularization solution is the minimizer of the problem

$$\min_{u \in X} \|Au - f\|^2 + \alpha \|u\|^2 \tag{32}$$

where the norms are the 2-norm induced by the inner products in X and Y spaces. This formulation of the Tikhonov regularization makes it possible for generalization to nonlinear case.

1.3.3 The conjugate gradient method

The CG method as a least squares solver is well-known. Its application to the ill-posed problem is a natural extension of this approach, and has shown in many cases its superiority to other methods, including the spectrum cutoff. In the following, we consider the finite dimensional case where A is a matrix. For the sake of argument, let's suppose for a moment that A is Hermitian so that the CG method can be directly applied to solve the system $Ax = b$ without employing the normal equations which will make the system Hermitian. Let

$$\{ \lambda_j, \; u_j \} \tag{33}$$

be the eigen-system of A so that $Au_j = \lambda_j u_j$. Due to the ill-posedness, namely, a large condition number $\kappa(A) = |\max(\lambda_j)/\min(\lambda_j)|$, the rate of convergence is slow. However, as is well-known, the iterative solution $x^{(k)}$ of the the CG method converges fast in the (orthogonal) directions u_j corresponding to the dominant eigen-values λ_j; in a few iterations, these dominant modes will be built up in $x^{(k)}$ whereas it takes time for modes of small eigen-values to grow up and converge. In other words, ill-posedness or ill-conditioning separates naturally the two types of modes. If we start the CG iteration with a initial guess void of the second type of modes, usually high-frequency modes, and if we terminate the iteration at a time before these modes begin to grow, a natural way of separation and of regularization is achieved, see Remark 1.11. It remains that for a non-Hermitian system, the normal equation is solved iteratively via the CG method.

Remark 1.13 The normal equation, dreadful for its high condition number, is actually a preferred format for the solution of the linear ill-posed problem via the conjugate gradient method. Obviously, the normal equation $A^* A \cdot x = A^* b$, while squaring the condition number, magnifies the difference between the two types of modes, and makes the separation and thus the regularization easier.

2 Applied approximation theory

We summarize some of the classical approximation theory for a record and in conjunction with our numerical solution of the scattering problems. Approximation theory is the very heart of numerical analysis. It is hardly possible to talk about it without mentioning some of the well-known manuscripts; among them are texts by E. Isaacson and H.B. Keller [3], and by G. Dahlquist and A. Bjorck [4]. Also in general, whenever classical and practical matters are concerned, Abramowitz and Stegun [1] is one of the excellent sources for reference.

The essence of the solution process for the differential equations is integration—procedures marked usually by bounded or compact (linear) operators. Nowhere has this been made more manifest than in the reformulation of the differential equations as integral equations via the classical potential theory. The recent advent in the fast algorithms and relentless desire and necessity for solving the wave equations accurately and reliably qualify and promote the use of the integral equations. Their numerical treatment calls for the design of high-order quadratures for smooth and singular functions.

A discretized integral operator is more than just a matrix with merely its algebraic properties. Its rows and columns are usually tabulated values of certain functions and usually of analytical functions. This makes it flexible and natural to choose efficient representations of these functions for their better approximation. For example, we have almost complete freedom to choose the locations of the quadrature nodes discretizing the operator L_{ji} for the merging and splitting of scattering matrices. These points can be unequispaced; they may be chosen independently of the underlying quadrature nodes discretizing the Lippmann-Schwinger volume integral equation. Similarly, we should have equal flexibility in arranging the locations of the interpolation points for the approximation of these analytical functions. Our summary of the approximation theory will begin with interpolation and numerical integration.

2.1 Interpolation, Lagrange and Chebyshev

Everyone knows that the polynomial interpolation at the $n+1$ roots

$$z_j = \cos\left(\frac{j+1/2}{n+1}\pi\right), \quad j = 0, 1, \dots, n \tag{34}$$

of the Chebyshev polynomial T_{n+1} is a special case of the Lagrange interpolation, and that the error for the latter assumes the form

$$f(x) - p_n(x) = \frac{f^{n+1}(\eta)}{(n+1)!}(x-x_0)(x-x_1)\cdots(x-x_n) \tag{35}$$

for the interpolant polynomial p_n of degree n which interpolates f at the $n+1$ distinct nodes in an arbitrary interval $[a,b]$. Obviously, the quality of the approximation hangs on the balance of the two factors

$$a_{n+1}(x, x_0, \dots, x_n) = \frac{f^n(\eta)}{n!}, \quad \omega_{n+1}(x) = (x-x_0)(x-x_1)\cdots(x-x_n). \tag{36}$$

