Lecture notes on Numerical Analysis

Instructor: Jie Liu[1]

We will cover

(1) Iterative methods for linear system of equations.

(2) Newton and Quasi-Newton methods for nonlinear equation and system of nonlinear equations.

(3) Monte Carlo methods including Monte Carlo integration and Markov Chain Monte Carlo.

The notes are based on the following books

- (Chap 1) J. Stoer and R. Bulirsch, "Introduction to numeric analysis", Springer-Verlag.
- (Chap 2) T. Sauer, "Numerical Analysis", 2nd edition, Pearson. Available at NUS Co-op@Forum.
- (Chap 3) R. W. Shonkwiler and F. Mendivil, "Explorations in Monte Carlo methods", Springer (available online from NUS library).
- (Chap 3) Jun S. Liu, "Monte Carlo Strategies in Scientific Computating", Springer.

Office hour: Monday afternoon or by appointment.

Grading policy:

- 3 computer project and 3 homework: 5% each
- midterm 20%
- final 50%

[1]email: matlj@nus.edu.sg, Office number: 65162798, Office: S17-08-15

Contents

1 Iterative methods for linear systems

1.1 A quick review of matrix decomposition from linear algebra

For any square matrix $A \in \mathbb{R}^{n\times n}$, there are 3 frequently used decompositions which always exist:

- QR factorization. $A = QR$ where Q is an orthogonal matrix (i.e., $Q^\top Q = I = QQ^\top$) and R is an upper triangular matrix. Let $A = (a_1, ..., a_n)$ and $Q = (q_1, ..., q_n)$. Because R is upper triangular, $\text{span}\{a_1, ..., a_k\} = \text{span}\{q_1, ..., q_k\}$ where the later basis vectors are orthogonal to each other. Hence QR factorization is Gram-Schmidt orthogonalization.

- Singular value decomposition. $A = U\Sigma V^\top$ where U and V are orthogonal matrices, and $\Sigma = \text{diag}(\sigma_1, ..., \sigma_n)$ is a diagonal matrix with non-negative diagonal entries.

- Jordan form. $A = TJT^{-1}$ where J are blockwise-diagonal matrix $J = \begin{bmatrix} J_1 & & \\ & \ddots & \\ & & J_k \end{bmatrix}$, $J_i = \begin{bmatrix} \lambda_i & 1 & & \\ & \lambda_i & \ddots & \\ & & \ddots & 1 \\ & & & \lambda_i \end{bmatrix}$.

Since $A = U\Sigma V^\top$, $A^\top A = V\Sigma^2 V^\top$. The eigenvalues of $A^\top A$ are therefore σ_1^2, ..., σ_n^2. Hence $\sqrt{\rho(A^\top A)} = \max_i \sigma_i$.

If A is symmetric, then A can be diagonalized, i.e, there are orthogonal matrix Ω and diagonal matrix Σ so that $A = \Omega\Sigma\Omega^\top$.

If $A = U\Sigma V^\top$. Then $AA^\top = U\Sigma^2 U^\top$ and $A^\top A = V\Sigma^2 V^\top$ are the diagonalization of symmetric matrix $AA^\top$ and $A^\top A$.

Given any $A \in \mathbb{C}^{n\times n}$, there exist unitary matrices U and V (i.e., $UU^* = U^*U = I$), and diagonal matrix Σ so that

$$A = U\Sigma V^*. \tag{1}$$

Then one can check that

$$\begin{pmatrix} 0 & A^* \\ A & 0 \end{pmatrix} \begin{pmatrix} V & V \\ U & -U \end{pmatrix} = \begin{pmatrix} V & V \\ U & -U \end{pmatrix} \begin{pmatrix} \Sigma & 0 \\ 0 & -\Sigma \end{pmatrix}$$

which means that the singular values of A becomes the eigenvalues of $[0, A^*; A, 0]$.

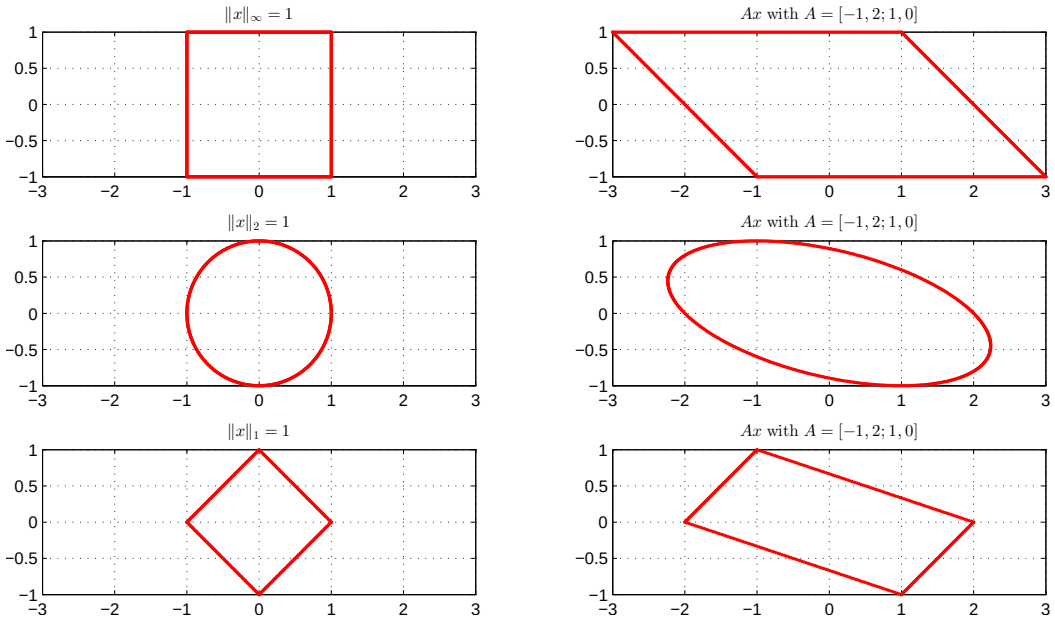

Figure 1: $A = [-1, 2; 1, 0]$. $A^\top A = [2, -2; -2, 4]$ with eigenvalues ≈ 0.7639 and 5.2361. $\|A\|_\infty = 3$. $\|A\|_2 \approx 2.2882$. $\|A\|_1 = 2$.

1.2 Matrix norm

Recall that a norm on a space X is a function $\|\cdot\|\colon X \to \mathbb{R}$ that assigns a real value to each vector and also satisfies the following three conditions

i) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.

ii) $\|x + y\| \leq \|x\| + \|y\|$.

iii) $\|\alpha x\| = |\alpha|\|x\|$.

When $X = \mathbb{R}^n$ or $\mathbb{C}^n$, we know the following function defines the p-norm on X

$$\|x\|_p = \|(x_1, ..., x_n)\|_p = \sqrt[p]{|x_1|^p + ... + |x_n|^p}.$$

When $X = \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$, we *want to* define the matrix norm in a way so that $\|Ax\| \leq \|A\|\|x\|$ for any $x \in \mathbb{R}^n$ or $\mathbb{C}^n$.

In order for this property to be true, we introduce the matrix norm induced by a vector norm:

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} = \sup_{\|x\|=1} \|Ax\|.$$

There are also matrix norms that satisfy *i*) to *iii*) yet they are not induced by a vector norm. For example, one can verify that $\|A\| = \mathrm{tr}(A^\top A) = \sqrt{\sum_{i,j} a_{ij}^2}$ is a matrix norm, but it is not induced by a vector norm.

The following is a list of the properties of the induced matrix norm.

(1) $\|Ax\| \leq \|A\|\|x\|$ for any x where $\|x\|$ and $\|Ax\|$ are vector norm and $\|A\|$ is the induced matrix norm. [Proof: by definition]

(2) $\|A\| = \sup_{\|x\|=1} \|Ax\|$. [Proof: Take $\alpha = \frac{1}{\|x\|}$ in the identity $\|\alpha Ax\| = |\alpha|\|Ax\|$ and note that the norm of $y = \frac{x}{\|x\|}$ equals 1.]

(3) $\|A\| \geq 0$ and $\|A\| = 0$ if and only if $A = 0$. [Proof: for any x, $\|Ax\| \leq \|A\|\|x\| = 0$ and therefore $Ax = 0$. So, $A = 0$. Note that we have used 0 to denote both number zero, zero vectors and zero matrices.]

(4) $\|A + B\| \leq \|A\| + \|B\|$. [Proof: It follows from $\|(A + B)x\| \leq \|Ax\| + \|Bx\|$.]

(5) $\|\alpha A\| = |\alpha|\|A\|$.

(6) $\|AB\| \leq \|A\|\|B\|$. [Proof: It follows from $\|(AB)x\| = \|A(Bx)\| \leq \|A\|\|Bx\| \leq \|A\|\|B\|\|x\|$ after using property (1) twice. So $\|AB\| = \sup_{x\neq 0} \frac{\|(AB)x\|}{\|x\|} \leq \|A\|\|B\|$.]

(7) $\|A\|_\infty = \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$. (max row-sum)

Proof of (7):

$$\begin{aligned}\|A\|_\infty &= \sup_{\|x\|_\infty=1} \|Ax\|_\infty = \sup_{\max_j |x_j|=1} \max_i |\sum_j a_{ij}x_j| \\ &\leq \sup_{\max_j |x_j|=1} \max_i \sum_j |a_{ij}||x_j| = \max_i \sum_j |a_{ij}|.\end{aligned}$$

So, we have proved $\|A\|_\infty \leq \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$ and are only left to prove $\|A\|_\infty \geq \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$. The idea is to choose a vector of the form $x = (\pm 1, \pm 1, ..., \pm 1)$ (Note that $\|x\|_\infty = 1$). So we assume $\max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right) = \sum_{j=1}^n |a_{kj}|$ and choose $x = (\text{sign}(a_{k,1}), \text{sign}(a_{k,2}), ..., \text{sign}(a_{k,n}))$.
Then $\|A\|_\infty = \sup_{\|x\|_\infty=1} \|Ax\|_\infty \geq \max_{1\leq i\leq n} \left(\sum_{j=1}^n |a_{ij}|\right)$. □

(8) $\|A\|_1 = \max_{1\leq j\leq n} (\sum_{i=1}^n |a_{ij}|)$. (max column-sum)

Proof of (8):

$$\begin{aligned}\|A\|_1 &= \sup_{\|x\|_1=1} \|Ax\|_1 = \sup_{\|x\|_1=1} \sum_i |\sum_j a_{ij}x_j| \le \sup_{\|x\|_1=1} \sum_i \sum_j |a_{ij}x_j| \\ &= \sup_{\|x\|_1=1} \sum_i \sum_j |a_{ij}||x_j| = \sup_{\|x\|_1=1} \sum_j \sum_i |a_{ij}||x_j| \\ &= \sup_{|x_1|+...+|x_n|=1} \sum_j |x_j| \left(\sum_i |a_{ij}| \right) \\ &\le \left(\max_j \sum_i |a_{ij}| \right) \sup_{|x_1|+...+|x_n|=1} \sum_j |x_j| = \max_j \sum_i |a_{ij}|.\end{aligned}$$

As long as we can also prove $\|A\|_1 \ge \max_j \sum_i |a_{ij}|$, we are done. That follows from taking $e = (0, ..., 0, 1, 0, ..., 0)^\top$ so that $\|e\|_1 = 1$ and Ax will select say, the $k-$th column vector of A whose 1-norm is assumed to be the largest among all the column vectors of A. So, $\|A\|_1 \ge \frac{\|Ae\|_1}{\|e\|_1} = \sum_i |a_{ik}| = \max_j \sum_i |a_{ij}|$. □

Now, introduce the spectral radius of a matrix:

$$\rho(A) = \max_{1 \le j \le n} |\lambda_j(A)|$$

which is the maximum of the absolute eigenvalues.

(9) $\|A\|_2 = \sqrt{\rho(A^\top A)}$.

Proof of (9): (Step 1. prove $\|A\|_2 \le \sqrt{\rho(A^\top A)}$.) Because $A^\top A$ is symmetric, it has n orthonormal eigen vector $u_1, ..., u_n$ with nonnegative eigen value $\lambda_1, ..., \lambda_n$. For any $x \in \mathbb{R}^n$ with $\|x\|_2 = 1$, we have $x = \sum_j \alpha_j u_j$ with $\|x\|_2^2 = x^\top x = \langle x, x \rangle = \sum_j \alpha_j^2 = 1$. So

$$\|Ax\|_2^2 = \langle Ax, Ax \rangle = \left\langle \sum_j A\alpha_j u_j, \sum_k A\alpha_k u_k \right\rangle \quad (2)$$

$$= \sum_{i,k} \alpha_j \alpha_k (u_j^\top A^\top A u_j) = \sum_j \alpha_j^2 \lambda_j \le \max_j \lambda_j. \quad (3)$$

(Step 2. prove $\|A\|_2 \ge \sqrt{\rho(A^\top A)}$.) Let $\lambda_s = \max_j \lambda_j$. Take $x = u_s$. Then $\|Ax\|_2^2 = \lambda_s$. □

Theorem 1 *If $\|A\| < 1$, then $(I - A)^{-1}$ exists and $\|(I - A)^{-1}\| \le \frac{1}{1-\|A\|}$*

Proof: For any y, consider the vector $x = \sum_{j=0}^\infty A^j y$ whose existence is guaranteed by $\|A\| < 1$. It is easy to verify that $(I - A)x = y$ and hence $(I - A)^{-1}$ exists [2]. Moreover,

[2]Recall that if an $n \times n$ matrix B satisfies "for any y, there is an x so that $Bx = y$", then B is invertible. The proof relies on choosing $y = e_i$ which is the i-th column of identity matrix I, and then finding x_i so that $Bx_i = y_i$. Let $C = [x_1, ..., x_n]$. Then $BC = [Bx_1, ..., Bx_n] = I$. Taking det on both sides, $\det(B)\det(C) = 1$. Hence $\det(B) \ne 0$ and B is invertible.

$\|(I-A)^{-1}y\| = \|x\| \le \frac{1}{1-\|A\|}\|y\|$ and hence

$$\|(I-A)^{-1}\| = \sup_{\|y\|=1} \|(I-A)^{-1}y\| \le \frac{1}{1-\|A\|}. \qquad \square$$

Theorem 2 *We have $\rho(A) \le \|A\|$ for any matrix norm induced by a vector norm. On the other hand, for any $\varepsilon > 0$, there is a matrix norm induced by a vector norm so that $\rho(A) > \|A\| - \varepsilon$.*

Proof: Let λ be the largest eigenvalue and u the associated eigenvector. Then

$$\|A\| = \sup_{x\neq 0} \frac{\|Ax\|}{\|x\|} \ge \frac{\|Au\|}{\|u\|} = |\lambda| = \rho(A).$$

(The rest of the proof is not required for the exam as it is too complicated.) On the other hand, let

$$TAT^{-1} = J = \text{diag}(C_1, ..., C_p)$$

where $C_j = \text{tridiag}(1, \lambda_j, 0)$ is the Jordan form. Let D_ε be an $n \times n$ diagonal matrix with diagonal entries being $[1, \varepsilon^{-1}, \varepsilon^{-2}, ..., \varepsilon^{1-n}]$. Then one can verify that

$$D_\varepsilon^{-1} J D_\varepsilon = \text{diag}(E_1, ..., E_p)$$

with $E_j = \text{tridiag}(\varepsilon, \lambda_j, 0)$. Hence $\|D_\varepsilon^{-1} J D_\varepsilon\|_\infty = \rho(A) + \varepsilon$. Define the vector norm

$$\|x\| = \|D_\varepsilon^{-1} T x\|_\infty.$$

Then

$$\begin{aligned} \|A\| &= \sup_{y\neq 0} \frac{\|Ay\|}{\|y\|} = \sup_{y\neq 0} \frac{\|D_\varepsilon^{-1}TAy\|_\infty}{\|D_\varepsilon^{-1}Ty\|_\infty} = \sup_{z\neq 0} \frac{\|D_\varepsilon^{-1}TAT^{-1}D_\varepsilon z\|_\infty}{\|z\|_\infty} \\ &= \|D_\varepsilon^{-1}TAT^{-1}D_\varepsilon\|_\infty = \rho(A) + \varepsilon \end{aligned}$$

where $z = D_\varepsilon^{-1} T y$. $\square$

Theorem 3 *The following three conditions are equivalent*

(1) $\lim_{k\to\infty} \|B\|^k = 0$ *for some matrix norm induced by a vector norm.*

(2) $\lim_{k\to\infty} \|B^k\| = 0$ *for any matrix norm induced by a vector norm.*

(3) $\rho(B) < 1$.

Proof: (1) $\Rightarrow$ (2): $\|B^k\| \le \|B\|^k \to 0$.

(2) $\Rightarrow$ (3): Let $\rho(B) = |\lambda|$, $Bu = \lambda u$. By (2), $\|B^k u\| = |\lambda|^k \|u\| \to 0$, hence $|\lambda| < 1$.

