Numerics Methods and Useful Facts

Tyler Chen

Contents

1 Calculus

1.1 Gradient and Jacobian

For $f : \mathbb{R}^n \to \mathbb{R}$ we define the gradient as,

$$\nabla f = \left[\frac{\partial f}{\partial x_1}, \frac{\partial f}{\partial x_2}, \ldots, \frac{\partial f}{\partial x_n}\right]$$

For $f : \mathbb{R}^n \to \mathbb{R}^m$ we define the Jacobian as,

$$J_f = \begin{bmatrix} \nabla f_1 \\ \nabla f_2 \\ \vdots \\ \nabla f_m \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

Note that the best linear approximation to f at x_0 is given by,

$$f(x_0) + J_f x$$

1.2 Gradients of Matrix Vector Forms

This can be useful for minimizing/maximizing expressions involving matrix vector quantities.

$$\nabla b^T A x = A^T b$$
$$\nabla x^T A x = (A + A^T)x$$

1.3 Taylor Expansions

$$f(t+k, x+h) = f + kf_t + hf_x + \frac{k^2}{2} f_{tt} + khf_{tx} + \frac{h^2}{2} f_{xx} + \mathcal{O}(k^2 + h^2)$$

1.3.1 Computing Expansions in Mathematica

Compute Taylor expansion of $f(t+k, x+h)$ to d-th order.

```
Normal[Series[f[t + z k, x + z h], {z,0,d}]] /. {z->1}
```

This can be written into a function like,

```
F[n_,j_] := Normal[Series[f[t + z n k, x + z j h], {z,0,d}]] /. {z->1}
```

Then `F[n, j]` computes the Taylor expansion of $f(t+nk, x+jh)$ about (t,x). This is useful for compute local truncation errors. For instance, to compute the LTE of a second order centered difference approximation $f'(x) \approx (f(t+k,x) - f(t-k,x))/2k$ we set $d = 3$ and use,

```
FullSimplify[(F[1, 0] - F[-1, 0])/(2 k)]
```

This gives that the difference methods is like $f_t(t,x) + \mathcal{O}(k^2)$.

1.4 Newton's Method

Suppose we wish to solve $G(x) = 0$ for some $G : \mathbb{R}^m \to \mathbb{R}$. One standard way to do this is using Newton's method, which iteratively finds the root of the first order linear approximation to $G(x)$ at points near the solution.

That is, we iteratively solve,

$$G(x_k) + J_G(x_k)(x_{k+1} - x_k) = 0$$

Explicitly,

$$x_{k+1} = x_k - J_G(x_k)^{-1} G(x_k)$$

1.5 Richardson Extrapolation

Suppose $\varphi(h)$ approximates quantity u to $\mathcal{O}(h^n)$. That is,

$$\varphi(h) = u + ch^n + \mathcal{O}(h^{n+k})$$

Then,

$$\varphi\left(\frac{h}{t}\right) = u + c\left(\frac{h}{t}\right)^n + \mathcal{O}\left(\left(\frac{h}{t}\right)^{n+k}\right) = u + t^{-n}ch^n + \mathcal{O}\left(h^{n+k}\right)$$

Therefore,

$$t^n \varphi\left(\frac{h}{t}\right) = t^n u + ch^n + \mathcal{O}(h^{n+k})$$

Therefore,

$$\frac{t^n\varphi(h/t) - \varphi(h)}{t^n - 1} = \frac{(t^n - 1)u + \mathcal{O}(h^{n+k})}{t^n - 1} = u + \mathcal{O}(h^{n+k})$$

2 Basic Linear Algebra

2.1 Useful Inequalities

Triangle Inequality:

$$\|x+y\| \leq \|x\| + \|y\|$$

Reverse Triangle Inequality:

$$|\,\|x\| - \|y\|\,| \leq \|x-y\|$$

Hölder's Inequality:

$$\|fg\|_1 \leq \|f\|_p \|g\|_q \,, \qquad 1/p + 1/q = 1, \qquad p,q >\geq 1$$

Cauchy-Schwarz Inequality:

$$|\langle u,v\rangle|^2 \leq \langle u,u\rangle \langle v,v\rangle = \|u\| \, \|v\|$$

Note that this is the Hölder Inequality with $p = q = 2$.

2.2 Invertible Matrix Theorem

The following are equivalent:

- A is invertible
- Exists B such that $BA = AB = I$
- $\det(A) \neq 0$
- A has full rank
- The columns of A are linearly independent
- The null space of A is zero.
- A is surjective
- $Ax = 0$ implies $x = 0$

2.3 Similar Matrices

Definition: Two matrices A and B are similar if $A = XBX^{-1}$ for some X.