It turns out, the first factor is not as sensitive as the second to the locations of $x, x_0, \dots, x_n$, and for a reasonable distribution of the interpolation nodes $\{x_j\}$ covering $[a,b]$, $|a_{n+1}|$ is always close to its maximum value. Having thus prepared ourselves by accepting the worst case estimate

$$|a_{n+1}| \le \frac{\max_{\eta\in[a,b]} |f^{n+1}(\eta)|}{(n+1)!} =: M_n, \tag{37}$$

we proceed by making the maximum absolute value of the second factor $\omega_{n+1}(x)$, a monic polynomial of degree $n+1$, as small as possible. And as is well-known, the Chebyshev polynomial

$$T_{n+1}(z) = \cos(n \arccos(z)), \quad z \in [-1, 1], \tag{38}$$

with its maximum absolute value unity on $[-1,1]$ and with the leading coefficient 2^n, is the unique optimal choice; namely,

$$\omega_{n+1}(x) = 2^{-n}\left(\frac{2}{b-a}\right)^{-(n+1)} T_{n+1}\left(\frac{2}{b-a}x - \frac{b+a}{b-a}\right), \quad x \in [a,b]. \tag{39}$$

We thus arrive at the standard error estimate

$$E_n^{chb}(f) =: \|f - p_n\|_\infty \le 2M_n\left(\frac{b-a}{4}\right)^{n+1}, \tag{40}$$

for the polynomial interpolation at the (scaled and shifted) Chebyshev nodes in the interval $[a,b]$

$$x_j = \frac{b+a}{2} + \frac{b-a}{2}z_j, \quad j = 0,1,\ldots,n. \tag{41}$$

This Chebyshev interpolation is meant to approximate smooth functions in the maximum norm, and indeed its performance is superb. Denoting by $E_n(f)$ the error in the maximum norm for the best approximation to f with polynomials of degree n, we have

Theorem 2.1 (**M.J.D. Powell**) *For an anbitrary continuous function f on a finite interval $[a,b]$,*

$$\begin{aligned} E_n^{chb}(f) &< 4E_n(f), \quad n \le 20, & (42)\\ E_n^{chb}(f) &< 5E_n(f), \quad n \le 100, & (43)\end{aligned}$$

and asymptotically, for large n, $E_n^{chb}(f) \sim (2/\pi)\ln(n)E_n(f)$.

The hazards associated with the equispaced interpolation nodes, also known as the Runge phenomenon–oscillations near the end points–further justify the use of unequispaced points for interpolation. The two end points a and b are where the smooth function f terminates, and can therefore be viewed practically as points of "singularity" of f. It means that

(I) There must be a higher concentration of points near the two ends than around the center of $[a,b]$, as if the function f is singular there. The $n+1$ Chebyshev nodes (34) are separated by $D_e \sim h_\theta^2/2$ near the ends, and by $D_c \sim h_\theta$ near the center, with $h_\theta = \pi/(n+1)$. The ratio of the concentrations of the Chebyshev nodes at the end v.s. at the center is

$$\rho = D_c/D_e \sim n+1. \tag{44}$$

Were equispaced points used to achieve such a high concentration, the number of such points would be

$$m = 2/D_e \sim (2/\pi)^2(n+1)^2 \approx 0.4(n+1)^2. \tag{45}$$

(II) The distribution of Chebyshev nodes is optimal in the sense of minimizing $\|\omega_{n+1}\|_\infty$ in $[a,b]$. Fortunately and as expected from the nature of optimization, any reasonably small perturbation to the Chebyshev nodes—any similar distribution, mainly obtained as the roots of orthogonal polynomials, will give essentially the same high performance. The roots of the Legendre polynomial