(3) $\Rightarrow$ (1): By Theorem 2, there is a vector norm induced matrix norm so that $\|B\| < \rho(B) + \varepsilon < 1$. $\square$

Remark: If (1) is true for some matrix norm, it may not true for another matrix norm as there is a matrix B so that $\|B\|_1 < 1 < \|B\|_\infty$ (see the homework). If (2) is true for some matrix norm, then it is true for any matrix norm as on the finite dimensional space $\mathbb{R}^{n\times n}$, any two norms are equivalent.

Theorem 4 $\lim_{k\to\infty} \|A^k\|^{1/k} = \rho(A)$.

Proof: $\rho(A)^k = \rho(A^k) \le \|A^k\|$, hence $\rho(A) \le \|A^k\|^{1/k}$. Let $B = \frac{A}{\rho(A)+\varepsilon}$. Then $\rho(B) < 1$ and hence $\|B^k\| \to 0$ by (2) of Theorem 3. Hence $\|B^k\| < 1$ when $k > K$ for some K, which can be rewritten as

$$\|A^k\|^{1/k} \le \rho(A) + \varepsilon. \qquad \square$$

1.2.1 Singular value decomposition and Frobenius norm

$$\|A\|_F = \left(\sum_{ij} |a_{ij}|^2\right)^{1/2} = \sqrt{\mathrm{tr}(A^\top A)}$$

Because $\mathrm{tr}(BC) = \mathrm{tr}(CB)$, it is easy to see that for any orthonormal matrix Ω, $\|A\|_F = \|\Omega A\|_F = \|A\Omega\|_F$. Actually, since $\|A\|_2 = \rho(A^\top A)$ and $\rho(\Omega B \Omega^\top) = \rho(B)$, we also have $\|A\|_2 = \|\Omega A\|_2 = \|A\Omega\|_2$.

Recall the singular value decomposition of $A \in \mathbb{R}^{m\times n}$: $A = U\Lambda V$ with $U \in \mathbb{R}^{m\times m}$, $V \in \mathbb{R}^{n\times n}$, $\Lambda \in \mathbb{R}^{m\times n}$, U and V are orthonormal and Λ is diagonal with nonegative entries $\{\sigma_1, ..., \sigma_p\}$ ($p = \min\{m, n\}$) which are called the singular value of A. Then $\|A\|_2 = \sqrt{\rho(A^\top A)}$ leads to $\|A\|_2 = \max_i \sigma_i$ and one can prove $\|A\|_F = \sqrt{\sigma_1^2 + ... + \sigma_p^2}$. For the proof, see "Numerical Linear Algebra" by Trefethen and Bau.

1.2.2 Condition number

The condition number of A with respect to p-norm is defined as

$$\kappa_p = \|A\|_p \|A^{-1}\|_p$$

Note that $\|A\|_p\|A^{-1}\|_p \ge \|AA^{-1}\|_p = 1$. Suppose we want to solve $Ax = b$ and suppose the input vector b has an error Δb and hence we are solving $A(x + \Delta x) = Ay = b + \Delta b$. Then $A\Delta x = \Delta b$ and because $\|b\|_p \le \|A\|_p\|x\|_p$

$$\frac{\|\Delta x\|_p}{\|x\|_p} \le \frac{\|A\|_p\|A^{-1}\Delta b\|_p}{\|b\|_p} \le \kappa(A)\frac{\|\Delta b\|_p}{\|b\|_p}$$

Remark: When A is symmetric, $A = \Omega^\top\Lambda\Omega$, where Ω is an orthogonal matrix and Λ is a diagonal matrix with diagonal entries being $[\lambda_1, ..., \lambda_n]$. Hence the condition number with respect to 2-norm can be written as

$$\kappa_2(A) = \sqrt{\rho(A^\top A)}\sqrt{\rho(A^{-\top}A^{-1})} = \frac{\max_j |\lambda_j|}{\min_j |\lambda_j|} \tag{4}$$

1.3 Jacobi and Gauss-Seidel iterations

Suppose we want to solve $Ax = b$ with $A = (a_{ij})$ and $b = (b_1, ..., b_n)^\top$.

Jacobi iteration
for $i = 1, ..., n$

$$x_i^{k+1} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^k - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) / a_{ii}$$

end

Let $A = L + D + U$. Then

$$Dx^{k+1} + (L + U)x^k = b. \tag{5}$$

Note that if x_k converges to y, then $Ay = b$ by (5). From (5), we have $x^{k+1} = -D^{-1}(L + U)x^k + D^{-1}b$. Let $e^k = x^k - x$, then $e^{k+1} = -D^{-1}(L+U)e^k$. So, Jacobi iteration converges if any only if $\lim_{k\to\infty} \left(-D^{-1}(L + U)\right)^k e^0 = 0$ for any e^0, which is equivalent to $\rho(D^{-1}(L+U)) < 1$ by Theorem 3.

Gauss-Seidel iteration
for $i = 1, ..., n$

$$x_i^{k+1} = \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) / a_{ii}$$

end

So, $(L + D)x^{k+1} + Ux^k = b$, $x^{k+1} = -(L + D)^{-1}Ux^k + (L + D)^{-1}b$. Gauss-Seidel iteration converges if any only if $\rho((L + D)^{-1}U) < 1$ by Theorem 3.

Theorem 5 (convergence of Jacobi and Gauss-Seidel iteration) *If A is strictly diagonal dominant* [3]*, then $\|G\|_\infty \le \|J\|_\infty < 1$ where $J = -D^{-1}(L+U)$ and $G = -(L+D)^{-1}U$.*

Proof: (Convergence of Jacobi) One can easily verify that

$$\kappa_J = \|J\|_\infty = \|D^{-1}(L + U)\|_\infty = \max_{1\le i\le n} \sum_{j=1,...n, j\neq i} \left| \frac{a_{ij}}{a_{ii}} \right| < 1. \tag{6}$$

(Convergence of Gauss-Seidel) Define $E = -D^{-1}L$ and $F = -D^{-1}U$, then $J = -D^{-1}(L+U) = -I^{-1}(D^{-1}L + D^{-1}U) = E + F$ and $G = -(L + D)^{-1}U = -(D^{-1}L + I)^{-1}D^{-1}U = (I - E)^{-1}F$.

Let's introduce the absolute value of a matrix: $|B| = (|b_{ij}|)$ if $B = (b_{ij})$. Let $e = (1, 1, ..., 1)^\top$ and note that $\|B\|_\infty = \| \, |B| \, e\|_\infty$.

The i-th component of $|J| \, e$ is $\sum_{j=1,...n, j\neq i} \left| \frac{a_{ij}}{a_{ii}} \right| \le \kappa_J$ by (6). Hence

$$|J| \, e \le \kappa_J e.$$

[3] which means for each i, $|a_{ii}| > \sum_{j,j\neq i} |a_{ij}|$

Because $|J| = |E| + |F|$,

$$|F|e \le (\kappa_J I - |E|)\, e.$$

Recall that $G = (I-E)^{-1}F$. Because E is strictly lower triangle matrix, $E^n = 0$, $(I-E)^{-1} = I + E + E^2 + ... + E^{n-1}$. Using this fact, we have

$$\left|(I-E)^{-1}\right| \le I + |E| + |E|^2 + ... + |E|^{n-1} = (I - |E|)^{-1} \tag{7}$$

because $|E|^n = 0$. Therefore

$$\begin{aligned} |G|e &\le (I - |E|)^{-1}|F|e \le (I - |E|)^{-1}(\kappa_J I - |E|)\, e \\ &\le \left(I + (\kappa_J - 1)(I - |E|)^{-1}\right) e = \left(I + (\kappa_J - 1)(I + |E| + |E|^2 + ... + |E|^{n-1})\right) e \\ &\le (I + (\kappa_J I - I))\, e = \kappa_J e. \end{aligned}$$

where we have used $\kappa_J - I \le 0$. This implies $\|G\|_\infty \le \kappa_J = \|J\|_\infty$. $\square$

1.4 Successive over-relaxation

for $i = 1, ..., n$

$$x_i^{k+1} = \omega \left(b_i - \sum_{j=1}^{i-1} a_{ij} x_j^{k+1} - \sum_{j=i+1}^{n} a_{ij} x_j^k \right) / a_{ii} + (1-\omega) x_i^k.$$

end

In matrix form, this can be written as

$$Dx^{k+1} = \omega \left(b - Lx^{k+1} - Ux^k \right) + (1-\omega) Dx^k.$$

So, $x^{k+1} = (D + \omega L)^{-1}(D - \omega(U + D)) + \omega(D + \omega L)^{-1} b$ which can be further written as

$$\begin{aligned} x^{k+1} &= (I + \omega D^{-1} L)^{-1} \left((1-\omega) I - \omega D^{-1} U \right) x^k + \omega (D + \omega L)^{-1} b \\ &:= R_\omega x^k + \tilde{R}_\omega b. \end{aligned} \tag{8}$$

By the way, because $x^{k+1} = x^k = x = A^{-1}b$ will always satisfy (8) for any b, R_ω and $\tilde{R}_\omega$ are related by $A^{-1} = R_\omega A^{-1} + \tilde{R}_\omega$, or $I = R_\omega + \tilde{R}_\omega A$.

Theorem 6 $\rho(R_\omega) \ge |\omega - 1|$

Proof: Let $\lambda_1, ..., \lambda_n$ be all the eigen value of R_ω. Then

$$\begin{aligned} &\prod_{j=1}^{n} |\lambda_j| = \left| \prod_{j=1}^{n} \lambda_j \right| = |\det R_\omega| \\ =& \left|\det(I + \omega D^{-1} L)^{-1}\right| \left|\det\left((1-\omega) I - \omega D^{-1} U\right)\right| \\ =& \frac{1}{|\det(I + \omega D^{-1} L)|} \left|\det\left((1-\omega) I - \omega D^{-1} U\right)\right| = \frac{1}{1^n} |1 - \omega|^n. \end{aligned} \tag{9}$$

In the very last step, we have used the fact that as $D^{-1}L$ is a strictly lower triangular, $(I + \omega D^{-1}L)$ is a lower triangle matrix with the diagonal entries all equal to 1. Hence $\det(I + \omega D^{-1}L) = 1^n$. For the same reason, $\det\left((1-\omega)I - \omega D^{-1}U\right) = (1-\omega)^n$. Therefore

$$\rho(R_\omega)^n = \left(\max_j |\lambda_j|\right)^n \geq \prod_{j=1}^{n} |\lambda_j| = |1-\omega|^n. \qquad \square$$

Theorem 7 *If A is symmetric positive definite, and $0 < \omega < 2$. Then the SOR method converges.*

For the proof, see Stoer and Bulirsch (theorem 8.3.7 in the 2nd edition).

1.5 Steepest descent and Conjugate gradient method

Now, we assume A is symmetric positive definite. The starting point in the derivation is to consider how we should minimize the function

$$f(x) = \frac{1}{2}x^\top Ax - x^\top b. \tag{10}$$

Because $D^2 f(x)$ is positive definite, by solving $\nabla f(x) = 0$, one can see that minimizing f and solving $Ax = b$ are equivalent problems.

1.5.1 Steepest descent

At any point x_c, the steepest descent direction $-\nabla f(x_c) = b - Ax_c$ is also the residue of solving $Ax = b$

$$r_c = b - Ax_c.$$

To minimize

$$f(x_c + \alpha r_c) = f(x_c) - \alpha r_c^\top r_c + \frac{1}{2}\alpha^2 r_c^\top A r_c.$$

$\alpha = r_c^\top r_c / r_c^\top A r_c$. This leads to steepest descent method:

1. Initialize with x_0

2. For $k = 0, 1, ...$ until some stopping criteria is satisfied

- $r_{k-1} = b - Ax_{k-1}$
- $\alpha_k = r_{k-1}^\top r_{k-1} / r_{k-1}^\top A r_{k-1}$
- $x_k = x_{k-1} + \alpha_k r_{k-1}$

Note that

$$r_{k-1}^\top r_k = r_{k-1}^\top (b - A(x_{k-1} + \alpha_k r_{k-1})) = r_{k-1}^\top r_{k-1} - \alpha_k r_{k-1}^\top A r_{k-1} = 0.$$

This indicate that the method is myopic in the sense that it often searches in similar directions to those searched before. In fact, it can be shown that the convergence is linear with rate $\gamma_{sd} = \frac{\kappa_2(A)-1}{\kappa_2(A)+1} = 1 - \frac{2}{\kappa_2+1}$ with $\kappa_2(A)$ is the condition number of A in the 2-norm. By (4), we know $\kappa_2(A)$ can be very large hence γ_{sd} can be very close to 1. In the next subsection, we will derive a method with rate $\gamma_{cg} = \frac{\sqrt{\kappa_2(A)}-1}{\sqrt{\kappa_2(A)}+1} = 1 - \frac{2}{\sqrt{\kappa_2}+1} < \gamma_{sd}$.

1.5.2 Conjugate gradient

In steepest descent method, the sequence $x_0, ..., x_k, ...$ is found by one-dimensional minimization of f in the direction of the gradient

$$f(x_{k+1}) = \min_u f(x_k + ur_k) \qquad \text{with } r_k = b - Ax_k. \tag{11}$$

We will now consider

$$f(x_{k+1}) = \min_{u_0,...,u_k} f(x_k + \sum_{j=0}^{k} u_j p_j) \qquad \text{with } \{p_j\} \text{ satisfy } p_i^\top A p_j = 0\ . \tag{12}$$

Those $\{p_j\}$ are said to be conjugate with respect to A and $p_0 = b - Ax_0$.

Note that p_i's are linearly independent (otherwise, assume $p_k = \sum_{j\neq k} c_j p_j$ and dot it with Ap_k to see the contradiction.) So, for a problem in $\mathbb{R}^n$, $p_0, ..., p_{n-1}$ will expand the whole $\mathbb{R}^n$ and therefore (12) will return the exact solution in at most n steps.

If (12) is true,

$$\begin{aligned} 0 = \frac{\partial f(x_{k+1})}{\partial u_i} &= \left\langle \nabla f(x_k + \sum_{j=0}^{k} u_j p_j), p_i \right\rangle = \left\langle A(x_k + \sum_{j=0}^{k} u_j p_j) - b, p_i \right\rangle \\ &= \langle -r_k + u_i A p_i, p_i \rangle \end{aligned}$$

From the above equation, we know two things: (1) when looking for $x_{k+1} = x_k + \sum_{j=0}^{k} u_j p_j$,

$$u_i = \frac{r_k^\top p_i}{p_i^\top A p_i}.$$

(2) $r_{k+1}^\top p_i = 0$ for $i = 0, ..., k$ because $\nabla f(x_{k+1}) = -r_{k+1}$. (2) implies that $u_i = 0$ for $i = 1, ..., k-1$ in (1). So we have $x_{k+1} = x_k + \alpha_k p_k$ with $\alpha_k = \frac{r_k^\top p_k}{p_k^\top A p_k}$.

The conjugate gradient method (CG) is as follows:

1. Initialize x_0, $r_0 = b - Ax_0$, $p_0 = r_0$.

2. For $k = 0, 1, ...$ until some stopping criteria is satisfied

- $\alpha_k = r_k^\top p_k / p_k^\top A p_k$ ($= r_k^\top r_k / p_k^\top A p_k$ as we will prove later)
- $x_{k+1} = x_k + \alpha_k p_k$
- $r_{k+1} = r_k - \alpha_k A p_k$
- $\beta_k = \frac{r_{k+1}^\top r_{k+1}}{r_k^\top r_k}$
- $p_{k+1} = r_{k+1} + \beta_k p_k$

Remark: We will prove $r_k^\top p_k = r_k^\top r_k$ in Theorem 8. The advantage of using the later is to save one inner product computation per iteration.

The matrix A doesn't have to be stored explicitly. It suffices to store the product Ap_k instead. In addition, we only need to store three other vectors x_{k+1}, r_{k+1} and p_{k+1}. We also need to store the inner product $r_k^\top r_k$ from the previous iteration.

When the conjugate gradient method is applied, n is usually so large that $O(n)$ iterations requires unacceptable amount of work. It is customary to regard the method as a genuinely iterative method with termination based on an iteration maximum $k_{\max}$ and the relative 2-norm of the residue $\frac{\|r_k\|_2}{\|b\|_2}$. Note that one reason to use 2-norm $\|r_k\|_2$ is because we have already computed it in $\beta_k = \|r_{k+1}\|_2^2/\|r_k\|_2^2$.

Remark: : If $r_k = b - Ax_k$, (which is true for $k = 0$)

$$r_{k+1} = b - Ax_k - \alpha_k Ap_k = b - Ax_{k+1}.$$

The way used to compute r_{k+1} saves one computation of Ax_{k+1} since we have already computed Ap_k.

We are only left to verify the following property:

Theorem 8 *$p_i^\top Ap_j = 0$ if $i < j$.*

Proof: We will prove the following statement S_k by induction

$$S_k: \quad \begin{array}{lll} (1) & r_j^\top p_i = 0 & \text{for } i < j \le k \\ (2) & r_i^\top p_i = r_i^\top r_i & \text{for } i \le k \qquad r_{k+1} \neq 0 \\ (3) & p_i^\top Ap_j = p_j^\top Ap_i = 0 & \text{for } i < j \le k \end{array}$$

S_0 is true because $r_0 = p_0$. Suppose S_k is true and we want to show S_{k+1} is true.