Why it is useful: The eigenvalues of similar matrices are the same.

3 Projectors

Definition: A matrix P is a projector if $P^2 = P$

If P is a projector then $I - P$ is a projector onto the null space of P.

Given any projector,

$$\text{range}(P) \cap \ker(P) = \{0\}, \qquad \text{range}(P) + \ker(P) = \mathbb{C}^m$$

Conversely, given any two subspaces S_1, S_2 of $\mathbb{C}^m$ satisfying, $S_1 \cup S_2 = \{0\}$ and $S_1 + S_2 = \mathbb{C}^m$, there is a projector P such that,

$$\text{range}(P) = S_1, \qquad \ker(P) = S_2$$

3.1 Orthogonal Projector

Definition: A projector is called orthogonal if its range and null space are orthogonal. Equivalently, if $P = P^*$.

In general $\|P\|_2 \geq 1$, and equality is attained if and only if P is orthogonal.

3.2 Constructing Projectors

Given a matrix A, the orthogonal projector onto the range of A is given by,

$$P_A = A(A^*A)^{-1}A^*$$

In the case that A has orthonormal columns, this reduces to $P_A = AA^*$

3.3 Gershgorin's Theorem

Note: TODO

4 Scalar Functions of Matrices

4.1 Matrix Norms

Definition: Given a matrix A, and vector norm $\|\cdot\|$, the induced matrix norm is defined as,

$$\|A\| = \sup_{u \neq 0} \frac{\|Au\|}{\|u\|} = \sup_{\|u\|=1} \|Au\|$$

Note that we could really use two different norms (one for the domain of A, and one for the range), but this is not common.

Equivalent definition:

$$\|A\| = \sup_{u,v \neq 0} \frac{\langle Au, v\rangle}{\|u\| \|v\|} = \sup_{\|u\|=\|v\|=1} \langle Au, v\rangle$$

If A is Hermitian,

$$\|A\| = \sup_{u \neq 0} \frac{\langle Au, u\rangle}{\|u\|^2} = \sup_{\|u\|=1} \langle Au, u\rangle$$

4.1.1 Inequalities

All norms are similar over finite dimensional vector spaces.

Give bounds.

For certain definitoin of matrix norm

$$\|AB\| \leq \|A\| \|B\|$$

4.1.2 Specific Properties

$$\|A\|_2 = \sigma_{\max}$$

$$\|A\|_F = \sqrt{\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_m^2}$$

4.2 Spectral Radius

Definition: Given a matrix A, the spectral radius is defined as,

$$\rho(A) = \max\{\lambda : \lambda \text{ is an eigenvalue of } A\}$$

The spectral radius is bounded above by any matrix norm. Equality with the 2-norm when A is Hermitian.

$$\rho(I - M^{-1}A) = \lim_{k\to\infty} \left\| (I - M^{-1}A)^k \right\|^{1/k}$$

4.3 Condition number

Definition: Given a matrix A, the condition number is defined as,

$$\kappa(A) = \frac{\|A\|}{\|A^{-1}\|}$$

We always have $\kappa(A) = \sigma_{\max}/\sigma_{\min}$, where $\sigma_{\max}$ and $\sigma_{\min}$ are the largest and smallest singular values.

4.4 Rayleigh Quotients

Definition: For a Hermitian matrix A and vector x, the Rayleigh quotient is defined as,

$$r(x) = \frac{x^* A x}{x^* x}$$

Why it is useful:

- Gives an estimate of eigenvalues.
  - If x is an eigenvector, then $r(x)$ is the corresponding eigenvalue.
  - Specifically, if q is an eigenvector, $r(x) - r(q) = \mathcal{O}(\|r - q\|)^2$ as $x \to q$. That is, the Rayleigh quotient is a quadratically accurate estimate to eigenvalues.
  - For any $z \in [\lambda_{\min}, \lambda_{\max}]$ there exists x such that $r(x) = z$.
- Eigenvectors are stationary points of $r(x)$. That is, $\nabla r(x) = 0$ when $Ax = r(x)x$.
- Can be used to estimate eigenvalues in inverse iteration (called Rayleigh quotient iteration)

5 Classification of Matrices

Matrices are assumed to be complex and unless specified otherwise.

5.1 Banded

Definition: A matrix A is banded with bandwidth $2k+1$ (semi-bandwidth k) if $a_{ij} = 0$ whenever $|i-j| > k$.