$$P_n(x) = \frac{1}{2^n n!}\frac{d^n}{dx^n}(x^2-1)^2, \tag{46}$$

whose leading coefficient is $A_n = (2n)!/[2^n(n!)^2]$ and whose absolute value is bounded by unity in $[-1,1]$, is worthy of our attention owing to their dual purposes for interpolation and (Gaussian) quadrature. The scaled and shifted (see (41)) Legendre nodes in $[a,b]$ give rise to the estimate

$$\begin{aligned} E_n^{leg}(f) =: \|f-p_n\|_\infty &\le M_n A_{n+1}^{-1}\left(\frac{2}{b-a}\right)^{-(n+1)} \\ &\sim [\pi(n+2)]^{1/2}\cdot M_n\cdot\left(\frac{b-a}{4}\right)^{n+1}, \end{aligned} \tag{47}$$

where p_n is the interpolant polynomial of degree n; see (41) for a comparison.

(III) In fact, for $n+1$ equispaced points in $[-1,1]$ with spacing $h = 2/n$,

$$\max |\omega_{n+1}(x)| \sim \begin{cases} 2/e^n, & x \text{ around the center} \\ (2/e)^{n+1/2}, & x \text{ near the ends} \end{cases} \tag{48}$$

whereas for the $n+1$ Chebyshev nodes in $[-1,1]$, the maximum value is uniformly $1/2^n$ across the interval. Therefore, if equispaced points are use for interpolation, the error is actually smaller near the center than with the Chebyshev nodes; the error is comparable to that of the Chebyshev away from the two ends, again as if f is singular there. Only near the ends does the Runge phenomenon rear its ugly head, what a "singularity"!

(IV) On the other hand and to amplify the argument made above, for the class of smooth, periodic functions over $[a,b]$, every point is the center, and the equispaced points are the choice both for the polynomial and trigonometric polynomial interpolation. Furthermore, without the periodicity but with the knowledge of a smooth extension of f outside $[a,b]$, the end points becomes "regular" because they are not the end points of f in the extended interval. This situation arises in certain applications, and should not be overlooked. Whenever we have access to both sides of a point on $[a,b]$, say a, we can make it the center of interpolation for f by using points from around it as opposed to from one side of it. It might be tempting for us to extend a smooth f given only in $[a,b]$ by its interpolating polynomial. Many have trid, no doubt, this artificial extension; it is of no use, it is unstable, except for certain extremely restricted class of functions and with careful maneuver.

(V) If we really have to use m equispaced points in $[a,b]$ for interpolation, the Runge phenomenon can be rid of if we lower the degree of the approximating polynomial to be considerably less than m. Equivalently, if we insist on using a degree n polynomial for interpolation, we'll have to require m considerably greater than n so as to resolve the "singularities" at the ends. Impirically and not surprisingly, the condition for a good resolution is

$$m \geq 0.25n^2, \tag{49}$$

see (45) for a comparison. Whenever $m > n+1$, the standard linear problem for interpolation is over determined, and it is solved as a least squares problem; see Section 2.6 for details. The resulting polynomial of degree n approximates f with a quality comparable to that of the interpolation by a degree n polynomial at the n Chebyshev nodes.

We will point out a simple and practical method to determine the rate of convergence for the error estimate (35), (40), or (47) before we put this section to an end. For an analytic function f, the term M_n corresponds to the distance d from $[a,b]$ to the nearest singularity of f, possibly located on the complex plane. More specifically, let

$$f(z) = \sum_{n \geq 0} \frac{f^{(n)}(x_0)}{n!}(z - x_0)^n \tag{50}$$

be the Taylor expansion around $x_0 \in [a,b]$. Then its radius of convergence is obviously no less than d. Therefore, for any small number $\epsilon > 0$, the power series (51) converges absolutely in the disk $|z - x_0| \leq d - \epsilon$, and

$$\frac{|f^{(n)}(x_0)|}{n!}(d-\epsilon)^n \tag{51}$$

is bounded by a constant c dependent only on ϵ; in other words,

$$M_n \leq c(d-\epsilon)^{-(n+1)}. \tag{52}$$

If f is generated by ensembles of monopoles, dipoles, or quadrupoles, corresponding to $\ln(r)$, $1/r$, and $1/r^2$ singularities, is it easy to check that

$$M_n \leq c \cdot (n+1)^{\lambda} d^{-(n+1)}, \quad \lambda = -1, 0, 1 \text{ for monopole, dipole, quadrupole.} \tag{53}$$