(1) By the definition of α_k, $r_{k+1}^\top p_k = r_k^\top p_k - \alpha_k p_k^\top Ap_k = 0$. Because of S_k-(1)(3), $r_{k+1}^\top p_j = 0$ for any $j < k$.

(2) $r_{k+1}^\top p_{k+1} = r_{k+1}^\top (r_{k+1} + \beta_k p_k) = r_{k+1}^\top r_{k+1}$.

(3) By S_k-(2), $\alpha_k \neq 0$, otherwise, $r_k = 0$ and we have already converged in the last step and don't need to go to step $k+1$. Then we have

$$\begin{aligned} p_{k+1}^\top Ap_k &= r_{k+1}^\top Ap_k + \beta_k p_k^\top Ap_k = \frac{1}{\alpha_k} r_{k+1}^\top (r_k - r_{k+1}) + \beta_k p_k^\top Ap_k \\ &= \frac{1}{\alpha_k} r_{k+1}^\top (p_k - \beta_{k-1} p_{k-1} - p_{k+1} + \beta_k p_k) + \beta_k p_k^\top Ap_k \\ &= -\frac{1}{\alpha_k} r_{k+1}^\top r_{k+1} + \beta_k p_k^\top Ap_k = 0. \end{aligned}$$

In the last step, we have used the definition of α_k, β_k. For $j < k$,

$$\begin{aligned} p_{k+1}^\top Ap_j &= r_{k+1}^\top Ap_j + \beta_k p_k^\top Ap_j = \frac{1}{\alpha_j} r_{k+1}^\top (r_j - r_{j+1}) \\ &= \frac{1}{\alpha_j} r_{k+1}^\top (p_j - \beta_{j-1} p_{j-1} - p_{j+1} + \beta_j p_j) = 0. \qquad \square \end{aligned}$$

Remark: Notice that we also have $r_i^\top r_j = 0$ for $i < j$. This is because $r_i = p_i - \beta_{i-1} p_{i-1}$ and S_k-(1).

1.5.3 Krylov space and convergence rate of conjugate gradient method

Krylov space $K_i(q, A)$ is the subspace spanned by the first i vectors of the sequence $\{A^i q\}_{i\geq 0}$.

Because $p_0 = r_0 = b - Ax_0$, $r_{k+1} = r_k - \alpha A p_k$ and $p_{k+1} = r_k - \alpha_k A p_k + \beta p_k$, by induction, one can prove that

$$r_k \text{ and } p_k \in \text{span}\{r_0, Ar_0, ..., A^k r_0\} = K_{k+1}(r_0, A).$$

Introduce the norm

$$\|x\|_A = \left(x^\top A x\right)^{1/2}$$

and note that if $Ax_e = b$, $\frac{1}{2}\|x - x_e\|_A^2 = \frac{1}{2}x^\top Ax - x^\top b + \frac{1}{2}x_e^\top Ax_e = f(x) + \text{Const}$ with f defined in (10). So (12) can be rewritten as

$$\|x_{k+1} - x\|_A = \min\{\|y - x\|_A \ ; \ y \in x_0 + K_{k+1}\} \tag{13}$$

where x is the exact solution of $Ax = b$.

If we introduce the error $e_j = x_j - x$, then because $r_0 = -Ae_0$, any $y \in x_0 + K_{k+1}$ satisfies

$$y - x \in x_0 - x + K_{k+1} = e_0 + \text{span}\{Ae_0, ..., A^{k+1}e_0\}.$$

Therefore, there is a real polynomial $p(t) = 1 + \alpha_1 t + ... + \alpha_{k+1}t^{k+1}$ with $y - x = p(A)e_0$. So (13) means

$$\|e_{k+1}\|_A = \min\{\|p(A)e_0\|_A \ ; \ p \in \bar{\boldsymbol{P}}_{k+1}\}$$

where $\bar{\boldsymbol{P}}_{k+1}$ denotes the set of all real polynomials of degree $\leq k+1$ with $p(0) = 1$. (In particular, once $k+1$ reaches the degree of the minimal polynomial of A, there is p_{k+1} so that $p_{k+1}(A) = 0$.) Because A is symmetric positive definite, it can be diagonalized by orthonormal matrix. Denote the eigenvalue and orthonormal eigenvector of A by $\lambda_1 \geq ... \geq \lambda_n > 0$ and $z_1, ..., z_n$. Let $e_0 = \sum_j \beta_j z_j$. So

$$\|e_0\|_A^2 = e_0^\top A e_0 = \sum_{j=1}^n \lambda_j \beta_j^2$$

$$\|p(A)e_0\|_A^2 = e_0^\top (p^2(A)A)e_0 = \sum_{j=1}^n p(\lambda_j)^2 \lambda_j \beta_j^2 \leq \left(\max_j p(\lambda_j)^2\right) \|e_0\|_A^2.$$

Therefore

$$\frac{\|e_{k+1}\|_A}{\|e_0\|_A} \leq \min_{p \in \bar{\boldsymbol{P}}_{k+1}} \max_j |p(\lambda_j)| \leq \min_{p \in \bar{\boldsymbol{P}}_{k+1}} \max_{\lambda \in [\lambda_n, \lambda_1]} |p(\lambda)| \tag{14}$$

Because it is $\min_{p \in \bar{\boldsymbol{P}}_{k+1}}$, we can obtain an upper bound of the right hand side of (14) by selecting a specific polynomial (which in fact is optimal). Recall the Chebyshev polynomial

$$T_{k+1}(x) = \cos(k+1)(\arccos x), \qquad x \in [-1, 1]$$

We need to rescale T_{k+1} so that it is defined on $[\lambda_n, \lambda_1]$, and then normalized it so that it is in $\bar{\boldsymbol{P}}_{k+1}$. So define

$$p_{k+1}(\lambda) = \frac{T_{k+1}\left(\frac{2\lambda-(\lambda_1+\lambda_n)}{\lambda_1-\lambda_n}\right)}{T_{k+1}\left(\frac{-(\lambda_1+\lambda_n)}{\lambda_1-\lambda_n}\right)}.$$

Because $\max_x |T_{k+1}(x)| = 1$,

$$\max_{\lambda\in[\lambda_n,\lambda_1]} |p_{k+1}(\lambda)| = \frac{1}{\left|T_{k+1}\left(\frac{-(\lambda_1+\lambda_n)}{\lambda_1-\lambda_n}\right)\right|} = \frac{1}{\left|T_{k+1}\left(\frac{\kappa_2+1}{\kappa_2-1}\right)\right|}$$

where $\kappa_2 = \frac{\lambda_1}{\lambda_n}$ is the condition number (see (4)) and we have used the fact that $|T_{k+1}(-x)| = |T_{k+1}(x)|$.

To obtain an upper bound of $\left|T_{k+1}\left(\frac{\kappa_2+1}{\kappa_2-1}\right)\right|^{-1}$, we use the following property of T_n

$$T_n\left(\frac{z+z^{-1}}{2}\right) = \frac{z^n+z^{-n}}{2}$$

and note that $\frac{\kappa_2+1}{\kappa_2-1} = \frac{z+z^{-1}}{2}$ with $z = \frac{\sqrt{\kappa_2}+1}{\sqrt{\kappa_2}-1}$. So $T_{k+1}\left(\frac{\kappa_2+1}{\kappa_2-1}\right) = \frac{z^{k+1}+z^{-k-1}}{2}$. Hence

$$\frac{\|e_{k+1}\|_A}{\|e_0\|_A} \le \frac{1}{\left|T_{k+1}\left(\frac{\kappa_2+1}{\kappa_2-1}\right)\right|} \le \frac{2}{z^{k+1}} = 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^{k+1}.$$

Theorem 9 *Consider conjugate gradient method for $Ax = b$ with A symmetric positive definite. Let $\kappa_2 = \frac{\lambda_{max}(A)}{\lambda_{min}(A)}$ be the condition number of A. Then*

$$\|x_k - x\|_A \le 2\left(\frac{\sqrt{\kappa_2}-1}{\sqrt{\kappa_2}+1}\right)^k \|x_0 - x\|_A.$$

1.6 Preconditioned conjugate gradient

Consider the $n \times n$ symmetric positive definite linear system $Ax = b$.

Left preconditioner $M^{-1}Ax = M^{-1}b$. So we need to compute $M^{-1}Ap$ in each iteration which means we have to solve an equation like $My = c$. A preconditioner is good if $M^{-1}A$ is not too far from normal and its eigenvalue are clustered.

Right preconditioner $AM^{-1}y = b$, $x = M^{-1}y$.

To preserve symmetric positive definite of A, we consider symmetric preconditioner: We apply the "regular" conjugate gradient method to the transformed system

$$\tilde{A}\tilde{x} = \tilde{b} \tag{15}$$

where $\tilde{A} = C^{-1}AC^{-1}$, $\tilde{x} = Cx$, $\tilde{b} = C^{-1}b$ and C is symmetric positive definite. (Hence $C^{-1} = \Omega\Lambda^{-1}\Omega$ is also symmetric positive definite.) The game is to choose C so that $\tilde{A}$ is well conditioned and for reasons that will soon emerge, the matrix C^2 must also be "easy to invert".

We apply algorithm in Section 1.5.2 to (15), we get (relation to $Ax = b$ is put inside [ ])

1. Initialize $\tilde{x}_0$, $\tilde{r}_0 = \tilde{b} - \tilde{A}\tilde{x}_0$ $[= C^{-1}(b - Ax_0) = C^{-1}r_0]$, $\tilde{p}_0 = \tilde{r}_0 [= C^{-1}r_0]$.

2. For $k = 0, 1, ...$ until some stopping criteria is satisfied

- $\alpha_k = \tilde{r}_k^\top \tilde{r}_k / \tilde{p}_k^\top \tilde{A}\tilde{p}_k$ $\left[= r_k^\top (C^{-2}r_k)/(C^{-1}\tilde{p}_k)^\top A(C^{-1}\tilde{p}_k) \text{ with } C^{-1}r_k = \tilde{r}_k\right]$
- $\tilde{x}_{k+1} = \tilde{x}_k + \alpha_k \tilde{p}_k$, [i.e. $x_{k+1} = x_k + \alpha_k C^{-1}\tilde{p}_k$]
- $\tilde{r}_{k+1} = \tilde{r}_k - \alpha_k \tilde{A}\tilde{p}_k$, [i.e. $r_{k+1} = r_k - \alpha_k A(C^{-1}\tilde{p}_k)$ with $C^{-1}r_{k+1} := \tilde{r}_{k+1}$]
- $\beta_k = \frac{\tilde{r}_{k+1}^\top \tilde{r}_{k+1}}{\tilde{r}_k^\top \tilde{r}_k}$ $\left[= \frac{r_{k+1}^\top (C^{-2}r_{k+1})}{r_k^\top (C^{-2}r_k)}\right]$
- $\tilde{p}_{k+1} = \tilde{r}_{k+1} + \beta_k \tilde{p}_k$, [i.e. $C^{-1}\tilde{p}_{k+1} = C^{-2}r_{k+1} + \beta_k C^{-1}\tilde{p}_k$]

Now, it is very clear that we can set $p_k = C^{-1}\tilde{p}_k$ and $q_k = C^{-2}r_k$ (by solving $Mq_k := C^2 q_k = r_k$) and rewrite the above algorithm into a more clever way:

1. Initialize x_0, $r_0 = b - Ax_0$, solve $Mq_0 = r_0$, $p_0 = q_0$.

2. For $k = 0, 1, ...$ until some stopping criteria is satisfied

- $\alpha_k = r_k^\top q_k / p_k^\top A p_k$
- $x_{k+1} = x_k + \alpha_k p_k$
- $r_{k+1} = r_k - \alpha_k A p_k$
- solve $Mq_{k+1} = r_{k+1}$
- $\beta_k = \frac{r_{k+1}^\top q_{k+1}}{r_k^\top q_k}$
- $p_{k+1} = q_{k+1} + \beta_k p_k$

Remark: Although the transformation C figured heavily in the derivation, its action is only left through the preconditioner $M = C^2$.

One of the most important preconditioning strategies involves computing an *incomplete* Cholesky factorization of A. We attempt to find the standard Choleskey factorization $A = GG^\top$. We approximate G by L where L is a lower triangular matrix and when going through the standard Choleskey factorization, we insist that at any stage, if $A(i,j) = 0$, then $L(i,j) = 0$ also. This is done to preserve the sparsity structure of A.

The preconditioner is then taken to be $M = LL^\top$. On the other hand, we can assume $M = C^2$. Because M is positive semi-definite, we have the existence of such a C. From linear algebra, we know for any matrix C, there is is orthogonal matrix Q and upper triangular matrix R such that $C = QR$. So, $M = R^\top R$ and therefore $R = L^\top$ by the uniqueness of Cholesky factorization of M. Then, we see that

$$\begin{aligned}\tilde{A} &= C^{-1}AC^{-1} = C^{-\top}AC^{-1} = (QL^\top)^{-\top}A(QL^\top)^{-1} = QL^{-1}AL^{-\top}Q \\ &= Q(L^{-1}GG^\top L^{-\top})Q \approx I\end{aligned}$$

1.7 Homework I

1.7.1 Part a.

1) Assume $A \in \mathbb{R}^{m\times n}$. Prove $\|A\|_1 = \max_{1\le j\le n}\left(\sum_{i=1}^m |a_{ij}|\right)$ and construct a square matrix A so that $\|A\|_1 < 1 < \|A\|_\infty$ and a square matrix B so that $\|B\|_\infty < 1 < \|B\|_1$. That's why we can have $\lim_{k\to\infty}\|A\|^k = 0$ for some norm $\|\cdot\|$ but not for all the norm.

   Remark: As long as $\lim_{k\to\infty}\|A\|^k = 0$ for some norm, $\lim_{k\to\infty}\|A^k\| = 0$ for all the norm (why? since $\|\cdot\|_* \le C\|\cdot\|_\bullet$ for *any* two norm $\|\cdot\|_*$ and $\|\cdot\|_\bullet$ on a finite dimensional vector space where C is a positive constant depending on the two norms we are comparing, but nothing else. That's a theorem you have not learned. As an example, see the problem from tutorial question set I. Using that theorem, as long as we have $\lim_{k\to\infty}\|A^k\|_\bullet = 0$ for some $\|\cdot\|_\bullet$, we have $\lim_{k\to\infty}\|A^k\|_* = 0$ for any norm $\|\cdot\|_*$.)

2) Given vector $a, b \in \mathbb{R}^n$, define

$$\langle a, b\rangle = \sum_{i=1}^n a_i b_i$$

   Prove

$$\|a\|_1 = \sup_{\|b\|_\infty=1}\langle a, b\rangle \qquad \text{and} \qquad \|a\|_\infty = \sup_{\|b\|_1=1}\langle a, b\rangle .$$

   Then using the above facts to prove that for any matrix $A \in \mathbb{R}^{m\times n}$

$$\|A\|_1 = \|A^\top\|_\infty.$$

   Remark: Recall the Holder inequality

$$\left|\sum_i x_i y_i\right| \le \left(\sum_i |x_i|^p\right)^{1/p}\left(\sum_i |y_i|^{p'}\right)^{1/p'} .$$

   when p and p' satisfy $\frac{1}{p} + \frac{1}{p'} = 1$. The quality is true when $y_i = x_i^{p-1}$ since $p'(p-1) = p$ when $1 < p < \infty$. Using Holder inequality, similar to the above case, one can prove

$$\|A\|_p = \|A^\top\|_{p'}.$$

   In particular, we have

$$\|A\|_2 = \|A^\top\|_2$$

   which again can be directly verified. Note that $\rho(A^\top A) = \rho(AA^\top)$ follows from the singular value decomposition of A: $A = U\Lambda V$ and therefore $\rho(A^\top A) = \rho(V^\top\Lambda^\top\Lambda V) = \rho(\Lambda^\top\Lambda) = \max_i|\sigma_i|^2 = \rho(U\Lambda\Lambda^\top U^\top) = \rho(AA^\top)$.

1.7.2 Part b.

Suppose we are numerically solving

$$-u''(x) = f(x) \text{ on } [0,1], \qquad u(0) = u(1) = 0 \tag{16}$$

(As a notation, $u'' = \frac{d^2u}{dx^2}$). If we know $u(x)$, we can easily find $f(x) = -u''(x)$. But in most situation, we are given $f(x)$ and are asked to find $u(x)$. How to do that? So, given interval $[0,1]$ and a number n, we divide $[0,1]$ into $n+1$ intervals of the same size. So the size of each interval is $h = 1/(n+1)$. Let $x_j = jh$. Then $x_0 = 0$ and $x_{n+1} = 1$. We know $u(x_0) = 0$ and $u(x_{n+1}) = 1$ from the given condition. We want to find $u(x_j)$ for all the x_j's in between. If we can do so, hopefully, we can connect those $u(x_j)$'s by straight lines and obtain a function which is close to $u(x)$.