Properties:

$\Rightarrow$ Eigenvalues are diagonal entries
$\Rightarrow$ Inverse, product, and sum of upper triangular matrices are upper triangular
$\Rightarrow$ Can solve linear systems in $\mathcal{O}((2k+1)m)$ time

5.2 Upper Triangular

Definition: A matrix R is upper triangular if $r_{ij} = 0$ for $i > j$. If $r_{ii} = 0$ the matrix is called strictly upper triangular.

Properties:

$\Rightarrow$ Eigenvalues are diagonal entries
$\Rightarrow$ Inverse, product, and sum of upper triangular matrices are upper triangular
$\Rightarrow$ Can solve linear systems in $\mathcal{O}(m^2)$ time with back substitution

5.3 Unitary

Definition: A matrix U is unitary if $U^*U = UU^* = I$.

Properties:

$\Leftrightarrow$ Columns are an orthonormal basis for $\mathbb{C}^n$
$\Rightarrow$ $\|AU\|_2 = \|UA\|_2 = \|A\|_2$

5.4 Hermitian

Definition: A matrix A is Hermitian if $A^* = A$

Properties:

$\Rightarrow$ Real eigenvalues
$\Rightarrow$ Normal

5.5 Skew symmetric

Definition: A real matrix A is skew symmetric if $A^T = -A$

Properties:

$\Rightarrow$ pure imaginary eigenvalues
$\Rightarrow$ $I + A$ is invertible

5.6 Normal (Unitarily Diagonalizable)

Definition: A matrix A is normal if $A^*A = AA^*$

Properties:

$\Leftrightarrow$ Unitarily diagonalizable (similar to a diagonal matrix by unitary similarity transform)
$\Rightarrow$ Hermitian if all eigenvalues are real
$\Rightarrow$ $\|A\|_2 = \rho(A)$

5.7 Positive definite

5.8 Hermitian Positive definite

Definition: A Hermitian matrix A is positive definite if $v^*Av > 0$ for all v.

Properties:

$\Leftrightarrow$ All eigenvalues are positive
$\Leftrightarrow$ Has Cholesky factorization

5.9 Diagonalizable

Definition: A matrix A is diagonalizable if it is similar to a diagonal matrix

5.10 Toeplitz

Definition: A matrix A is Toeplitz if each diagonal is constant.

Properties:

$\Rightarrow$ Can solve linear systems in $\mathcal{O}(m^2)$ time
$\Rightarrow$ If A is tridiagonal, $y_j = \sin(kj\pi/(m+1))$ is an eigenvector for $k = 1, 2, \ldots, m$

6 Matrix Decompositions

6.1 SVD

Definition: For any matrix $A \in \mathbb{C}^m$, the singular value decomposition (SVD) is a decomposition,

$$A = U\Sigma V^* = \sum_{i=1}^{m} \sigma_i u_i v_i^*$$

- U unitary
- Σ diagonal, with real positive entries in non-increasing order
- V unitary

Existence: Always

Uniqueness: *Note:* double check Unique up to complex sign of columns of U and V

Computing:

Why it is useful:

- Gives geometric interpretation for linear transforms on $\mathbb{C}^n$
- Rank revealing
- Numerical stability of algorithms using SVD

6.1.1 Reduced SVD

If A is rank deficient some singular values will be zero. We can drop these singular values and the corresponding singular vectors.

Why it is useful:

- Saves storage compared to regular SVD

6.1.2 Rank Reduced SVD

We can always define a new matrix A_k by,

$$A_k = \sum_{i=1}^{k} \sigma_i u_i v_i^*$$

This gives the best rank-k approximation to A in the sense that when $\|\cdot\|$ is the 2-norm of Frobenius norm,

$$\|A - A_k\| \leq \inf\{\|A - B\| : B \text{ is rank } k\}$$

6.2 (P)LU

Definition:

Existence:

Uniqueness:

Computing:

Why it is useful:

-

Gaussian Elimination

6.2.1 Partial Pivoting

When is pivoting needed?

6.2.2 Cholesky

Definition:

Existence: If A is Hermitian positive definite

Uniqueness: Unique up to sign

Computing: Same as LU decomposition, except don't make L unit lower triangular.

Why it is useful:

- Save storage space vs. LU decomposition

6.3 QR

Definition:

Existence:

Uniqueness:

Computing:

Why it is useful:

-

6.4 Eigen

Definition:

Existence:

Uniqueness:

Computing:

Why it is useful:

-

6.5 Shur

6.6 Jordan Normal

7 The Eigenproblem

7.1 Direct Methods

Do we even have direct methods?