In this case, which is most relevant to the scattering problem, the interpolation error for the Chebyshev and Legendre notes are

$$E_n^{chb}(f) \leq 2c \cdot (n+1)^{\lambda} \left(\frac{b-a}{4d}\right)^{n+1}, \tag{54}$$

$$E_n^{leg}(f) \sim c\sqrt{\pi}(n+2)^{\lambda+1/2} \left(\frac{b-a}{4d}\right)^{n+1}. \tag{55}$$

With the preceding formulae, we can estimate when the interpolation converges as $n \to \infty$ and the rate of convergence. If the interpolation diverges, or it converges slowly, the procedure on $[a, b]$ should be divided into several sub-intervals according to (54) and (55); this is known as the composite interpolation.

Remark 2.1 In principle, there is the problem of interpolation of singular functions, based on polynomial, rational, or other functions. Fortunately, such a problem rarely arises in the classical potential theory for the scattering problem. On the other hand, the discretization of the integral equations via the Nystrom method requires nothing other than quadratures for singular functions, mainly of the logarithmic ($\ln(r)$), coulombic ($1/r$) types. In the case of solving scattering problem in regions with corners, such as square or triangular subdomains where the scatterer is smooth inside and jumps to zero outside, singularities of the type r^{α} will have to be addressed. Happily, the numerical integration of singular functions is not more expensive or difficult than that of the smooth functions. These will be the topics of the next section.

2.2 Gaussian quadrature, maximum degree of precision

One of the fundamental facts from approximation theory is that numerical integration is a lot easier than interpolation. The Gaussian quadrature is to replace the integral

$$I(f) = \int_a^b \omega(x) f(x) dx \tag{56}$$

where ω is a non-trivial non-negative weight, with the n-term sum

$$G_n(f) = \frac{b-a}{2} \sum_{j=1}^{n} w_j f(x_j). \tag{57}$$

Remark 2.2 The Gaussian quadrature attains the maximum degree of precision in the sense that it is the only precise formula integrating f as polynomials of degree up to $2n-1$. Together with the trapezoidal rule (see (62)) and quadratures of Gaussian type (e.g., Radau's and Lobatto's, etc., see [1], page 888), it is representative of a class of quadratures–linear functionals–that are bounded, and therefore stable, in the maximum norm. This is a direct consequence of the fact that their weights w_j are all positive (non-negative is sufficient), and they integrate $f(x) = 1$ exactly; therefore, the maximum norm of these functionals is

$$\|G_n\|_{\infty} = \|T_n\|_{\infty} = b - a. \tag{58}$$

The Gaussian quadrature may be viewed as interpolation based. None of the other interpolation based quadratures on the equispaced points is stable because they develop large weights with alternating signs as n increases. It is advised that these quadratures not be used for n substantially greater than 7.

For $\omega(x) \equiv 1$, $x_j = y_j \cdot (b-a)/2 + (b+a)/2$ with y_j the roots of the Legendre polynomial P_n; the weights are $w_j = 2/(1-y_j^2)[P_n'(y_j)]^2$. The error of the approximation is

$$I(f) - G_n(f) = \frac{f^{2n}(\eta)}{(2n)!} \cdot \frac{(b-a)^{2n+1}(n!)^4}{(2n+1)[(2n)!]^2} \tag{59}$$

Assuming that f is generated by distributions of ensembles of monopoles, dipoles, or quadrupoles which are separated from $[a, b]$ by a distance $d > 0$, we combine (59) and (53) to obtain

$$|I(f) - G_n(f)| \sim c \cdot \pi \cdot \frac{b-a}{2} \cdot (2n+1)^{\lambda} \left(\frac{b-a}{4d}\right)^{2n}, \tag{60}$$

with $\lambda = -1, 0, 1$ for monopole, dipole, and quadrupole cases respectively. If the quadrature diverges, or it converges slowly, the procedure on $[a, b]$ should be divided into several sub-intervals according to (60); this is known as the composite quadratures.