The idea is to set up a system of equations for $u(x_j)$'s and solve the resulting system of equations. How to set up the system of equations? Recall that by definition $u'(x) = \lim_{h\to 0} \frac{u(x+h)-u(x)}{h}$. So, we have

$$u'(x) \approx \frac{u(x+h)-u(x)}{h} \qquad \text{e.g. } u'(x_j) = \frac{u(x_{j+1})-u(x_j)}{h}.$$

$$\begin{aligned} u''(x_j) &\approx \frac{u(x_j+\frac{h}{2}) - u(x_j-\frac{h}{2})}{h} \approx \frac{\frac{u(x_j+h)-u(x_j)}{h} - \frac{u(x_j)-u(x_j-h)}{h}}{h} \\ &= \frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2}. \end{aligned}$$

1 Prove by Taylor expansion that

a)
$$\frac{u(x+h)-u(x)}{h} = u'(x) + \frac{u''(\theta)}{2}h$$
for some θ between x and $x+h$, assuming $u \in C^2([0,1])$.

b)
$$\frac{u(x+h)-2u(x)+u(x-h)}{h^2} = u''(x) + \frac{u^{(4)}(\theta)}{12}h^2$$
for some θ between $x-h$ and $x+h$, assuming $u \in C^4([0,1])$.

As a consequence, if $u''(x)$ and $u^{(4)}(x)$ remain bounded in $[0,1]$ (thinking about $u = x^8$, $\cos(x)$, $\sin(x)$ or e^x), then as h goes to zero, the finite difference approximation approaches the associated derivatives.

This leads to the so called central difference scheme for $-u''(x) = f(x)$: Fix n, Let $h = \frac{1}{n+1}$ and $x_i = hi$ with $i = 1, ..., n$. We get a system of equation for $u(x_i)$'s with $i = 1, ..., n$

$$-\frac{u(x_{i+1}) - 2u(x_i) + u(x_{i-1})}{h^2} = f(x_i).$$

Note the because $u(x_0)$ and $u(x_{n+1})$ are already known, the number of equations equal to the number of unknowns.

2 Write down a matrix $A \in \mathbb{R}^{n\times n}$ and translate the above system of equations into the form $Ay = b$. If you are interested (just for fun. But I cannot give you extra credit for it, sorry about that), prove that A is positive definite. Note that the operator $-\frac{d^2}{dx^2}$ is a positive operator in the sense that $\int_0^1 \left(-\frac{d^2}{dx^2}u\right) u dx = \int_0^1 |u'|^2 dx \geq 0$ for any $u \in C^2$ with $u(0) = u(1) = 0$. So, the discrete operator preserves this property. One proof is to mimic this property and consider

$$y^\top Ay = \sum_{i=1}^{n} -y_i \frac{y_{i+1} - 2y_i + y_{i-1}}{h^2}$$

with $y_0 = y_{n+1} := 0$. Then prove the above quantity > 0 if $y \neq 0$.

3 Prove that the eigenvalues of of the symmetric matrix h^2A is $\{\lambda_k = 2 - 2\cos\theta_k \ ; \ \theta_k = k\pi/(n+1), k = 1, ..., n\}$. (Note that if you multiply A by h^2, every eigenvalue is simply multiplied by h^2 and hence the condition number remain the same (why?).)

- One way to prove it is to verify that the associated eigenvector for λ_k is $u_k = (\sin(\theta_k), \sin(2\theta_k), ...)^\top$, i.e. verify that $h^2Au_k = \lambda_k u_k$.
- Another way to prove it is to study $f_n(\lambda) = \det(\lambda I - A_n)$ with $A_n = \text{tridiag}(-1, 2, -1)$. Let us study a result for more general tridiagonal matrix first. Let $g_k = \det \text{tridiag}(a, b, c)$. Then by Laplace expansion, one can have $g_n = bg_{n-1} - acg_{n-2}$ with $g_1 = b$. By a general procedure of solving equations of this type, $g_n = \frac{r_1^{n+1} - r_2^{n+1}}{r_1 - r_2}$ with r_1, r_2 the root of $x^2 = bx - ac$. (In particular, one can verify $g_1 = \frac{r_1^2 - r_2^2}{r_1 - r_2} = r_1 + r_2 = b$ and $g_2 = \frac{r_1^3 - r_2^3}{r_1 - r_2} = b^2 - ac$.)

  Apply this result to $A_n = \text{tridiag}(-1, 2, -1)$, we have $f_n(\lambda) = \frac{r_1(\lambda)^{n+1} - r_2(\lambda)^{n+1}}{r_1(\lambda) - r_2(\lambda)}$ with $r_1(\lambda)$, $r_2(\lambda)$ the roots of $x^2 = (2 - \lambda)x - 1$. Notice that when $r_1(\lambda) = r_2(\lambda)e^{i\frac{2k\pi}{n+1}}$, $r_1^{n+1} = r_2^{n+1}$ and we have $f_n(\lambda) = 0$. So, we have to find λ so that the two solutions (called r_1 and r_2) of $x^2 = (2 - \lambda)x - 1$ satisfies $r_1 = r_2 e^{i\frac{2k\pi}{n+1}}$. From $r_1 r_2 = 1$, we know $r_2^2 e^{i\frac{2k\pi}{n+1}} = 1$. Then $2 - \lambda = r_1 + r_2 = e^{i\frac{k\pi}{n+1}} + e^{i\frac{-k\pi}{n+1}}$. So $\lambda = 2 - 2\cos\left(\frac{k\pi}{n+1}\right)$.

4 Estimate the condition number of A with respect to the 2-norm and estimate how fast it increases as $n + 1$ increase. What does this imply when you apply conjugate gradient to solve $Au = f$?

1.8 Computer project I

Let $f(x) = \pi^2 \sin(\pi x)$ so that the exact solution of (16) is $u(x) = \sin(\pi x)$.

Implement the conjugate gradient method with matlab, and solve (16) numerically with $n + 1 = 10, 20, 40, 80$. Plot the results and compare it with the exact solution.

Note that (to save storage) you don't need to store A. You only need a function that will return Au when it is given a u as input.

1.9 Homework related: $-\frac{d^2}{dx^2}$ and $\frac{1}{h^2}$tridiag$(-1,2,-1)$

In the homework and tutorial problems, we have discussed how to solve

$$-\frac{d^2}{dx^2}u(x) = f(x) \text{ on } [0,1], \qquad u(0) = u(1) = 0. \tag{17}$$

The discretization of the above problem leads to

$$A\vec{u} = \vec{f}$$

where $\vec{f} = (f(x_1), ..., f(x_n))^\top \in \mathbb{R}^{n\times 1}$ and $A = \frac{1}{h^2}\text{tridiag}(-1,2,-1) \in \mathbb{R}^{n\times n}$. Here $[0,1]$ is divided into $n+1$ subintervals and $x_i = ih$ with $h = \frac{1}{n+1}$.

For any number ξ, we know

$$-\frac{d^2}{dx^2}\sin(\xi x) = \xi^2 \sin(\xi x).$$

If we require $\sin(\xi x)$ to satisfies the boundary condition, $\sin(\xi 0) = 0 = \sin(\xi 1)$. So, $\xi = k\pi$ for $k \in \{1, 2, ...\}$. We will not consider negative k and $k = 0$ any more. Because for negative k, $\sin(-k\pi x) = -\sin(k\pi x)$ and eigenfunctions are determined up to a constant coefficient; for $k = 0$, the eigenvalue is zero and we know eigenvalues are assumed to be non-zero from the very beginning. (Otherwise, 0 is always an eigenvalue.)

So, the kth eigenfunction of the "operator" $-\frac{d^2}{dx^2}$ is $(k\pi)^2$ and its associated eigenfunction is

$$v^k(x) = \sin(k\pi x).$$

Note that when k increase, the graph of $\sin(k\pi x)$ as a function of x becomes more and more oscillatory on $[0,1]$.

Now we fix an n and fix a $k \le n$. If we evaluate the kth eigenfunction of $-\frac{d^2}{dx^2}$, which is v^k, at the grid points $x_1, ..., x_n$, we obtain a vector

$$\vec{v^k} = \left(\sin(k\pi\frac{1}{n+1}), \sin(k\pi\frac{2}{n+1})..., \sin(k\pi\frac{n}{n+1})\right)^\top \in \mathbb{R}^{n\times 1}.$$

With the $\theta_k = \frac{k\pi}{n+1}$ that we have introduced in the homework, one immediately see

$$\vec{v^k} = (\sin(\theta_k), \sin(2\theta_k)..., \sin(n\theta_k))^\top$$

which is exactly the kth eigenvector of A that we have discussed in the homework assignment. As we know when k increase, the graph of $\sin(k\pi x)$ becomes more and more oscillatory on $[0,1]$, the vector $\vec{v^k}$ becomes more and more oscillatory at k increase.

As $\{\vec{v^1}, ..., \vec{v^n}\}$ form a base of $\mathbb{R}^n$, any vector e^0 can be expanded by

$$e^0 = \sum_j \beta_j^0 \vec{v^j}.$$

We know $\vec{v}^k$ is also the eigenvector of $J = \frac{1}{2}(2I - A)$ which is the Jacobi iteration matrix. Hence if we expand the initial error e^0 as a linear combination of $\vec{v}^j$ as above,

$$e^m = J^m e^0 = J^m \left(\sum_j \beta_j^0 \vec{v}^j \right) = \sum_j \beta_j^0 \lambda_j^m \vec{v}^j$$

where $\lambda_j = \cos(\theta_j)$ is the eigenvalues of J that we have seen in the tutorial problems. For some j, $\lambda_j = \cos(\theta_j)$ is close to zero and hence $\beta_j^0 \lambda_j^m \vec{v}^j \to 0$ very fast when $m \to \infty$. For some j, (like $j = 1$ and $j = n$), $\lambda_j = \cos(\theta_j)$ is close to 1 or -1 and hence $\beta_j^0 \lambda_j^m \vec{v}^j \to 0$ very slow when $m \to \infty$. The weighted Jacobi iteration discussed in the tutorial problem will change the iteration matrix J and hence the λ_j. One can make smart choice so that the new λ_j's associated with very oscillatory $\vec{v}^k$'s are very close to zero. So, "the rather oscillatory components of the initial error e^0 decay very fast". This paves the way for the development of multigrid method.

In the end, let us also look at the kth eigenvalue values of A. By the homework, it is $\frac{1}{h^2}(2 - 2\cos(\theta_k))$. By Taylor expansion

$$\begin{aligned} &\frac{1}{h^2}(2 - 2\cos(\theta_k)) = \frac{1}{1/(n+1)^2}(2 - 2\cos(\theta_k)) \\ \approx &(n+1)^2(2 - 2(1 - \frac{1}{2}\theta_k^2)) = (n+1)^2 \left(\frac{k\pi}{n+1} \right)^2 = (k\pi)^2. \end{aligned}$$

So the kth eigenvalue value of A is close to the kth eigenvalue value of $-\frac{d^2}{dx^2}$. With the previous discussion about the relation between the eigenvectors and eigenfunctions, we see that the matrix A is indeed a good approximation of the "operator" $-\frac{d^2}{dx^2}$.

Moreover, in the homework, I have mentioned that you can prove

$$y^\top A y = \sum_j -y_i \frac{y_{i+1} - 2y_i + y_{i-1}}{\Delta x^2} = \sum_j \left| \frac{y_{j+1} - y_j}{\Delta x} \right|^2 \geq 0 \tag{18}$$

with $y_0 = 0 = y_{n+1}$. This is exactly the discrete version of

$$\int_0^1 \left(-\frac{d^2}{dx^2} y(x) \right) y(x) dx = \int_0^1 \left| \frac{d}{dx} y(x) \right|^2 dx \geq 0$$

which is true because of the integration by part formula. So the last equality in (18) is an application of the so called "summation" by part formula which mimics the integration by part formula.

1.10 Tutorial Question Set: I

1. If A is an $n \times n$ matrix. To solve $Ax = f$ by Gauss elimination, how many flops are involved?

2. Recall that a norm on a space X is a function $\|\cdot\|$: $X \to \mathbb{R}$ that assigns a real value to each vector and also satisfies the following three conditions

   I) $\|x\| \geq 0$ and $\|x\| = 0$ if and only if $x = 0$.

   II) $\|x + y\| \leq \|x\| + \|y\|$.

   III) $\|\alpha x\| = |\alpha|\|x\|$.

   We have learned vector norm ($X = \mathbb{R}^n$ or $\mathbb{C}^n$) and matrix norm ($X = \mathbb{R}^{m\times n}$ or $\mathbb{C}^{m\times n}$) so far. Suppose x is a $m-$vector and A is an $m \times n$ matrix. Prove

   a) $\|x\|_\infty \leq \|x\|_2$

   b) $\|x\|_2 \leq \sqrt{m}\|x\|_\infty$

   c) $\|A\|_\infty \leq \sqrt{n}\|A\|_2$

   d) $\|A\|_2 \leq \sqrt{m}\|A\|_\infty$

3. Let A be an $m \times n$ matrix and let B be a submatrix of A, that is, a matrix obtained by selecting certain rows and columns of A. Explain how B can be obtained from A by certain tow and column "deletion matrices". Using this product to show that $\|B\|_p \leq \|A\|_p$ for any p with $1 \leq p \leq \infty$.

4. When A and B are both $m \times n$ matrices, prove

$$\|A\| - 2\|B\| + \|2B - A\| \leq 2\|A - B\|.$$

5. (a) Prove that if A is invertible, then $\|Ax\| \geq \|x\|\|A^{-1}\|^{-1}$ where $\|A^{-1}\|$ is the matrix norm induced by the vector norm $\|x\|$.

   (b) Prove that if A is invertible and if λ is an eigenvalue of A, then $\|A^{-1}\|^{-1} \leq |\lambda| \leq \|A\|$. Here the matrix norm is induced by any vector norm.

1.11 Tutorial Question Set: II

1. Suppose you apply Jacobi iteration to the system

$$\begin{cases} 3x_1 + x_2 &= 5, \\ x_1 + 2x_2 &= 5. \end{cases}$$

starting from $x^0 = (0,0)^\top$. Find x^1, x^2, x^3, x^4. What if you apply it to

$$\begin{cases} x_1 + 2x_2 &= 5, \\ 3x_1 + x_2 &= 5. \end{cases}$$

with the same initial guess? Do you expect convergence in the second case? If you write the above system of equations as $Ax = b$, in which case the matrix A is strictly diagonally dominant?

2. Preconditioner and Jacobi and Gauss-Seidel iteration for $Ax = b$: We can write $Ax = b$ as $x = (I - A)x + b$. A preconditioner P is a matrix that is *close to* A and is *easy to invert*. Suppose we have such a P, then we can use

$$Px = (P - A)x + b.$$

and iterate by $Px^{k+1} = (P - A)x^k + b$. In each step that goes from x^k to x^{k+1}, we have to solve a system $Px^{k+1} = \cdots$. This is feasible because P can be easily inverted. In the case $P = A$, this goes back to solving $Ax = b$ in one step.

Write down the equation satisfies by $e^k = x - x^k$.

Naturally, one can think of taking $P = D$ or $P = L + D$ where $A = L + D + U$. They are easy to invert and hopefully can be close to A since they are part of A. Show that they are Jacobi and Gauss-Seidel iteration respectively.

3. Suppose we take $A = \text{tridiag}(-1, 2, -1) \in \mathbb{R}^{n\times n}$ that we have seen in the homework. If we use Jacobi iteration, write down the matrix J so that $e^{k+1} = Je^k$. Show that $\|J\|_1 = 1 = \|J\|_\infty$. So, by looking at these two norms, we can not say $e^k = J^k e^0$ will converge to zero. In order to prove this, compute the spectral radius $\rho(J)$ and determine how does it depend on n which is the size of the matrix? [Hint: to compute the eigen vlaues, you can make use of the fact that the eigen vlaues of matrix $\text{tridiag}(-1, 2, -1)$ is $2 - 2\cos(\theta_k)$. Look at your homework.]

4. One can expand the initial error e^0 in terms of the eigen vectors of J. We say that it contains different modes. Determine which modes decay the slowest. If we do a weighted Jacobi iteration, namely, take $P = \frac{1}{w}D$ with $w = \frac{2}{3}$, what would you say about the decaying of different modes? [Actually, thoughts along this line lead to the so called multigrid method.]

5. Suppose you want to solve $Ax = b$ by the Gauss-Seidel iteration with $A = \begin{pmatrix} 1 & \alpha \\ \alpha & 1 \end{pmatrix}$. Determine the necessary and sufficient condition for α so that the Gauss-Seidel iteration converges to the exact solution.

1.12 Lab tutorial: I

1) Generate an $(n-1)\times(n-1)$ sparse matrix $A = \text{tridiag}(-1, 2, -1)$ and compute its condition number using `condest`. Compare the result with your solution of problem 4 from the part b of Homework I.