7.2 Power iteration

7.3 Simultaneous Power Iteration

8 Direct Methods for Linear Systems

8.1 QR

8.2 Gaussian Elimination

8.3 SVD

9 Iterative Methods for Linear Systems

9.1 Simple Iteration

Simple iteration can be described as,

$$x_{k+1} = x_k + M^{-1}r_k = x_k + M^{-1}(b - Ax_k) = M^{-1}b - (I - M^{-1}A)x_k$$

where M^{-1} is some matrix which approximates A^{-1}.

9.1.1 Algorithm

Pick M as one of,

$$M = \begin{cases} \operatorname{diag}(A) & \text{Jacobi Iteration} \\ \operatorname{tril}(A) & \text{GS Iteration} \\ \omega^{-1}\operatorname{diag}(A) - \operatorname{tril}(A, k = -1) & \text{SOR} \end{cases}$$

Iteratively, compute the residual $r_k = b - Ax_k$, solve the system $Mz_k = r_k$ for z_k, and update $x_{k+1} = x_k + z_k$

9.1.2 Convergence

Simple iteration converges if and only if $\rho(I - M^{-1}A) < 1$.

To prove "if" direction, use theorem on spectral radius as limit of matrix norms. To prove "only if" direction, look at largest eigenvector of $I - M^{-1}A$.

9.2 Multigrid Methods

Simple iteration converges really slowly for low frequency components. However, by adjusting the mesh size we can solve an approximation to the low frequency components much quicker.

9.3 Conjugate Gradient

At each step the A-norm of the error is minimized over successive Krylov spaces,

$$\mathcal{K}_k = \operatorname{span}\{r_0, Ar_0, \ldots, A^k r_0\}$$

9.4 When/why it is used

CG is the standard method for Hermitian positive definite systems $Ax = b$.

Lower storage and computation cost than GMRES

Good for use with PDE methods.

9.4.1 Algorithm

9.4.2 Convergence

In exact arithmetic CG will converge in at most m steps. In finite precision arithmetic, orthogonality of search directions is *not* maintained, so exact convtem[$\Leftrightarrow$] Has Cholesky factorization ergence may never be obtained.

9.5 GMRES

Minimizes 2-norm of the residual over successive Krylov spaces.

9.6 Other methods

10 Solving Least Squares

The linear least squares problem is,

$$\min_x \|b - Ax\|_2$$

This is solved when x solve the linear system (called the normal equations),

$$A^T Ax = A^T b$$

10.1 Derivations of Normal Equations

Using Projectors: We know that the image of x solving the least squares problem will be the orthogonal projection of b onto the range of A. That is,

$$Ax = A(A^*A)^{-1}A^*b$$

Multiplying both sides on the left by A^* yields the normal equations.

Using Calculus: Note that

$$\|b - Ax\| = (b - Ax)^*(b - Ax) = b^*b + -2b^*Ax + x^*(A^*A)x$$

Therefore, since A^*A is Hermitian, solving $\nabla \|b - Ax\| = 0$ gives $2A^*Ax - 2A^*b = 0$.

10.2 Solving Least Squares Numerically

11 Boundary Value Problems

How do we solve boundary value ODEs?

11.1 Error and Convergence

11.1.1 Local Truncation Error

Definition: The LTE of a method is defined by replacing U_j in the method with the true solution $u(x_j)$. The discrepancy is the local truncation error. Denoting the true solution evaluated on the mesh by $\hat{U}$ we have,

$$\tau = A\hat{U} - F$$

11.1.2 Global Error

Definition: The global error of a method is defined as $E = U - \hat{U}$.

11.1.3 Stability

Explicitly denoting the dependence of the equations on the mesh spacing h we have,

$$A^h E^h = -\tau^h$$

Therefore,

$$\left\|E^h\right\| = \left\|(A^h)^{-1}\tau^h\right\| \leq \left\|(A^h)^{-1}\right\| \left\|\tau^h\right\|$$

If $\left\|(A^h)^{-1}\right\|$ is bounded for h sufficiently small, then the global error will go to zero provided the LTE goes to zero.

Definition: A method is stable if $(A^h)^{-1}$ exists and is bounded in norm for all h sufficiently small.

11.1.4 Convergence

Definition: A method is said to be convergent if $\left\|E^h\right\| \to 0$ as $h \to 0$.

We have condition,

$$\text{consistency} + \text{stability} \implies \text{convergence}$$

11.2 Green's Functions

11.3 Laplacian

11.4 Finite Element Methods

12 Integrators and IVPs

How do we solve $u'(t) = f(t, u(t))$ given $u(0)$?