Remark 2.3 Following up our theme of "singularity" of f at the two ends of the interval $[a, b]$ (see Section 2.1), we observe that this is also the case for numerical integration–higher concentration of points are required to remove the "singular" behavior at the ends although the functions ω and f are perfectly smooth there. On the other hand, the Gaussian quadrature nodes for the singular weight function

$$\omega(x) = (1 - x^2)^{-\frac{1}{2}} \tag{61}$$

are the Chebyshev points, which are essentially the same in terms of point concentration as the Legendre points for $\omega(x) \equiv 1$. We may be sure, therefore, that the "singularity" at the two ends is at least as strong as $1/\sqrt{r}$. We may further expect that numerical integration for singular functions via quadrature of Gaussian type is essentially not more difficult than for the smooth functions.

2.3 Trapezoidal rule, Euler-Maclaurin summation formula

If there ever is a Gaussian quadrature for the periodic case, it is the trapezoidal rule

$$T_n(f) = h\frac{f(a) + f(b)}{2} + h\sum_{j=1}^{n-1} f(a + jh), \quad h = (b-a)/n. \tag{62}$$

Since there are no "singular" end points on a torus, an equispaced distribution of quadrature nodes is the choice for the periodic class. The error analysis for the trapezoidal quadrature is a consequence of the well-known result

Theorem 2.2 *(***Euler-Maclaurin summation formula***) Suppose that a, b are a pair of real numbers such that $a < b$, and that $m \geq 1$ is an integer. Further, let B_k denote the Bernoulli numbers*

$$B_2 = \frac{1}{6}, \;\; B_4 = -\frac{1}{30}, \;\; B_6 = \frac{1}{42}, \;\; \ldots, \;\; B_{2m} \sim (-1)^{m+1}\frac{2(2m)!}{(2\pi)^{2m}}. \tag{63}$$

If $f \in c^{2m+2}[a, b]$, then there exists a real number ξ, with $a < \xi < b$, such that

$$\int_a^b f(x)dx = T_n(f) - \sum_{l=1}^{m} \frac{h^{2l}B_{2l}}{(2l)!}(f^{(2l-1)}(b) - f^{(2l-1)}(a)) - (b-a)B_{2m+2}\frac{f^{(2m+2)}(\xi)}{(2m+2)!}h^{2m+2}. \tag{64}$$

We shall divide the summation formula into three cases, and examine them.

(I) If f is periodic so that all the odd order derivatives up to $2m-1$ match and cancel in the summation formula, then for fixed m, the error of the trapezoidal rule drop like $n^{-2(m+1)}$ as n increases, namely, the order of trapezoidal rule is equal to the smoothness of the periodic function f. If f is periodic and in C^{∞} on the real line, the rate of convergence is beyond any order–it is super algebraic. We omit the error estimate here in favor of a stronger result to be presented below, see (67) and (70).

(II) If f is periodic and smooth, then for fixed n the summation formula as an asymptotic expansion may converge as m goes to infinity. Let us look at a simply case by assuming that $f(x) = e^{ikx}$, with

k integer, in $[0, 2\pi]$ (here we really mean the real or imaginary part of the function in order to use the summation formula). As $m \to \infty$, the error

$$(b-a)B_{2m+2}\frac{f^{(2m+2)}(\xi)}{(2m+2)!}h^{2m+2} \sim 4\pi \cdot e^{ik\xi}\left(\frac{k}{n}\right)^{(2m+2)} \tag{65}$$

decays to zero if $n > |k|$. In other words, the trapezoidal rule is exact for trigonometric polynomials of degree k if $n > k$, or equivalently, if the sampling interval $h = 2\pi/n$ is less than the wavelength $\lambda = 2\pi/k$.

Remark 2.4 Roughly speaking this is *one* point per wavelength for *integration*. As for *interpolation* (or trigonometric polynomial interpolation, to be more specific; see Section 2.5 for more details) there is the famous rule of *two* points per wavelength at the minimum. It seems that in order to achieve comparable performance with optimal methods, the number of samples required for interpolation is twice as many as that for integration; also compare (54) against (60) for the case of polynomial based interpolation and integration.