2) Solve $-u'' = f$ numerically on $[0, 1]$ with n subintervals. The boundary condition is $u(0) = 0 = u(1)$ and $f = \pi^2 \sin(\pi x)$. The exact solution is $u(x) = \sin(\pi x)$.

   Let $n + 1 = 10; 20; 40; 80$. Plot the results and compare it with the exact solution. Suppose for a fixed $h = 1/n$, the numerical solution is u_h^i at grid point $x_i = ih$. Plot the error $e_h = \max_i |u_h^i - u(x_i)|$ for different h and see how it changes when h decreases.

   First, by Matlab function backslash (\) to solve $Ax = b$. Then, in the Computer Project I, you need to replace the backslash by conjugate gradient method.

2 Nonlinear equation and system of nonlinear equations

In this chapter, we want to study how to solve nonlinear equation like

$$\cos(x) - x = 0$$

or system of nonlinear equation like

$$\begin{cases} x_1 & + & x_2^2 & = & 4 \\ \sin(x_1) & - & 4x_2 & = & 0 \end{cases}$$

2.1 Fixed-point iteration

Suppose you have a scientific calculator. You enter an arbitrary positive number, say, 2, and then press the square root button. After a while, you obtain 1. If instead you keep pressing cos, you obtain 0.7390851332, at least to the first 10 decimal places.

Definition 1 *A number r is a fixed point of the function g if $g(r) = r$.*

Definition 2 *Given any x_0, the following iteration is called fixed-point iteration*

$$x_{i+1} = g(x_i) \qquad \text{for } i = 0, 1, 2, 3, ... \tag{1}$$

If g is continuous and there is an x^* so that $x_i \to x^*$ when $i \to \infty$. Letting $i \to \infty$ in (1), one immediately obtain $x^* = g(x^*)$.

If the iteration converges, fixed-point iteration solves the fixed-point problem $x = g(x)$. So one may ask, can any equation $f(x) = 0$ be turned into a fixed point problem $x = g(x)$? The answer is yes, and in many different ways. For example, if we want to solve $x^3+x-1 = 0$, we can rewrite it as

$$x = 1 - x^3 = g_1(x), \tag{2}$$

or

$$x = \sqrt[3]{1 - x} = g_2(x), \tag{3}$$

or

$$x = \frac{1 + 2x^3}{1 + 3x^2} = g_3(x), \tag{4}$$

or

$$x = 2x - 1 + x^3 = g_4(x). \tag{5}$$

i	$x_i = g_1(x_{i-1})$	$x_i = g_2(x_{i-1})$	$x_i = g_3(x_{i-1})$	$x_i = g_4(x_{i-1})$
0	0.50000000	0.50000000	0.50000000	0.50000000
1	0.87500000	0.79370053	0.71428571	0.12500000
2	0.33007812	0.59088011	0.68317972	-0.74804688
3	0.96403747	0.74236393	0.68232842	-2.9146814
4	0.10405419	0.63631020	0.68232780	-31.590654
5	0.99887338	0.71380081	0.68232780	-31590.687
6	0.00337606	0.65900615	0.68232780	-3.1526605e13
7	0.99999996	0.69863261	0.68232780	-3.1335139e40
8	0.00000012	0.67044850	0.68232780	-3.0767690e121
9	1.00000000	0.69072912	0.68232780	-Inf
24	0.00000000	0.68227157	0.6823278	
25	1.00000000	0.68236807	0.6823278	

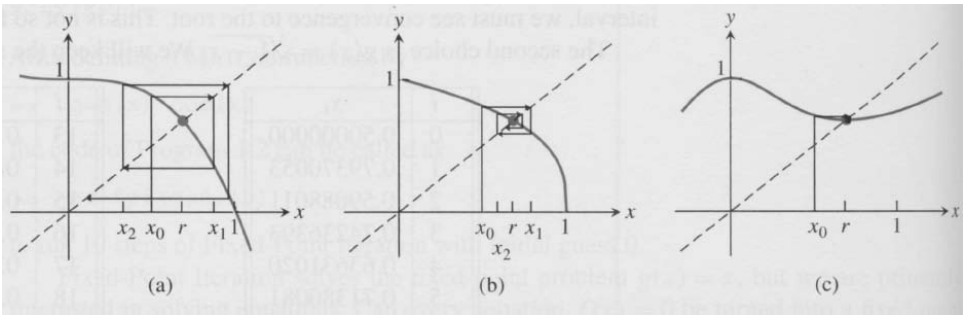

Figure 2: Geometry of fixed-point iteration. (a) $g(x) = 1 - x^3$. (b) $g(x) = (1 - x)^{1/3}$. (a) $g(x) = \frac{1+2x^3}{1+3x^2}$.

The following table shows the fixed-point iteration for the preceding four choices of $g(x)$. The starting point $x = 0.5$ is chosen somewhat arbitrary.

Figure 2 shows the three different $g(x)$ along with the first few steps of fixed-point iteration in each case. The fixed point r is the same for each $g(x)$. It is represented by the point where the graphs $y = g(x)$ and $y = x$ intersect.

The convergence properties of fixed-point iteration can be easily explained by a careful look at the algorithm in the simplest possible situation. Figure 3 shows fixed-point iteration for two linear functions (a) $g_1(x) = -\frac{3}{2}x + \frac{5}{2}$; (b) $g_2(x) = -\frac{1}{2}x + \frac{3}{2}$. In each case the fixed point is $x = 1$. Let $e_i = |x_i - r|$. Then

(a) $x_{i+1} - 1 = g_1(x_i) - 1 = g_1(x_i) - g_1(1) = -\frac{3}{2}(x_i - 1)$ which means $e_{i+1} = \frac{3}{2}e_i$. Hence the error becomes larger and larger and the iteration diverges.

(b) $x_{i+1} - 1 = g_2(x_i) - 1 = g_2(x_i) - g_2(1) = -\frac{1}{2}(x_i - 1)$ which means $e_{i+1} = \frac{1}{2}e_i$. Hence the error becomes smaller and smaller and the iteration converges.

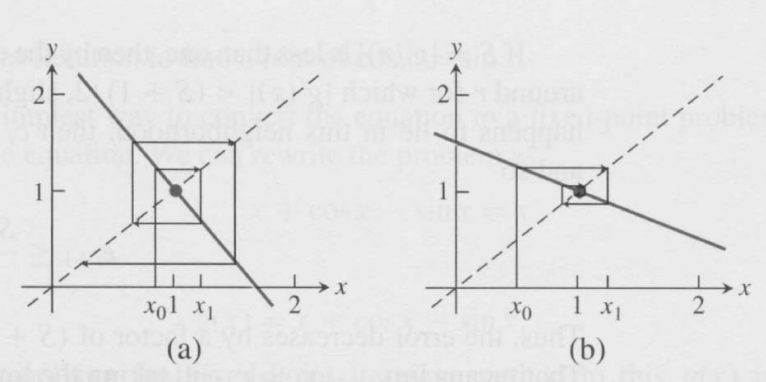

Figure 3: Fixed-point iteration. (a) $g_1(x) = -\frac{3}{2}x + \frac{5}{2}$. (b) $g_2(x) = -\frac{1}{2}x + \frac{3}{2}$.

Definition 3 *Let e_i denote the error at step i of an iterative method. If*

$$\lim_{i\to\infty} \frac{e_{i+1}}{e_i} = S < 1, \tag{6}$$

the method is said to obey linear convergence with rate S.

Theorem 1 *Assume that g is continuous differentiable, that $g(r) = r$, and that $S = |g'(r)| < 1$. Then fixed-point iteration converges linearly with rate S to the fixed point r for initial guesses sufficiently close to r.*

Proof: The iteration is $x_{i+1} = g(x_i)$. Hence $x_{i+1} - r = g(x_i) - r = g(x_i) - g(r) = g'(c_i)(x_i - r)$ where c_i is between x_i and r. We have used the mean value theorem in the last step. Let $e_i = |x_i - r|$. We have

$$e_{i+1} = |g'(c_i)|e_i.$$

If $|g'(r)| = S < 1$, then there is a δ so that whenever $|x - r| \le \delta$, $|g'(x)| < \frac{S+1}{2} < 1$. If x_k, for some integer k, happens to lie in this δ-neighborhood of r (call it D), then $c_k \in D$ and $e_{k+1} = |g'(c_k)|e_k < \frac{S+1}{2}e_k < e_k$. Hence $x_{k+1}, c_{k+1} \in D$ and $e_{k+2} = |g'(c_{k+1})|e_{k+1} < \frac{S+1}{2}e_{k+1} < e_{k+1}$. By induction, we know $x_i \in D$ for any $i \ge k$ and $e_{i+k} \le \frac{S+1}{2}e_{i+k-1} \le \left(\frac{S+1}{2}\right)^2 e_{i+k-2} \le \cdots \le \left(\frac{S+1}{2}\right)^i e_k \to 0$ as $i \to \infty$. Hence $x_{i+k} \to r$ as $i \to \infty$. So is c_{i+k}. Letting $i \to \infty$ in $\frac{e_{i+1}}{e_i} = |g'(c_i)|$, we proved (6).

Definition 4 *An iterative method is called locally convergent to r if the method converges to r for initial guesses sufficiently close to r.*

Example: Explain why the fixed-point iteration with $g(x) = \cos(x)$ converges.
Solution: The fixed-point r satisfies $r = \cos(r)$. $r \approx 0.74$ and $|g'(r)| = |-\sin(r)| < 1$. Hence by Theorem 1, the iteration is locally convergent. In fact, for any initial guess the iteration would converge to r.

2.2 Newton's method for a single nonlinear equation

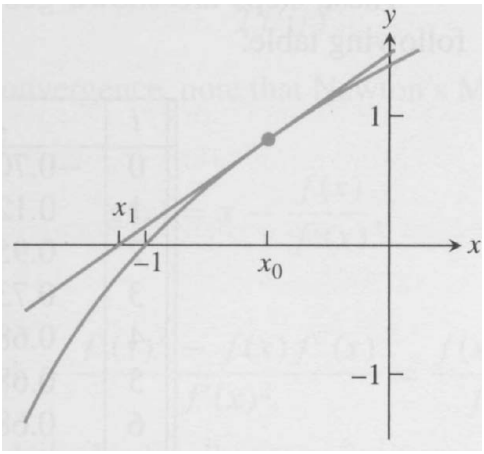

Figure 4: One step of Newton's method

Suppose x_* is the root of function $f(x)$. Since $0 = f(x_*) \approx f(x_n) + f'(x_n)(x_n - x_*)$, $x_* \approx x_n - f(x_n)/f'(x_n)$. Hence we obtain the so called Newton iteration with initial guess x_0:

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}. \tag{7}$$

Example Use Newton's method to solve for $x^3 + x - 1 = 0$ with initial guess $x_0 = -0.7$.
Solution: $x_{n+1} = x_n - \frac{x_n^3+x_n-1}{3x_n^2+1} = \frac{2x_n^3+1}{3x_n^2+1}$.

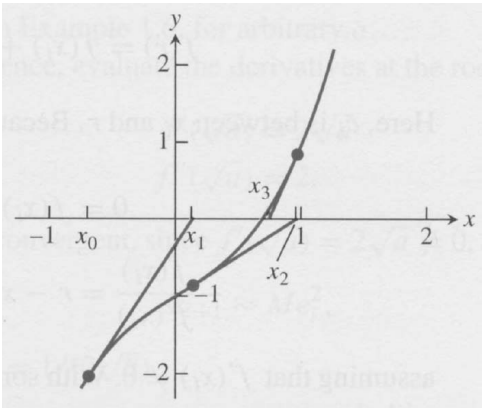

Figure 5: Three steps of using Newton's method to solve for $x^3 + x - 1 = 0$.

Inserting (7) from initial guess $x_0 = -0.7$ yields $x_1 \approx 0.1271$, $x_2 \approx 0.9577$. Further steps are given in the following table. After only 6 steps, the root is known to 8 correct digits.

i	x_i	$e_i = \lvert x_i - x_* \rvert$	e_i/e_{i-1}^2
0	−0.70000000	1.38232780	
1	0.12712551	0.55520230	0.2906
2	0.95767812	0.27535032	0.8933
3	0.73482779	0.05249999	0.6924
4	0.68459177	0.00226397	0.8214
5	0.68233217	0.00000437	0.8527
6	0.68232780	0.00000000	0.8541
7	0.68232780	0.00000000	

Definition 5 *Let e_i denotes the error after step i of an iterative method. The iteration is quadratically convergent if*

$$M = \lim_{i\to\infty} \frac{e_{i+1}}{e_i^2} < \infty.$$

Theorem 2 *Let f be twice continuous differentiable and $f(x_*) = 0$. If $f'(x_*) \neq 0$, then Newton's method is locally and quadratically convergent to x_*. The error e_i satisfies*

$$\lim_{i\to\infty} \frac{e_{i+1}}{e_i^2} = \frac{f''(x_*)}{2f'(x_*)}.$$

Proof: To prove local convergence, note that Newton's method is a particular form of fixed-point iteration, where

$$g(x) = x - \frac{f(x)}{f'(x)},$$

with derivative $g'(x) = 1 - \frac{f'(x)^2 - f(x)f''(x)}{f'(x)^2} = \frac{f(x)f''(x)}{f'(x)^2}$. Since $g'(x_*) = 0$, by Theorem 1, Newton's method is locally convergent.

Now, we prove the quadratic convergence. By Taylor expansion, $0 = f(x_*) = f(x_i) + f'(x_i)(x_* - x_i) + \frac{f''(c_i)}{2}(x_* - x_i)^2$. Hence

$$x_{i+1} - x_* = x_i - \frac{f(x_i)}{f'(x_i)} - x_* = (x_* - x_i)^2 \frac{f''(c_i)}{2f'(x_*)}.$$

So $e_{i+1} = e_i^2 \left| \frac{f''(c_i)}{2f'(x_*)} \right|$. Since c_i lies between x_* and x_i, it converges to x_* just as x_i does. We prove the quadratic convergence by letting i go to infinity.

2.3 Newton's method in $\mathbb{R}^n$

Definition 6 (Order of convergence) *A sequence of vector $x_0, x_1, ... \in \mathbb{R}^n$ converges locally to x_* with order r if*

$$\|x_{k+1} - x_*\| \leq C\|x_k - x_*\|^r$$

for all sufficient large k, and some positive constant C independent of k.

- If $r = 1$ and $C < 1$, we have so called linear convergence.
- $r > 1$ means superlinear convergence.
- In the special case of $r = 2$, it is called quadratic convergence. Note that if define $e_k = C\|x_k - x_*\|$, we get $e_{k+1} \leq e_k^2$.

Given $F : \mathbb{R}^n \to \mathbb{R}^n$, Newton's method for finding zeros of the nonlinear function $F(x) = 0$ is given by

$$x_{k+1} = x_k - F'(x_k)^{-1}F(x_k). \tag{8}$$

Given $f : \mathbb{R}^n \to \mathbb{R}$, Newton's method for finding a minimizer of f is given by

$$x_{k+1} = x_k - H(x_k)^{-1}g(x_k). \tag{9}$$

where $g = \nabla f \in \mathbb{R}^n$, $H = D^2 f \in \mathbb{R}^{n\times n}$.

The requirement on the initial guess x_0 that we need in order to guarantee the convergence of Newton's method is stated in the next section. Bad initial guess can make x_k diverge.

2.3.1 Proof of quadratic convergence

(This proof is not required for the exam. It is provided because I feel it is a good exercise for multivariable calculus and it makes use of the matrix norm that we have learned before.)

Assume F has a zero value at x_*, $F'(x_*)$ is invertable, F' is Lipschitz continuous in the region $\{x, \|x - x_*\| \leq a\}$ with a Lipschitz constant K. Define the solution neighbor $\Omega = \{x \in \mathbb{R}^n, \|x - x_*\| \leq \min(a, \frac{1}{2K\|F'(x_*)^{-1}\|})\}$ and $e_k = \|x_k - x_*\|$.

