Numerical Analysis

Kexing Ying

May 15, 2020

Contents

1 Introduction

This course is an introduction to numerical analysis and is built on top of last term's linear algebra on which many concepts will reappear. This time however, we will mostly work in specific spaces rather than arbitrary inner product spaces. We will also consider issues of implementing algorithms and their ability to scale to large problems. This will be achieved by examining their efficiency, accuracy and stability. We will also consider typical numerical concepts such as iterations, conditioning, error analysis and operations count.

As a outline, we will first delve into numerical linear algebra, in which we will study orthogonalisation, least-squares problems, linear equations and factorisations. We will the move on to gradients and hessians, interpolations with orthogonal and non-orthogonal polynomials, Fourier series and lastly, numerical integration.

We will in this course mostly deal with the vector space $\mathbb{R}^n$ while unless mentioned otherwise, algorithms can easily extended to $\mathbb{C}^n$. We will use inner product and dot product interchangeably and define the outer product between two vectors $\mathbf{a}, \mathbf{b}$ to be the matrix $\mathbf{a}\mathbf{b}^T = A \in M_n(\mathbb{R})$ where $A_{ij} = a_i b_j$. We also note that the dot product on the reals form a bilinear form, and so all associated properties apply.

Lastly, we define the i-th standard basis $e_i \in \mathbb{R}^n$ as the vector with the j-th entry equal to the ij-Kronecker delta. That is, $[e_i]_j = 1$ is $i = j$ and 0 otherwise. This is a nice set of basis as it is orthonormal and given some vector $\mathbf{v}$, the dot product $\langle e_i, \mathbf{v} \rangle$ is the i-th entry of $\mathbf{v}$.

2 Numerical Linear Algebra

2.1 Orthogonalisation

We would like to find an orthogonal basis from a set of linearly independent vectors. As, simply by normalising the resulting vectors, we also obtain an orthonormal basis. From first year, we know that this can be achieved through the Gram-Schmidt process while we will also take a look at another method which utilises the Householder transformation. In both cases, the methods are related to the QR-decomposition of a square matrix.

Suppose we have a set of n linearly independent vectors $\{a_k\}_{k=1}^n$ in the m-dimensional space $\mathbb{R}^m$ where $n \leq m$. It is often advantageous to convert this set of vectors into an orthonormal basis such that the span of this basis is the same as the span of $\{a_k\}_{k=1}^n$.

To achieve this, we propose the first procedural method – the *classical Gram-Schmidt* (cGS) procedure.

Algorithm 1 (Classical Gram-Schmidt Procedure)**.** Given a set of n linearly independent vectors $\{\mathbf{a}_k\}_{k=1}^n$ in the m-dimensional space $\mathbb{R}^m$, we obtain an orthonormal basis of $\mathrm{sp}\{\mathbf{a}_k\}_{k=1}^n$.

1. Let $\mathbf{v}_1 := \mathbf{a}_1; \mathbf{q}_1 := \mathbf{v}_1/\|\mathbf{v}_1\|$. We call $\mathbf{v}_1$ the preliminary vector.
2. For $k = 2, \cdots, n$, let $\mathbf{v}_k := \mathbf{a}_k - \sum_{l=1}^{k-1}\langle \mathbf{a}_k, \mathbf{q}_l \rangle \mathbf{q}_l$; $\mathbf{q}_k := \mathbf{v}_k/\|\mathbf{v}_k\|$, that is, we define $\mathbf{v}_k$ such that it is orthogonal to all previous $\mathbf{q}_l$ while we normalise $\mathbf{v}_k$ resulting in $\mathbf{q}_k$.

Then, by the above procedure $\{\mathbf{q}_k\}_{k=1}^n$ is the required set of vectors.

Let us recall the proof for the correctness of the classical Gram-Schmidt procedure.

Proof. Normality and span (as $\mathbf{q}_k$ is a linear combination of $\mathbf{a}_i$ where $i \leq k$) is trivial so we shall show orthogonality.

We induction on n. For $n = 1$, orthogonality is trivial so let us assume the inductive hypothesis and suppose $n = i + 1$. By the inductive hypothesis, the Gram-Schmidt procedure will result us with $\{\mathbf{q}_k\}_{k=1}^i$ – an orthonormal basis of $\mathrm{sp}\{\mathbf{a}_k\}_{k=1}^i$, and so, it suffices to show,

$$\langle \mathbf{q}_j, \mathbf{v}_{i+1} \rangle = \left\langle \mathbf{q}_j, \mathbf{a}_{i+1} - \sum_{l=1}^{i} \langle \mathbf{a}_{i+1}, \mathbf{q}_l \rangle \mathbf{q}_l \right\rangle = 0,$$

for all $j = 1, \cdots, i$. But, this is true as,

$$\left\langle \mathbf{q}_j, \mathbf{a}_{i+1} - \sum_{l=1}^{i} \langle \mathbf{a}_{i+1}, \mathbf{q}_l \rangle \mathbf{q}_l \right\rangle = \langle \mathbf{q}_j, \mathbf{a}_{i+1} \rangle - \sum_{l=1}^{i} \langle \mathbf{a}_{i+1}, \mathbf{q}_l \rangle \langle \mathbf{q}_j, \mathbf{q}_l \rangle = \langle \mathbf{q}_j, \mathbf{a}_{i+1} \rangle - \langle \mathbf{q}_j, \mathbf{a}_{i+1} \rangle = 0,$$

where the second equality holds since by the inductive hypothesis $\langle q_j, q_l \rangle = \delta_{jl}$. $\square$

While we have proved the correctness of the algorithm, the question remains on whether or not we can always perform such an procedure. We see that, by induction, to see that the algorithm can always be performed, it suffices to show that there does not exists a case where $\mathbf{v}_2 \neq 0$ (we can then apply induction).