12.1 Runge-Kutta Methods

12.2 Linear Multistep Methods

A linear multistep method is a method of the form,

$$\sum_{j=0}^{r} \alpha_j U^{n+r} = k \sum_{j=1}^{r} \beta_j f(t_{n+r}, U^{n+r})$$

The local truncation error is,

$$\tau_n = \frac{1}{k}\left(\sum_{j=0}^{r} \alpha_j\right) u(t_n) + \sum_{q=1}^{\infty} k^{q-1}\left(\sum_{j=0}^{r}\left(\frac{1}{q!} j^q \alpha_j - \frac{1}{(q-1)!} j^{q-1}\beta_j\right)\right) u^{(q)}(t_n)$$

Therefore the method is consistent if,

$$\sum_{j=0}^{r} \alpha_j = 0, \qquad \sum_{j=0}^{r} j\alpha_j = \sum_{j=0}^{r} \beta_j$$

The method is p-th order accurate if,

$$\sum_{j=0}^{r}\left(j^q \alpha_j - q j^{q-1}\beta_j\right) = 0, \qquad q = 1, 2, \ldots, p$$

12.2.1 Characteristic Polynomials

The characteristic polynomials for a LMM are defined as,

$$\rho(\zeta) = \sum_{j=0}^{r} \alpha_j \zeta^j, \qquad \sigma(\zeta) = \sum_{j=0}^{r} \beta_j \zeta^j$$

12.3 Stability

12.4 Zero Stable

An r-step LMM is said to be zero-stable if the roots of the characteristic polynomial $\rho(\zeta)$ all have modulus at most one, and are simple if they have modulus one.

$$\text{consistency} + \text{zero-stability} \iff \text{convergence}$$

12.5 Absolute Stability

The region of absolute stability for a method is the values of $k\lambda$, if when applied to the test equation $u' = \lambda u$, the solution doesn't blow up. That is, $\{U^n\}_{n=0}^{\infty}$ is bounded.

Note: double check this

The region of absolute stability for a LMM is the set of points z for which $\pi(\zeta, z) = \rho(\zeta) - z\sigma(\zeta)$ satisfy the root condition.

12.5.1 Regions of Absolute Stability of Common Methods

Forward Euler : $\{z : |z + 1| \leq 1\}$

Backward Euler : $\{z : |z - 1| \geq 1\}$

Trapezoid : $\{z : \text{Re}(z) \leq 0$

Midpoint : $\{z : \text{Im}(z) \in (-1, 1)\}$

12.5.2 Plotting Regions of Stability

boundary locus method for LMM

contour method for one step methods

12.6 Stiff ODEs

A stable L stable etc

13 PDEs

13.1 Method of Lines

13.2 Von Neumann Analysis

1. Replace U_j^n with $g(\xi)^n e^{i\xi j\Delta x}$
2. Solve for $g(\xi)$ and compute $|g(\xi)|$
3. Method is stable if and only if for all ξ, $|g(\xi)| \leq 1 + \mathcal{O}(\Delta x)$

13.2.1 Von Neumann Analysis Using Mathematica

We illustrate how to apply Von Neumann Analysis to the Lax–Wendroff method using Mathematica.

Define our replacement,

```
U[n_, j_] := g[\[Xi]]^n Exp[I \[Xi] j h]
```

Now solve for $g(\xi)$ after replacement in the method.

```
gxi = FullSimplify[ Solve[U[1, 0] ==  U[0, 0] - a k (U[0, 1] - U[0,
   -1])/(2 h) + (a^2 k^2)/2 (U[0, -1] - 2 U[0, 0] + U[0, 1])/h^2, g[\[
   Xi]]]]
```

We would like to figure out what values of ν give $|g(\xi)| \leq 1$. To do this we will replace ak/h with ν and $h\xi$ with a continuous parameter t and plot in the complex plane. If the entire parametric plot is contained in the unit circle, then each $|g(\xi)|$ will have modulus at most one.

```
g[\[Nu]_, t_] :=  ReplaceRepeated[g[\[Xi]] /. gxi[[1]], {a -> (h \[Nu])
   /k, h \[Xi] -> t}]
```

```
Manipulate[ParametricPlot[ReIm[g[\[Nu], t]], {t, 0, 2 \[Pi]}, PlotRange
    -> {-1, 1}], {\[Nu], -2, 2}]
```

This reveals we need $|\nu| \leq 1$. This agrees with the CFL condition. We now prove this symbollically by,

```
FullSimplify[Conjugate[g[\[Nu], t]] g[\[Nu], t] <= 1, Assumptions -> {0
    <= t <= 2 \[Pi], -1 <= \[Nu] <= 1}]
```