Lemma 1 $x_{k+1} - x_* = F'(x_k)^{-1}\int_0^1 [F'(x_k + t(x_* - x_k)) - F'(x_k)]dt(x_* - x_k)$

Proof: Because $\int_0^1 F'(x_k + t(x_* - x_k))(x_* - x_k)dt = F(x_*) - F(x_k) = -F(x_k)$, the right hand side equals to

$$F'(x_k)^{-1}\left(-F(x_k) - F'(x_k)(x_* - x_k)\right) = -F'(x_k)^{-1}F(x_k) - (x_* - x_k). \qquad \square$$

Lemma 2 *If $x \in \Omega$, then $\|F'(x)^{-1}\| \leq 2\|F'(x_*)^{-1}\|$.*

Proof: The idea is to use Theorem 1 (If $\|A\| < 1$, then $(I-A)^{-1}$ exists and $\|(I-A)^{-1}\| \leq \frac{1}{1-\|A\|}$).

$$\|I - F'(x)F'(x_*)^{-1}\| = \| \left(F'(x_*) - F'(x)\right) F'(x_*)^{-1}\| \leq K\|x - x_*\|\|F'(x_*)^{-1}\| \leq \frac{1}{2}.$$

So, $F'(x)F'(x_*)^{-1}$ is invertible and $\| \left(F'(x)F'(x_*)^{-1}\right)^{-1} \| \leq \frac{1}{1-\frac{1}{2}} = 2$. In particular, $F'(x)$ is non-singular. So,

$$\|F'(x)^{-1}\| = \|F'(x_*)^{-1}F'(x_*)F'(x)^{-1}\| \leq \|F'(x_*)^{-1}\|2. \qquad \square$$

Theorem 3 *If $x_k \in \Omega$, then $e_{k+1} \leq K\|F'(x_*)^{-1}\|e_k^2$ and $x_{k+1} \in \Omega$.*

Proof: By Lemma 1 and then Lemma 2,

$$e_{k+1} \leq \|F'(x_k)^{-1}\| \int_0^1 Ktdt\|x_* - x_k\|^2 \leq K2\|F'(x_*)^{-1}\|\frac{1}{2}e_k^2.$$

If $x_k \in \Omega$, $\|x_{k+1} - x_*\| \leq \frac{1}{4K\|F'(x_*)^{-1}\|}$ and so $x_{k+1} \in \Omega$. $\square$

2.4 Secant method for a single nonlinear equation

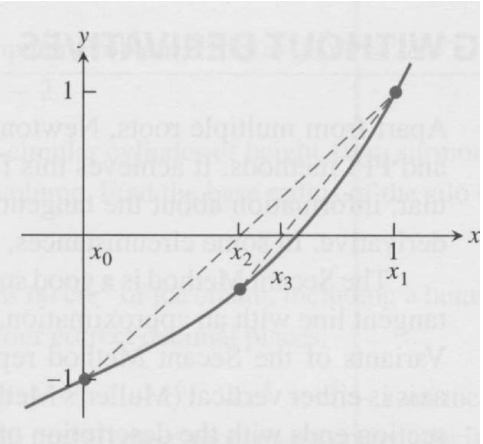

Figure 6: Secant method.

Given a nonlinear function f, suppose we want to find x_* so that $f(x_*) = 0$. The secant method is as follows: Given x_{k-1} and x_k, draw a secant line through $(x_{k-1}, f(x_{k-1}))$, $(x_k, f(x_k))$ which intersect the x-axis at x_{k+1}

$$x_{k+1} = x_k - \frac{f(x_k)}{f(x_k) - f(x_{k-1})}(x_k - x_{k-1}). \tag{10}$$

Because $f(x_*) = 0$, from (10), we get

$$\begin{aligned} x_{k+1} - x_* &= x_k - x_* - \frac{f(x_k) - f(x_*)}{f(x_k) - f(x_{k-1})}(x_k - x_{k-1}) \\ &= x_k - x_* - \frac{f(x_k) - f(x_*)}{f[x_{k-1}, x_k]} = (x_k - x_*)\left(1 - \frac{f[x_k, x_*]}{f[x_{k-1}, x_k]}\right) \\ &= (x_k - x_*)(x_{k-1} - x_*)\frac{f[x_{k-1}, x_k, x_*]}{f[x_{k-1}, x_k]} \\ &= (x_k - x_*)(x_{k-1} - x_*)\frac{f''(\xi_2)}{2f'(\xi_1)} \end{aligned}$$

where for the proof, we have assume $f \in C^2$ and we further assume x_* is a simple root of f. Hence in a sufficient small neighborhood of x_*, there is a $M > 0$ so that

$$\left|\frac{f''(\xi_2)}{2f'(\xi_1)}\right| \leq M.$$

Define $e_k = M\,|x_k - x_*|$, then

$$e_{k+1} \leq e_k e_{k-1} \Longrightarrow \eta_{k+1} \leq \eta_k + \eta_{k-1} \text{ with } \eta_k = \ln e_k$$

One can then prove that $\eta_k \leq c_0 r^k$ with $r = \frac{1+\sqrt{5}}{2}$ by induction where c_0 depends on e_0 and e_1. Note that r is the positive root of the characteristic equation $\eta^2 = \eta + 1$. Hence $e_k \leq \alpha^{r^k}$ with $\alpha = e^{c_0} < 1$ if $e_0 < 1$, $e_1 < 1$. Note that if $e_k = \alpha^{r^k}$, then $e_{k+1} = \alpha^{r^k r} = \left(\alpha^{r^k}\right)^r = (e_k)^r$. So, we sometimes say that the secant method has order of convergence $\frac{1+\sqrt{5}}{2}$.

Remark: Note that $\frac{f(x_k)}{f(x_k)-f(x_{k-1})}(x_k - x_{k-1})$ can be viewed as an approximation of $f(x_k)/f'(x_k)$ used in the Newton method which has order of convergence 2. We will go back to this point when we talk about quasi-Newton in $\mathbb{R}^n$.

2.5 Review of Numerical Analysis I: Newton's formula for interpolation

Divided difference

$$\begin{aligned} f[x_0] &= y_0 \\ f[x_0, x_1] &= \frac{f[x_1] - f[x_0]}{x_1 - x_0} \\ &\cdots \\ f[x_0, x_1, ...x_k] &= \frac{f[x_1, ..., x_k] - f[x_0, ..., x_{k-1}]}{x_k - x_0} \end{aligned}$$

Theorem 4 *Given the value of f at $x_0, ..., x_n$. We want to find an nth degree polynomial so that it equals $f(x_i)$ at point x_i for $i = 0, ..., n$. This interpolation can be written as*

$$p_n(x) = d_0 + d_1(x - x_0) + ... + d_n(x - x_0)...(x - x_{n-1}) \tag{11}$$

where $d_k = f[x_0, ..., x_k]$.

Remark: The above formula is the Newton's formula for the interpolation. You should have learned the Lagrange polynomial for interpolation. They are the same thing but written in different form.

By the above theorem, when we want to add one more interpolation point x_{n+1}, we only need to add one more term. This is the advantage of Newton's formula over Lagrange formula.

Proof: By induction. It is true when $n = 1$. Suppose it is true for $n - 1$, namely $p_{n-1}(x) = f[x_0] + f[x_0, x_1](x - x_0) + ... + f[x_0, ..., x_{n-1}](x - x_0)...(x - x_{n-2})$ interpolate at $x_0, ..., x_{n-1}$. Then $q_{n-1}(x) = f[x_1] + f[x_1, x_2](x - x_1) + ... + f[x_1, ..., x_n](x - x_1)...(x - x_{n-1})$ interpolate at $x_1, ..., x_n$. Define

$$q_n = \frac{(x - x_0)q_{n-1}(x) - (x - x_n)p_{n-1}(x)}{x_n - x_0}.$$

Then q_n interpolate f and $x_0, ..., x_n$. By the uniqueness,

$$q_n(x) = d_0 + d_1(x - x_0) + ... + d_n(x - x_0)...(x - x_{n-1})$$

By comparing the leading coefficient, we get

$$d_n = \frac{f[x_1, ..., x_n] - f[x_0, ..., x_{n-1}]}{x_n - x_0}. \qquad \square$$

Theorem 5 *Let x_0,...,x_n be $n + 1$ distinct points on $[a, b]$. If $f \in C^{n+1}([a, b])$ and p_n its interpolation at $x_0, ..., x_n$. Then for any x, there is a ξ depends on x so that*

$$f(x) - p_n(x) = \frac{f^{n+1}(\xi)}{(n + 1)!}(x - x_0)...(x - x_n). \tag{12}$$

Proof: Let $g(t) = f(t) - p_n(t) - \lambda\omega(t)$ where $\omega(t) = \prod_{i=0}^{n}(x - x_i)$ and λ is a constant so that $g(x) = 0$. So, we are left to prove that $\lambda = \frac{f^{n+1}(\xi)}{(n+1)!}$ for some ξ depends on x.

Note that $g(t) = 0$ at $t = x_0, ..., x_n, x$ which are $n + 2$ points. By Rolle's theorem, there is a ξ so that

$$g^{(n+1)}(\xi) = 0$$

which implies $f^{(n+1)}(\xi) - \lambda(n + 1)! = 0$. $\square$

Theorem 6 *If $f \in C^{n+1}([a, b])$ and $x \neq x_i$, then*

$$f[x_0, ..., x_n, x] = \frac{f^{(n+1)}(\xi)}{(n + 1)!}$$

Proof: Let $p_{n+1} \in \boldsymbol{P}_{n+1}$ which interpolate f at $x_0, ..., x_n, x$. Then $p_{n+1}(t) = p_n(t) + f[x_0, ..., x_n, x](t - x_0)...(t - x_n)$. Let $t = x$ and then use Theorem 5. $\square$

2.6 Quasi-Newton method in $\mathbb{R}^n$

The computation of $F'(x_k)$ (8) can be expensive. One can try to use finite difference to replace the derivative. An example that we have already see is the secant method. We will use matrix B_k to approximate $F'(x_k)$. Since

$$F'(x_{k+1})(x_{k+1} - x_k) \approx F(x_{k+1}) - F(x_k),$$

we will enforce the following secant condition (also called quasi-Newton condition)

$$B_{k+1}(x_{k+1} - x_k) = F(x_{k+1}) - F(x_k). \tag{13}$$

If we define $s_k = x_{k+1} - x_k$, $v_k = F(x_{k+1}) - F(x_k)$, we have

$$B_{k+1}s_k = v_k. \tag{14}$$

The secant condition is just a system of n equations and therefore is not sufficient to uniquely determine the n^2 unknowns from B_{k+1}. It is necessary to enforce further conditions on B_{k+1}.

2.6.1 Broyden's method for system of nonlinear equations

We would like to update B_{k+1} by using B_k from the last iteration. Some formula can be worked out if we consider the following rank one update:

$$B_{k+1} = B_k + u_k w_k^\top. \tag{15}$$

From secant condition, we have

$$v_k = B_{k+1}s_k = (B_k + u_k w_k^\top)s_k$$

which implies

$$u_k = \frac{v_k - B_k s_k}{w_k^\top s_k}. \tag{16}$$

One example of w_k that ensure $w_k^\top s_k \neq 0$ is $w_k = s_k$. This leads to the following Broyden's method:

Broyden's method for solving system of nonlinear equations $F(x) = 0$ is given by

- Solve s_k so that $B_k s_k = -F(x_k)$
- $x_{k+1} = x_k + s_k$, $v_k = F(x_{k+1}) - F(x_k)$
- $B_{k+1} = B_k + \frac{(v_k - B_k s_k)s_k^\top}{s_k^\top s_k}$

2.6.2 Convergence proof of Broyden's method

(This proof is not required for the exam.)

We will use L^2 matrix norm in the following discussion because it involves projection matrix.

A projection matrix P is an $n \times n$ square matrix that gives a vector space projection from $\mathbb{R}^n$ to a subspace W. A square matrix P is a projection matrix if and only if $P = P^2$. Since the length of a vector has been reduced after projection, $\|P\| \leq 1$. On the other hand, any vector w in W is preserved by the projection matrix P since $Pw = w$. Consequently, a projection matrix P has 2-norm equal to one, unless $P = 0$.

Assume F has a zero value at x_*, $F'(x_*)$ is invertable, F' is Lipschitz continuous in the region $\{x, \|x - x_*\| \leq a\}$ with a Lipschitz constant K. Let $A = F'(x_*)$ and $e_k = \|x_k - x_*\|$. Define the solution neighbor $\Omega = \{(x, B), \|x - x_*\| \leq a, \|B - A\| + \frac{3K}{2}\|x - x_*\| \leq \frac{1}{3\|A^{-1}\|}\}$.

To prove the convergence of Broyden's method, we need the following lemma, which is a immediate consequence of Theorem 1

Lemma 3 *If A^{-1} exist, $\|A^{-1}\|\|A - B\| < 1$, then B^{-1} exists and*

$$\|B^{-1}\| \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|A - B\|}$$

Proof: By Theorem 1, because $\|I - A^{-1}B\| < \|A^{-1}\|\|A - B\| < 1$, $I - (I - A^{-1}B) = A^{-1}B$ is invertible, which implies B is non-singular. And

$$\| \left(A^{-1}B\right)^{-1} \| \leq \frac{1}{1 - \|I - A^{-1}B\|} \leq \frac{1}{1 - \|A^{-1}\|\|A - B\|}.$$

Then the result follows from $\|B^{-1}\| \leq \|B^{-1}A\|\|A^{-1}\|$. $\square$

Theorem 7 *If $(x_k, B_k) \in \Omega$, then $(x_{k+1}, B_{k+1}) \in \Omega$ and $e_{k+1} < \frac{1}{2}e_k$.*

Proof: The proof contain the following 6 steps:

(a) Claim that

$$x_{k+1} - x_* = B_k^{-1}\left(B_k - A - \int_0^1 [F'(x_k + t(x_* - x_k)) - F'(x_*)]dt\right)(x_k - x_*). \quad (17)$$

The proof is very simple, because the right hand side equals $(x_k - x_*) - B_k^{-1}\left(F(x_k) - F(x_*)\right) = (x_k - x_*) - B_k^{-1}F(x_k)$.

(b) We are going to use the Lipschitz condition on the integral. If $(x_k, B_k) \in \Omega$, then by (a) and Lemma 3

$$\begin{aligned} e_{k+1} &\leq \|B_k^{-1}\|\left(\|B_k - A\| + \frac{K}{2}e_k\right)e_k \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\|\|A - B_k\|}\left(\|B_k - A\| + \frac{K}{2}e_k\right)e_k \\ &\leq \frac{\|A^{-1}\|}{1 - \frac{1}{3}}\frac{1}{3\|A^{-1}\|}e_k = \frac{1}{2}e_k. \end{aligned}$$

We are almost done, except we need to estimate $\|B_{k+1} - A\|$ in order to finish the induction. That's why the next few steps.

(c) One can easily verify that the last equation of Broyden's method can be immediately rewritten as

$$B_{k+1} - A = (B_k - A)\left(I - \frac{s_k s_k^\top}{s_k^\top s_k}\right) + \frac{(v_k - As_k)s_k^\top}{s_k^\top s_k}$$

(d) One recognize that $\frac{s_k s_k^\top}{s_k^\top s_k}$ and $I - \frac{s_k s_k^\top}{s_k^\top s_k}$ are simply projection matrices. Therefore $\left\|\frac{s_k s_k^\top}{s_k^\top s_k}\right\| \leq 1$ and $\left\|I - \frac{s_k s_k^\top}{s_k^\top s_k}\right\| \leq 1$

(e) Simply plugging in the definition, one can verify that

$$v_k - As_k = \int_0^1 [F'(x_k + t(x_{k+1} - x_k)) - F'(x_*)]dt s_k.$$

(f) If $\|x_k - x_*\| \leq a$ and $\|x_{k+1} - x_*\| \leq a$, plugging (e) into (c), using (d) and the Lipschitz condition we get

$$\begin{aligned}\|B_{k+1} - A\| &\leq \|B_k - A\| + \int_0^1 K\|x_k + t(x_{k+1} - x_k) - x_*\|dt \\ &\leq \|B_k - A\| + \frac{K}{2}\left(\|x_k - x_*\| + \|x_{k+1} - x_*\|\right)\end{aligned}$$

Here we have used the inequality $\int_0^1 \|a + tb\|dt \leq \frac{1}{2}\left(\|a\| + \|a + b\|\right)$ which follows from the fact that the function $t \mapsto f(t) = \|a + tb\|$ is a convex function. Recall that $f(t)$ convex function means $f(\theta t_1 + (1-\theta)t_2) \leq \theta f(t_1) + (1-\theta)f(t_2)$ for any $\theta \in [0, 1]$ and any t_1, t_2. So if $\|x_{k+1} - x_*\| \leq \frac{1}{2}\|x_k - x_*\|$, from the above inequality, we can easily get

$$\|B_{k+1} - A\| + \frac{3K}{2}\|x_{k+1} - x_*\| \leq \|B_k - A\| + \frac{3K}{2}\|x_k - x_*\|.$$

By (b) and the above inequality, $(x_{k+1}, B_{k+1}) \in \Omega$. This finishes the proof. $\square$

2.6.3 Symmetric Broyden's method for minimization problem

If we want to solve $x = \text{argmin} f(x)$ with $f : \mathbb{R}^n \to \mathbb{R}$, since the Hessian is symmetric, it seems reasonable to ask that each B_{k+1} be symmetric as well. If we still use rank one update (15), and of B_k is symmetric, we would require $w_k = \gamma_k u_k$ with γ_k a scalar. From (16), we know u_k (and hence w_k) is in the direction of $v_k - B_k s_k$. We can then obtain the unique symmetric rank one update

$$B_{k+1} = B_k + \frac{(v_k - B_k s_k)(v_k - B_k s_k)^\top}{(v_k - B_k s_k)^\top s_k}.$$

Symmetric Broyden's method for minimization problem $x = \text{argmin} f(x)$ is given by

- Solve p_k so that $B_k p_k = -\nabla f(x_k)$.
- Perform a line search along the quasi-Newton direction, i.e., compute the α_k that minimizes $f(x_k + \alpha p_k)$ over $\alpha > 0$.
- $x_{k+1} = x_k + \alpha_k p_k$
- $s_k = x_{k+1} - x_k = \alpha_k p_k$, $v_k = \nabla f(x_{k+1}) - \nabla f(x_k)$
- $B_{k+1} = B_k + \frac{(v_k - B_k s_k)(v_k - B_k s_k)^\top}{(v_k - B_k s_k)^\top s_k}$

Remark: It is not obvious what is the best method for minimizing $f(x_k + \alpha p_k)$ for general f. Perhaps we could use a *one-dimensional* Newton or secant method.