Now, a general smooth periodic function f can always be split into two parts

$$f(x) = t_k(x) + \delta_k(x), \tag{66}$$

where t_k is a trigonometric polynomial of degree k, for example, consisting of the first k Fourier modes of f; δ_k is the remainder that is small. Obviously, the error of the trapezoidal rule is bounded by

$$|I(f) - T_n(f)| \le \|T_n\|_\infty \|\delta_k\|_\infty = (b-a)\|\delta_k\|_\infty. \tag{67}$$

for any $n > k$. Finally, suppose f is periodic and smooth on the real line and analytically extendible to the complex plane (e.g., $e^{ikr\cos(\theta)}$ as a function of θ). Suppose further that the nearest singularity is separated from the real line by $d > 0$; namely, f is analytic in the strip between the parallel lines $z = \pm id$. Then it is well-known (see, e.g., [5]) that the Fourier series

$$f(z) = \sum_m a_m e^{imz} \tag{68}$$

converges absolutely in the strip. In particular, at $z = -ib$ with $0 < b < d$, $|a_m e^{imz}| = |a_m|e^{mb}$ must go to zero as $m \to \infty$ and therefore is bounded. We assume, without real loss of generality, that $|a_m|e^{md}$ is bounded. It follows that

$$|a_m| = O(e^{-md}). \tag{69}$$

Let $\delta_k(x) = \sum_{|m|>k} a_m e^{imx}$ in $f(x) = t_k(x) + \delta_k(x)$; obviously δ_k is bounded by a constant multiple of e^{-md}. It follows from (67) that trapezoidal rule converges exponentially

$$|I(f) - T_n(f)| = O(e^{-nd}). \tag{70}$$

Of course, if $f(z)$ is an entire function whose restriction on the real line is periodic, such as $e^{ikr\cos(\theta)}$, $\theta \in \mathbb{R}^1$ (this is the plane wave $\phi_0(x_1, x_2) = e^{ikx_1}$ restricted on a circle of radius r), the trapezoidal rule converges faster than any exponential; shall we call it super exponential?

(III) For the non-periodic case, the trapezoidal rule is second order accurate, and the error is strictly from the end points, see (64). The corrected trapezoidal rule is a method which compensates this "singular" behavior by simply incorporating the terms, in the summation formula and involving the odd order derivatives of f at the two ends, into the trapezoidal rule. What is actually added is of course their numerical approximation via standard finite difference formulae for the derivatives, using the equispaced points of the trapezoidal rule. A more systematic treatment can be found, for example, in [6] where the procedure is thoroughly tested and fully understood.

2.4 Quadratures for singular functions

Unlike interpolation, numerical integration for singular functions is actually needed, especially for the numerical solution of integral equations where the form of singularity is often $\ln(r)$, r^α, $\alpha \in [-1, 1)$ for two dimensions. Furthermore, the numerical integration of singular functions is not more expensive or difficult than that of the smooth functions; the "singularity" for smooth functions at the two ends of the interval (1-D), or on the boundary of a square (2-D), is already no weaker than $\ln(r)$ or r^α, see Remark 2.3. The design of the quadratures is technical but the principles and considerations are classical. These are the latest results, the efforts are still being made, and new results are still forthcoming. The trapezoidal rule based and the Gaussian formula based quadratures have been investigated, and designed; numerical tables of the quadrature nodes (for Gaussian based only) and weights are available. See [6], [7], and [8] for details.

2.5 Trigonometric polynomial interpolation

Trigonometric polynomial interpolation is for smooth periodic functions and can be viewed as a special case of polynomial interpolation when the is the trigonometric polynomials are replaced by their counterpart of complex exponentials. When the interpolation points are equispaced, the interpolation procedure is FFT which computes the trapezoidal rule for $f(x)e^{imx}$. Because of this product of two periodic functions $f(x)$ and e^{imx} where m is bounded by k (see (66), the number of points per wavelength should be 2 in order for the trapezoidal rule to converge to the integral of $t_k(x)e^{imx}$. Therefore, if k is the highest frequency of the significant Fourier modes of f, namely, if t_k is the dominant part of $f = t_k(x) + \delta_k(x)$, the number of equispaced interpolation points n must be no less than $2k$. When this happens, the interpolation error is given by (67); in other words, it is proportional to $\|\delta_k\|_\infty$. If more points are used ($n > 2k$), the convergence is super-algebraic; see (I) of Section 2.3. If fewer points are used, the computed trigonometric polynomial still interpolates f at the n points, but elsewhere the error is not small. The trigonometric is the same as the discrete least squares approximation (see Section 2.6) based on the orthogonal family e^{imx} where the Fourier coefficients are computed. It is very tempting to think that when the problem is under-resolved ($n < 2k$), we should be able to obtain the low-frequency Fourier coefficients accurately—only the high-frequency ones are in question. It could not be farther from truth: when there is no convergence, no Fourier coefficient is computed accurately. This is due to the so-called "aliasing": e^{imx} is exactly the same on the n equispaced points as e^{ilx} when $l \equiv m$, $\mathrm{mod}(n)$. Thus, an under-resolved Fourier mode e^{ilx} with $2l > n$ will contaminate the mode e^{imx} with $2m < n$ which by itself is indeed well resolved. If f is periodic on the real line and analytic on the strip between the parallel lines $z = \pm id$ in the complex plane, the error estimate for the interpolation with n equispaced points is