Suppose $\mathbf{v}_2 = 0$, then $\mathbf{a}_2 - \langle \mathbf{a}_2, \mathbf{q}_1 \rangle \mathbf{q}_1 = 0$. But this would mean $\mathbf{a}_2$ is a multiple of $\mathbf{q}_1$ which is in turn a multiple of $\mathbf{a}_1$, contradicting the linearly independent assumption. So, this cannot occur and so cGS can always be performed.

While the cGS is mathematically correct, when implementing the algorithm on computers, it is possible to find special cases where the cGS suffers from accuracy and stability. As we shall see on in an exercise, it is possible to construct a slightly different version of the Gram-Schmidt procedure – *modified Gram-Schmdit* (mGS) such that this is no longer a problem.

2.2 QR-Decomposition I

The QR-decomposition is a very important decomposition in numerical analysis and we shall, in fact, look at two methods of achieving the QR-decomposition, hence the name of this section. The QR-decomposition decomposes a matrix into the product of two matrices Q and R where Q is orthogonal and R is upper triangular.

Suppose we have the linearly independent sequence of vectors $\{\mathbf{a}_k\}_{k=1}^n$ and the orthonormal basis of this resulted from Gram-Schmidt $\{\mathbf{q}_k\}_{k=1}^n$ in $\mathbb{R}^m$. Then by defining

$$A := (\mathbf{a}_1, \mathbf{a}_2, \cdots, \mathbf{a}_n) \in \mathbb{R}^{m\times n},$$

and similarly,

$$Q := (\mathbf{q}_1, \mathbf{q}_2, \cdots, \mathbf{q}_n) \in \mathbb{R}^{m\times n},$$

we seek to establish the relation $R \in \mathbb{R}^{n\times n}$ such that $A = QR$. Indeed, by considering the classical Gram-Schmidt procedure, where

$$\mathbf{v}_k := \mathbf{a}_k - \sum_{l=1}^{k-1} \langle \mathbf{a}_k, \mathbf{q}_l \rangle \mathbf{q}_l,$$

we see that $\mathbf{a}_j$ is a linear combination of $\mathbf{q}_i$ where $j \le i$, and so it follows that R is upper triangular.

Suppose we denote the ij-th entry of R as r_{ij}, then $\mathbf{a}_k = \sum_{l=1}^k r_{lk}\mathbf{q}_l$, by the definition of matrix multiplication. Since $\mathbf{q}_1 = \mathbf{a}_1/\|\mathbf{a}_1\|$, we have $\mathbf{a}_1 = \|\mathbf{a}_1\|\mathbf{q}_1$, and so, $r_{11} = \|\mathbf{a}_1\|$. Similarly, for $\mathbf{a}_k$, we have

$$\mathbf{a}_k := \mathbf{v}_k + \sum_{l=1}^{k-1} \langle \mathbf{a}_k, \mathbf{q}_l \rangle \mathbf{q}_l,$$

where $\mathbf{v}_k = \|\mathbf{v}_k\|\mathbf{q}_k$, so,

$$\mathbf{a}_k := \|\mathbf{v}_k\|\mathbf{q}_k + \sum_{l=1}^{k-1} \langle \mathbf{a}_k, \mathbf{q}_l \rangle \mathbf{q}_l.$$

Thus, by comparing coefficients, we find $\|\mathbf{v}_k\| = r_{kk}$ and $\langle \mathbf{a}_k, \mathbf{q}_l \rangle = r_{lk}$ for $1 \le l < k$. With this, using Gram-Schmdit, we have found a method to decompose a matrix as a product of an orthogonal and an upper triangular matrix.

However, the question of why this decomposition is importan remains. Suppose we would like to solve the linear system $A\mathbf{x} = \mathbf{b}$ (where A has full rank). If we have a QR-decomposition on A, say $A = QR$, then the problem becomes $QR\mathbf{x} = \mathbf{b}$. As Q is orthogonal, $Q^TQ = I$ and so,

$$\mathbf{d} := Q^T\mathbf{b} = Q^TQR\mathbf{x} = R\mathbf{x}.$$

Now, as R is upper triangular, the linear system $R\mathbf{x} = \mathbf{d}$ becomes easy to solve by **backwards substitution**; that is, since R is upper triangular, we have

$$x_n = d_n / r_{nn},$$

and

$$x_k = \frac{1}{r_{kk}} \left(d_k - \sum_{i=k+1}^{n} r_{ki} x_i \right).$$

We note that we claimed Q is orthogonal throughout the argument. This is true as the column vectors are orthonormal, and hence, by the definition of matrix multiplication, we have

$$[Q^T Q]_{ij} = \langle \mathbf{q}_i, \mathbf{q}_j \rangle = \delta_{ij},$$

and so $Q^T Q = I$.

Indeed, the orthogonal matrices are a nice set of matrices and and the dot product and hence the norm is invariant under transformations by orthogonal matrices. Indeed,

$$\langle Q\mathbf{x}, Q\mathbf{y} \rangle = \mathbf{x}^T Q^T Q \mathbf{y} = \mathbf{x}^T \mathbf{y} = \langle \mathbf{x}, \mathbf{y} \rangle.$$

By thinking in Euclidean spaces, we see that these types of transformations are rotations and so, the transformations induced by an orthogonal matrix is often refereed as a rotation.