The number of iterations required for convergence depends on the problem and it is common to restart the problem after say every n or $2n$ iterations.

If we want B_k not just symmetric, but also symmetric positive definite, we can go to rank two update, and have the so called BFGS and DFP formula. See S. G. Nash and A. Sofer, "Linear and nonlinear programming" (1996) for details.

2.7 Homework II

1) Show that the Newton's method for the function $f(x) = x^r - a$, $x > 0$, where $r > 1$ and $a > 0$, converges globally to $b = a^{\frac{1}{r}} > 0$ as long as the initial guess $x_0 \geq b$. [Hint: You can use the fact that a bounded monotone sequence converge to a finite number. The r is not necessarily an integer.]

2) Consider the following iteration

$$x_{k+1} = G(x_k) \tag{18}$$

where $x_k \in \mathbb{R}^n$ and $G(x_k) \in \mathbb{R}^n$. When it converges, we obtain x that satisfies $x = G(x)$. If we want to solve $H(x) = 0$, we can write it as $x = \alpha H(x) + x$ for some scalar or matrix α and use the above method to try to find the solution. The issue is whether such iteration will converge.

   a) Show that the Newton's method to solve $F(x) = 0$ can be written into (18) and figure out the specific form of G.

   b) Prove the following fact: If the mapping $G : \mathbb{R}^n \to \mathbb{R}^n$ is contractive, i.e., there is a constant $\lambda < 1$ such that

$$\|G(x) - G(y)\| \leq \lambda \|x - y\|$$

   for some vector norm $\|\cdot\|$, then there is an x and only one x such that

$$\lim_{k \to \infty} \|x_k - x\| = 0$$

   and this x satisfies $G(x) = x$.

3) Verify the following formula for rank one updated matrix

$$\left(A + uv^\top\right)^{-1} = A^{-1} - \frac{A^{-1}uv^\top A^{-1}}{1 + v^\top A^{-1}u}.$$

4) Let us introduce the Golden section search method to find the minimizer of $f(x)$.

   Suppose we are given x_1, x_3 and x_2 so that

$$x_1 < x_3 < x_2 \qquad \text{and } f(x_1) > f(x_3) < f(x_2) \tag{19}$$

   and suppose $x_3 - x_1 < x_2 - x_3$. Certainly, a minimizer exists in the interval $[x_1, x_2]$. We want to shrink this interval of uncertainty. So we take an $x_4 \in (x_3, x_2)$ (which is the longer interval) and according to the size of $f(x_4)$ we choose either $[x_1, x_4]$ (when $f(x_3) < f(x_4)$) or $[x_3, x_2]$ (when $f(x_3) > f(x_4)$) as the new "existence" interval. We should choose x_4 so that the two possible uncertain intervals have the same length:

$$x_4 - x_1 = x_2 - x_3.$$

This requirement is based on the following intuition: If they are not, a run of "bad luck" could lead to the wider interval being used many times, thus slowing down the rate of convergence.

Then we rename the three points in this new interval of uncertainty and repeat.

The golden section search chooses the spacing between x_1, x_3, x_2 in such a way that these three points have the same proportion of spacing as the subsequent triple x_1, x_3, x_4 or x_3, x_4, x_2. By maintaining the same proportion of spacing throughout the algorithm, we avoid a situation in which x_3 is very close to x_1 or x_2, and guarantee that the interval width shrinks by the same constant proportion in each step.

Based on the above requirements, figure out given x_1 and x_2, how the point x_3 and x_4 should be located. Then prove that the length of the interval of uncertainty decreases at a constant rate $\frac{\sqrt{5}-1}{2} \approx 0.618$.

2.8 Computer project II

a) Mandelbrot set

First, consider the region in the complex plane consisting of the values z_0 for which the sequence defined by

$$z_{n+1} = z_n^2$$

remains bounded when $n \to \infty$. It is easy to see that this set for z_0 is simply the unit disc. The boundary of the unit disc is the unit circle. There is nothing very difficult or exciting here.

The Mandelbrot set is defined as the region in the complex plane consisting of the values z_0 for which the sequence defined by

$$z_{n+1} = z_n^2 + z_0, \qquad n = 0, 1, 2, ...$$

remains bounded when $n \to \infty$.

Draw the Mandelbrot set in the region $[-2, 1] \times [-1.5, 1.5]$ and then in the region $[-0.4, 0.2] \times [-1.2, -0.6]$. You may need to use Matlab function `imagesc.m`. You may partition the domain into 1000×1000 meshes and iterate 300 times at each point. If the $|z_{300}| > 2$, assume it diverges.

b) Many interesting physical systems can be expressed as a system of partial differential equations. The following equation can be used to model the the heat transport in a thin bar $\Omega = [0, 1]$ which radiates heat and at the same time is heated by an external heat source f:

$$-k\frac{d^2u}{dx^2} + \sigma u^4 = f + \sigma u_{\text{ext}}^4 \qquad \text{in } \Omega \tag{20}$$

$$u = g \qquad \text{on } \partial\Omega. \tag{21}$$

Here k is the constant heat conductivity, σ is the Stefan-Boltzmann constant and g is the temperature at end points of the bar. u_{ext} is the temperature of the surrounding space (absolute temperature in Kelvin).

Take $u_{\text{ext}} = 0$, $k = 1$ and $\sigma = 1$. Find f and g so that the exact solution is $u(x) = \sin(\pi x)$.

Discretize it like in the first project and obtain a system of equations $F(x) = 0$. Solve it by Newton's method. The initial guess can be the zero vector. Like in the first project, solve (20)+(21) numerically with $n + 1 = 10, 20, 40, 80$. Plot the differen between the numerical solution and the exact solution. Also show the loglog plot of the maximum error with respect to $h = \frac{1}{n+1}$. When you turn in the hard copy of your solution, also include your Matlab code.

2.9 Tutorial Question Set: III

1. Which of the following three fixed-point iterations converges to $\sqrt{5}$? Rank the ones that converge from fastest to slowest.

   (a) $x \to f_1(x) = \frac{4}{5}x + \frac{1}{x}$.

   (b) $x \to f_2(x) = \frac{x}{2} + \frac{5}{2x}$.

   (c) $x \to f_3(x) = \frac{x+5}{x+1}$.

2. (i) Consider fixed-point iteration with $g(x) = x - x^3$. (a) Show that $x = 0$ is the only fixed-point. (b) Show that if $0 < x_0 < 1$, then $x_0 > x_1 > x_2 ... > 0$. (c) Show $g'(0) = 1$ and in this case fixed-point iteration converges to the fixed point $r = 0$.

   (ii) Consider fixed-point iteration with $g(x) = x + x^3$. (a) Show that $x = 0$ is the only fixed-point. (b) Show that if $0 < x_0 < 1$, then $x_0 < x_1 < x_2 ....$ (c) Show $g'(0) = 1$ and in this case fixed-point iteration fails to converge.

3. Apply Newton's method to find a root of $f(x) = 4x^4 - 6x^2 - \frac{11}{4}$ with initial guess $x = 1/2$. Show that the method fails to find a root.

4. Use Newton's method to find a root of $f(x) = (x-1)^2$ and show that the method is linear convergent.

5. Prove that if f is $m+1$-times continuously differentiable on $[a, b]$, which contains a root r of multiplicity $m > 1$, then the modified Newton's method

$$x_{i+1} = x_i - \frac{mf(x_i)}{f'(x_i)} \tag{22}$$

   converges locally and quadratically to r.

2.10 Tutorial Question Set: IV

1. Show that $\lim_{n\to\infty} \underbrace{\sqrt{2+\sqrt{2+\cdots+\sqrt{2}}}}_{\text{n square roots}} = 2$

2. Prove that Newton's method applied to the vector equation $F(x) = Ax - b = 0 \in \mathbb{R}^n$ converges in one step for any initial guess x_0.

3. Let $D \subset \mathbb{R}$ be an open interval and let $f : D \to D$ be m times continuously differentiable. Under the assumption that the sequence $x_{n+1} = f(x_n)$ converges to some $x \in D$ with $f'(x) = f''(x) = \cdots = f^{(m-1)}(x) = 0$, show that the convergence is of order m

4. Show that $x_{n+1} = \frac{x_n(x_n^2+3a)}{3x_n^2+a}$ is a method of order three for computing the square root of a positive a.

5. Show that for a nonsingular $n \times n$ matrix A, the sequence

$$A_{n+1} = A_n(2I - AA_n)$$

converges quadratically to the inverse A^{-1} provided that A_0 is sufficiently close to A^{-1}.

2.11 Tutorial Question Set: V

1. Derive the Newton's method for

$$\begin{cases} x_1 x_2 &= x_3^2 + 1 \\ x_1 x_2 x_3 + x_2^2 &= x_1^2 + 2 \\ e^{x_1} + x_3 &= e^{x_2} + 3 \end{cases}$$

and then carry out one iteration of Newton's method using calculator, staring with $x^0 = (1, 1, 1)^\top$.

2. Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$. Prove that if $|a_{ij}| \le M$ for any i and j, then $\|A\|_p \le nM$ for any $p \ge 1$. Then, for the F used in Question 1, Find a function $K = K(R)$ so that

$$\|F'(x) - F'(y)\|_2 \le K(R)\|x - y\|_2$$

whenever $x, y \in \Omega := \{x | \|x\|_2 \le R\}$.

3. We know the mean value theorem says that for any scalar function $f \in C^1([a, b])$, for any $x, y \in [a, b]$, there is a $\theta \in [0, 1]$ so that

$$f(x) - f(y) = f'((1 - \theta)x + \theta y)(x - y).$$

Explain why the mean value theorem cannot be extended to the case when f is a vector-valued function. However, for $F : \mathbb{R}^n \to \mathbb{R}^n$, the following inequality is true

$$\|F(x) - F(y)\| \le \left(\sup_{\theta\in[0,1]} \|F'((1 - \theta)x + \theta y)\| \right) \|x - y\|. \tag{23}$$

Prove the above inequality for the special case when $F = (f_1(x_1), f_2(x_2))^\top$ which maps $x = (x_1, x_2)^\top$ to $(f_1, f_2)^\top \in \mathbb{R}^2$.

4. Show that $\lim_{x\to 0} xe^{-x} = 0$. Prove that the Newton method for solving $xe^{-x} = 0$ will diverge if $x_0 > 1$.

5. Let $f \in C^2([a, b])$ and x_* is a simple root of $f(x) = 0$. Consider the iteration

$$y = x_k - \frac{f(x_k)}{f'(x_k)} \tag{24}$$

$$x_{k+1} = y - \frac{f(y)}{f'(x_k)}. \tag{25}$$

Show that $\lim_{k\to\infty} \frac{e_{k+1}}{e_k^3} = S < \infty$.

2.12 Lab tutorial: II

Let us introduce the bisection method to find the root of a function.

The basic idea is that if $f(a)$ and $f(b)$ have different sign, then there is a root between a and b. So, the algorithm is as follows:

Given initial interval $[a, b]$ such that $f(a)f(b) < 0$
while $(b - a)/2 >$ Tolerance
    $c = (a + b)/2$
    if $f(c) = 0$, **stop**, **end**
    if $f(a)f(c) < 0$
        $b = c$
    else
        $a = c$
    end
end
The final interval $[a, b]$ contains a root. The approximate root is $(a + b)/2$.

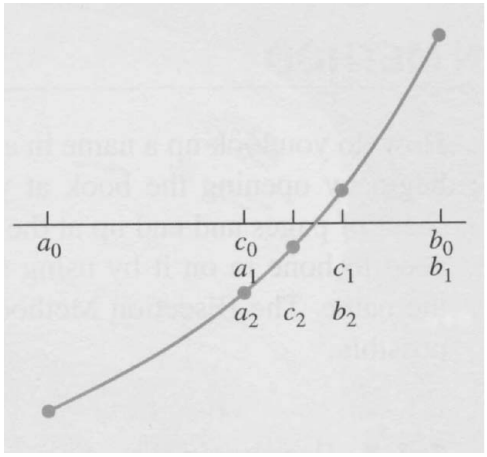

Figure 7: The bisection method.

Use Matlab and the bisection method to find the root of the function $f(x) = x^3 + x - 1$ on the interval $[0, 1]$. The error is required to be less than 0.001. (How would you guarantee the error is less than 0.001, if you do not know the exact solution?)

3 Monte Carlo methods

This chapter is based on the book "Explorations in Monte Carlo Methods" by Shonkwiler and Mendivil. A more advanced reference for this subject is "Monte Carlo strategies in scientific computing" by Jun S. Liu.

3.1 Some basic probability

An event E in a probabilistic experiment is some designed set of outcomes. The set of all possible outcomes is called a sample space, denoted as Ω.

The probability of E, written as $P(E)$ is the fraction of times E occurs in an infinitely long sequence of trials of the experiment. Mathematically, E is a measurable subset of Ω and P is a function on those sets. $P(E_1 \cup E_2 \cup \cdots)$ is the probability that at least one events E_1, E_2, ... occurs and $P(E_1 \cap E_2 \cap \cdots)$ is the probability that all the events E_1, E_2, ... occur. The following simple axiom completely determine how probabilities operate:

a) $P(\emptyset) = 0$ and $P(\Omega) = 1$.

b) If $E \subset \Omega$, then $P(E) \geq 0$.

c) If E_1, E_2, ... are disjoint events (i.e., $E_i \cap E_j = \emptyset$ whenever $i \neq j$), then $P(\cup_{i=1}^{\infty} E_i) = \sum_{i=1}^{\infty} P(E_i)$.

The conditional probability of A given B (denoted by $P(A|B)$) is *defined* as $P(A|B) = P(A \cap B)/P(B)$. Very often people write $P(AB)$ to mean $P(A \cap B)$. The interpretation can be easily seen from a picture like ⓪: $P(A \cap B)/P(B)$ is the probability of event A happens when event B already happens; Or equivalently the probability of A given the outcome must be restricted to B. Obviously, we have the Bayes's formula

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

We say two events A and B are mutually independent if any only if $P(A \cap B) = P(A)P(B)$.

A discrete random variable (r.v.) is one that can assume only finitely many or countably infinitely many outcomes. A continuous r.v. X is one for which $P(X = x) = 0$ for every real value x. We can say that a (one-dimensional real) r.v. is a mapping form the outcome space Ω to the real line $\mathbb{R}$.

The cumulative distribution function (cdf) is $F(x) = P(X \leq x)$. $F(x)$ is right continuous, monotonly nondecreasing and its value is between 0 and 1. Obviously $\lim_{x \to -\infty} F(x) = 0$ and $\lim_{x \to +\infty} F(x) = 1$.

For a discrete r.v., the probability density function (pdf) is $f(x) = P(X = x) = \lim_{\varepsilon \to 0} (F(x) - F(x - \varepsilon))$.

$$P(X \in A) = \sum_{x \in A} f(x).$$

For a continuous r.v., the probability density (pdf) is $f(x) = \frac{d}{dx}F(x)$.

$$P(X \in A) = \int_A f(x)dx = \int_A dF(x).$$

For example, we sat $X \sim U(a, b)$ (the uniform distribution on interval $[a, b]$) if its pdf is $f = \frac{1}{b-a}$. We say $X \sim N(\mu, \sigma^2)$ (the normal distribution), if its pdf is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{|x-\mu|^2}{2\sigma^2}}$.