$$|f - p_n| = O(e^{-nd/2}), \tag{71}$$

indicating the fact that effort must be doubled here to achieve a comparable accuracy with that of numerical integration; see Remark 2.4.

2.6 Least squares approximation

The least squares approximation of a function f is to find coefficients c_m so as to minimize the residual

$$r_n = \|f(x) - \sum_{j=0}^{n} c_j B_j(x)\|_2 \tag{72}$$

for a set of prescribed basis functions. When discretized this form of approximation offers a greater flexibility in the choice of points to sample f, and a more rapid convergence, than the interpolation format. When over-determined (the number of points greater than the number of basis functions), the interpolation format is a special case of discretized least squares approximation; whereas when discretized, the least squares format is intended to be solved as over-determined linear system although it may be advantageous computationally to match the number of points with the number of basis functions to create a square linear system. Furthermore, with orthonormal basis $B_j(x)$, the least squares approximation is simply the truncated (generalized) Fourier expansion of f. Among the most popular choices for orthonormal basis functions are complex exponentials and spherical harmonics, Legendre, Chebyshev, and other orthogonal polynomials, Bessel families, corresponding to various weighted inner products for the 2-norm (72), for example,

$$(f,g) = \int_a^b \omega(x) f(x) g(x) dx, \quad c_j = (B_j, f) \tag{73}$$

Equispaced, Legendre, and Chebyshev points are used to discretize the integral (73) for complex exponentials ($\omega = 1$), Legendre ($\omega = 1$), and Chebyshev ($\omega = (1-x^2)^{-1/2}$) basis functions. In the case of equispace nodes for the complex exponentials as basis functions, this discrete problem is identical to the trigonometric polynomial interpolation, and thus is the FFT. Its error is therefore privided by (71), namely, the square root of the error (71) for numerical intergration. Similarly, the errors of the least squares approximation for the Legendre and Chebyshev cases are essentially the same as (71), except that the exponent $2n$ is replaced by n; see Remark 2.4.

References

[1] M. Abramowitz and I. Stegun (1965), *Handbook of Mathematical Functions,* Dover, New York.

[2] D. Colton and R. Kress (1992), *Integral Equation Methods in Scattering Theory,* Applied Mathematical Science vol. 93, Springer, Berlin.

[3] E. Isaacson, H.B. Keller (1966), *Analysis of numerical methods,* New York, Wiley

[4] G. Dahlquist, A. Bjorck (1974) *Numerical methods,* Prentice Hall, Inc., Englewood Cliffs, N.J.

[5] T. W. Korner (1987) *Fourier analysis,* Cambridge

[6] S. Kapur, V. Rokhlin (1997) *High-Order Corrected Trapezoidal Quadrature Rules for Singular Functions, SIAM Journal of Numerical Analysis*, v. 34, No. 4, pp. 1331-1356, 1997.

[7] J. Ma, V. Rokhlin, and S. Wandzura (1996) *Generalized Gaussian Quadrature Rules for Systems of Arbitrary Functions, SIAM Journal of Numerical Analysis*, v. 33, No. 3, pp. 971-996, 1996.

[8] N. Yarvin, V. Rokhlin (1996) *Generalized Gaussian Quadratures and Singular Value Decompositions of Integral Operators, Yale University Technical Report*, YALEU/DCS/RR-1109, 1996, to appear in *SIAM Journal of Scientific Computing.*