Example: If X has the standard normal distribution ($N(0, 1)$), what the pdf of X^2?
Solution: $P(X^2 \leq x) = \sqrt{1}\sqrt{(2\pi)} \int_{-\sqrt{x}}^{\sqrt{x}} e^{-t^2/2}dt = \frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-t^2/2}dt$ for $x > 0$. So, the pdf of random variable X^2 is $\frac{d}{dx}\left(\frac{2}{\sqrt{2\pi}} \int_0^{\sqrt{x}} e^{-t^2/2}dt\right) = \frac{2}{\sqrt{2\pi}} e^{-x/2} \frac{1}{2\sqrt{x}} = \frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{x}} e^{-x/2}$ when $x > 0$ and 0 when $x \leq 0$. $\square$

If X is a discrete r.v. with pdf function $f(x)$. Then

$$E(X) = \sum_{x \in \Omega} xf(x),$$

where Ω is the set of all possible outcomes. More generally, if $r(X)$ is a function depends on X, then define the expected value of $r(X)$ as follows:

$$E\left(r(X)\right) = \sum_{x \in \Omega} r(x)f(x).$$

$E\left(r(X)\right)$ can be estimated empirically. If X_1, ..., X_n is a sequence trials each with outcome $X = X_i$, then

$$E\left(r(X)\right) \approx \frac{1}{n} \sum_{i=1}^{n} r(X_i).$$

The above approximation is valid for the following reasons: Let n trials be carried out and suppose n_1 outcomes were x_1, n_2 outcomes were x_1 and so on until n_k outcomes were x_k. (Suppose there are only k different kinds of outcomes, i.e., $|\Omega| = k$.) Then

$$\frac{1}{n} \sum_{i=1}^{n} r(X_i) = \sum_{j=1}^{k} r(x_j) \frac{n_j}{n} \approx \sum_{j=1}^{k} r(x_j)P(X = x_j).$$

If X is a continuous r.v. with pdf function $f(x)$. Then

$$E(X) = \int_{\mathbb{R}} xf(x)dx = \int_{\mathbb{R}} xdF(x).$$

More generally, if $r(X)$ is a function depends on X, then

$$E\left(r(X)\right) = \int_{\mathbb{R}} r(x)f(x)dx = \int_{\mathbb{R}} r(x)dF(x).$$

It is very important to realize that expectation is linear, i.e,

$$E\left(af(X)+bg(X)\right)=aE\left(f(X)\right)+bE\left(g(X)\right).$$

Example

(1) Show that if X has nonnegative integers as values, then $E(X)=\sum_{n=1}^{\infty}P(X\geq n)$.
Proof: $E(X)=\sum_{n=1}^{\infty}nP(X=n)=\sum_{n=1}^{\infty}n(P(X\geq n)-P(X\geq n+1))=\sum_{n=1}^{\infty}P(X\geq n)$. $\square$

(2) Show that if X is a nonnegative random variable, then $E(X)=\int_0^{\infty}P(X\geq x)dx$
Proof: $E(X)=\int_0^{\infty}xf(x)dx=\int_0^{\infty}xdP(X\leq x)=\int_0^{\infty}xd(1-P(X\geq x))=-\int_0^{\infty}xdP(X\geq x)=\int_0^{\infty}P(X\geq x)dx$. $\square$

The variance of r.v. X is defined as $\text{var}(X)=E\left((X-E(X))^2\right)$. Note that $\text{var}(aX)=a^2\text{var}(X)$ and

$$E\left((X-E(X))^2\right)=E\left(X^2\right)-\left(E(X)\right)^2.$$

Example If $X\sim U(a,b)$, then $E(X)=\frac{a+b}{2}$ and $\text{var}(X)=\frac{1}{12}(b-a)^2$. If $X\sim N(\mu,\sigma^2)$, then $E(X)=\mu$ and $\text{var}(X)=\sigma^2$. $\square$

Proposition 1 (Chebyshev inequality)

$$P(|X-EX|\geq a)\leq\frac{var(X)}{a^2}. \tag{1}$$

Proof: Let $\mu=EX$.

$$\frac{\text{var}(X)}{a^2}=\frac{1}{a^2}\int_{\mathbb{R}}(x-\mu)^2dP(x)\geq\frac{1}{a^2}\int_{\{x:|x-\mu|\geq a\}}(x-\mu)^2dP(x)\geq\int_{\{x:|x-\mu|\geq a\}}1dP(x).$$

The last term above is exactly $P(|X-EX|\geq a)$. $\square$

Let us now consider the r.v. $X=(X_1,...,X_d)$ taking values in $\mathbb{R}^d$. X is called multivariate random variable. Its cdf is defined as $F(x)=P(\bigcap_{i=1}^d\{X_i\leq x_i\})$ with $x\in\mathbb{R}^d$. If X is a discrete r.v., then $f(x)=P\left(\bigcap_{i=1}^d\{X_i=x_i\}\right)$. If X is a continuous r.v., its density is $f(x)=\frac{\partial^dF(x)}{\partial x_1\cdots\partial x_d}$. For example, a multivariate normal distribution has a

$$f(x)=|2\pi\Sigma|^{-1/2}\exp\{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)\}$$

where $\mu\in\mathbb{R}^d$ and $\Sigma\in\mathbb{R}^{d\times d}$ is positive definite.

When we say X_i's are mutually independent, it means the event $\{X_1 \le x_i\}$'s are are independent. Then for discrete r.v. $f(x) = P\left(\bigcap_{i=1}^d \{X_i = x_i\}\right) = \prod_i P(X_i = x_i) = \prod_i f_{X_i}(x_i)$ and for continuous r.v. $f(x) = \frac{\partial^d F(x)}{\partial x_1 \cdots \partial x_d} = \prod_i \frac{\partial P(X_i \le x_i)}{\partial x_i} = \prod_i f_{X_i}(x_i)$. If X_i are mutually independent,

$$\mathrm{var}(\sum_i X_i) = \sum_i \mathrm{var}(X_i) + \sum_{i<j} 2E\left((X_i - EX_i)(X_j - EX_j)\right) = \sum_i \mathrm{var}(X_i).$$

A weaker assumption which guarantees $\mathrm{var}(X_1+X_2) = \mathrm{var}(X_1)+\mathrm{var}(X_2)$, which is equivalent to $E((X_1 - EX_1)(X_1 - EX_1)) = 0$, is that X_1 and X_2 are uncorrelated.

Proposition 2 (L^2 weak law of large number) *Let X_1, X_2, ... be mutually uncorrelated random variables with $E(X_i) = \mu$ and var$(X_i) \le C < \infty$. Let $S_n = X_1 + \ldots + X_n$. Then, as $n \to \infty$, $S_n/n \to \mu$ in L^2 (which means $E(S_n/n - \mu)^2 \to 0$).*

Proof: Observe that $E(S_n/n) = \mu$.

$$E(S_n/n - \mu)^2 = \mathrm{var}(S_n/n) = \frac{1}{n^2}\mathrm{var}(S_n) = \frac{1}{n^2}\left(\mathrm{var}(X_1) + \cdots + \mathrm{var}(X_n)\right) \le \frac{Cn}{n^2} \to 0.$$

3.2 Random number generator

We only discuss the linear congruential (pseudo-)random number generator $x_{n+1} = f(x_n)$ with

$$f(x) = (ax + c) \mod m$$

where x_n is an integer and $a > 1$, $c \ge 0$, $m >> 1$ are fixed integers. Once x_i is obtained, $u_i = x_i/m$ will be output and be used as a random number between 0 and 1. We use the notation $m|x$ to mean that m divides x (exactly, with no remainder) and we write $a = b \mod m$, if $m|(a - b)$. One says that a is congruent to b modulo m. For example, $21 = 5 \mod 8$ and $-19 = 5 \mod 8$.

Example: Let $a = 5$, $c = 1$, and $m = 8$ and let the generator be seeded with $x = 0$. The sequence of samples will be

$$0, 1, 6, 7, 4, 5, 2, 3, 0,$$

whereupon the list cycles.

The fact that $a > 1$ makes the iteration behave chaotically which makes it a good random number generator. Setting $a > 1$ also ensures that if you run the linear congruential random number generator twice with two slightly different initial seed values, even though the first few terms might be similar, the two sequences will quickly diverge.

Obviously, the iteration $x_{n+1} = f(x_n)$ eventually cycles: there is a positive integer p such that $x_i = x_{i+p}$ for all sufficiently large i. The smallest such p is called the cycle length or period of the random number generator. Ideally we want p as large as possible. But obviously, it cannot bigger than m. There are sufficient and necessary conditions for the above iteration to has period equal to m. See Theorem 1.3 of the book.

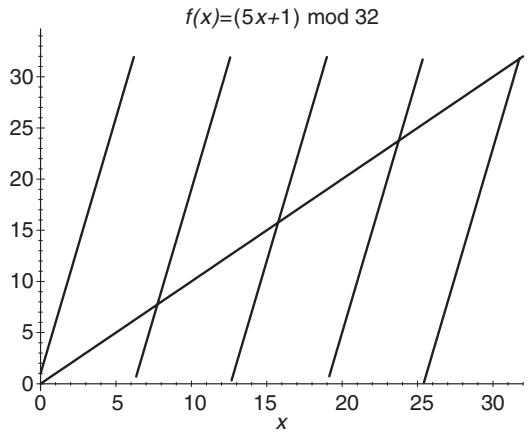

Figure 8: Fixed-point iteration on a sawtooth. The sequence of samples x_n is in reality a fixed-point iteration sequence $x_{n+1} = f(x_n)$ on this graph. Starting with x_n on the x-axis, move vertically up to the sawtooth graph, this is $y = f(x_n)$. Now move horizontally at this value of y until you reach the identity function $y = x$, drop vertically down to the x-axis, and this is x_{n+1}.

Example: Let $a = 7^5 = 16807$, $c = 0$, and $m = 2^{31} - 1$ (which is a prime number). This is the random number generator used in Matlab version 4 in the 1990s. The cycle length is m which is around 2×10^9, perhaps sufficient for the 20th century, but not sufficient nowadays.

3.3 Some probability distributions and their uses

We have already learned the uniform distribution $U(a, b)$ and Gauss distribution $N(\mu, \sigma^2)$.

(1) Suppose a r.v. takes only values 0 and 1. It has chance p to be 1 and chance $1 - p$ to be 0. Then it is mean is p and variance is $p(1-p)^2 + (1-p)(0-p)^2 = p(1-p)$.

(2) Binomial distribution $B(n, p)$: The *number of ways* in which there can be x successes in n trials is given by $C(n, x) = \frac{n(n-1)\cdots(n-x+1)}{x(x-1)\cdots 2\cdot 1}$. Note that the numerator $x!$ is the number of permutations of x symbols. Suppose p is the probability of success for each trial. Then the probability that there are x successes out of n trials is $f(x) = C(n, x)p^x(1-p)^{n-x}$. It is means is np as $X = \sum_{i=1}^{n} X_i$ with $X_i = 1$ with change p and $= 0$ with chance $1 - p$. X_i's are mutually independent. Hence $EX = np$ and $\text{var}X = n\text{var}X_1 = npq$.

(3) Poisson distribution $Poi(\lambda t)$: Let X denotes the number of events that occur in an interval of length t. The event could be "a custom arrived", "an incoming call", "the failure of some mechanical component we want to exam" or "a biological birth". We know on average the event happens at rate λ (the number of events in a unit time). Partition the interval into n subintervals of equal length t/n where n very large and we assume the probability of exactly one event in interval t/n is $\lambda t/n$ (and some other

minor assumptions: see page 61 of the book). Then the probability of x events in the n subintervals is

$$C(n,x)\left(\frac{\lambda t}{n}\right)^x\left(1-\frac{\lambda t}{n}\right)^{n-x} \to \frac{(\lambda t)^x e^{-\lambda t}}{x!} \qquad \text{as} \quad n\to\infty$$

after some simple computation[4]. So $P(X=x)=\frac{(\lambda t)^x e^{-\lambda t}}{x!}$. Since $X=\sum_{i=1}^n X_i$ with $EX_i=\lambda t/n$ and $\text{var}X_i=\left(\frac{\lambda t}{n}\right)\left(1-\frac{\lambda t}{n}\right)$, it is easy to believe that $EX=\lambda t$ and $\text{var}X=\lambda t$. Indeed this is true. See page 63 of the book for the computations. See also the tutorial problems.

(4) Exponential distribution $X\sim E(\lambda)$: The exponential distribution is used to model the random waiting time for event to happen. Since $P(Y_t=0)=e^{-\lambda t}$ for a Poisson distribution Y_t,

$$\begin{aligned} P(X\le t) &= P(\text{at least one event occurs in } [0,t)) \\ &= 1-P(\text{no event occurs in}[0,t)) = 1-e^{-\lambda t}. \end{aligned}$$

So the pdf is the derivative of the above function and hence is $\lambda e^{-\lambda t}$. By simple calculation, $E(X)=\frac{1}{\lambda}$ which is obvious since in unit time λ events happen. We can also verify $\text{var}(X)=\frac{1}{\lambda^2}$.

The exponential distribution is "memoryless" (also called Markovian). For example, if an event has not yet occurred by time a, then the probability that it occurs in the interval a to $a+b$ is the same as that of its occurring in 0 to b:

$$P(X\le a+b|X>a)=P(X\le b)$$

This is so because in terms of the cfd, $P(X\le a+b|X>a)=\frac{P(\{X\le a+b\}\cap\{X>a\})}{P(X>a)}=\frac{F(a+b)-F(a)}{1-F(a)}=\cdots=1-e^{-\lambda b}$.

Theorem 1 *Suppose $U\sim U(0,1)$ and F is a cdf. Then $X=F^{-1}(U)$ has the distribution F. Here we define $F^{-1}(u)=\inf\{x,F(x)\ge u\}$.*

Proof: Since F is monotonically non-decreasing,

$$P(X\le x)=P(F^{-1}(U)\le x)=P(U\le F(x))=F(x).$$

See page 67 of Shonkwiler and Mendivil for rigorous proof. □

[4]Indeed, $C(n,x)\frac{x!}{n^x}\to 1$ and $\left(1-\frac{\lambda t}{n}\right)^{n-x}\to e^{-\lambda t}$.

3.3.1 Sampling from the exponential and Poisson distribution

The problem with the above theorem is that in most case we do not have an explicit formula for the cdf. For example, what is the cfd for $N(0,1)$? Fortunately, the method for obtaining a sample from $E(\lambda)$ is simple. From $U = 1 - e^{-\lambda X}$, we get $X = -\frac{1}{\lambda}\ln(1-U)$. But as U and $1-U$ has the same pdf, we can just take $X = -\frac{1}{\lambda}\ln U$.

Since the Poisson r.v. is the number of exponential events that occur in the interval $[0,t)$, we may simulate the Poisson distribution by sampling exponentials until time t is exceeded. On page 67 of the book of Shonkwiler and Mendivil, three methods are given, the last one as follows:

- Let $U_0 = 1$ and $U_i \sim U(0,1)$ for $i = 1, 2, ...$ Let k be the largest integer such that $\prod_{i=0}^{k} U_i > e^{-\lambda t}$. Then let $X = k$ and the X satisfies the Poisson distribution $Poi(\lambda t)$.

Note what behind the above algorithm is the simple fact $\prod_{i=0}^{k} U_i > e^{-\lambda t}$ if and only if $\sum_{i=1}^{k} -\frac{1}{\lambda}\ln U_i < t$. So k is the largest integer that the sum of k Poisson r.v.'s $< t$ which means exactly k events occurs in the time interval $[0,t)$.

Application: Discrete event simulation. A software simulation is managed by a "simulation executive" that is responsible for stepping time through the simulation and invoking the other modules as necessary. These include modules for processing the entities of the simulation, implementing their interdependencies, and data display and logging modules.

A timetable of events is maintained, always in chronological order, and the simulation proceeds from one event to the next event in the schedule. Each event has a tag describing when it occurs in absolute time and its nature. When an event is handled, it may generate new events each of which is scheduled by sampling from the appropriate waiting time distribution and merging it into the master schedule at the appointed time.

Consider the example of simulating cell tissue growth, say starting with one cell at time 0 and carrying the simulation through for 8 hours. Assume that cell maturation times are exponentially distributed with parameter $\lambda = 0.5$ per hour. The master schedule needs to contain event times only, since all the events are the same, namely binary fission of identical cells. A more ambitious simulation could invoke crowding and other effects between the cells as desired. Drawing a sample from $E(0.5)$, we might get $W = 3$ hours, say. So our original cell divides at time $t = 3$ and our master schedule is $S[1] = 3$. Now move time forward to $t = 3$ and handle the cell division of the first cell. Sample $E(0.5)$ again to get $W = 4$, say, for the first daughter cell, so now $S[2] = 7$ (the present time is $t = 3$ and the cell division time is 4, so the event is at absolute time 7). And again for the second daughter cell, get $W = 1$, say. Now we must shift the $S[2] = 7$ down to keep the schedule in time-sorted order, so $S[2] = 4$ and $S[3] = 7$. The next event on the schedule is at time $t = 4$, the division of a first-generation daughter cell. Sample from $E(0.5)$ to get $W = 2$, say, and again to get $W = 5$. For the $W = 2$ event, its absolute time will be $4 + 2 = 6$, so the master schedule becomes $S[3] = 6$ and $S[4] = 7$. The $W = 5$ event occurs at absolute time t = 9, which is beyond our study, so we need not include it in the master schedule. (But it would have to be accounted for if cell interactions were being monitored.) Continuing in this fashion, we finally reach

the point where all subsequent events occur beyond absolute time t = 8 and the simulation is over. At this point the statistics of the run are tabulated and presented.

$$\begin{pmatrix} 3 \\ \\ \\ \end{pmatrix} \rightarrow \begin{pmatrix} 3 \\ 7 \\ \\ \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 7 \\ \\ \end{pmatrix} \rightarrow \begin{pmatrix} 3 \\ 4 \\ 6 \\ 7 \\ \\ \end{pmatrix} \begin{pmatrix} 3 \\ 4 \\ 6 \\ 7 \\ 9 \\ \end{pmatrix} \cdots .$$

3.3.2 Sampling from normal distribution (Box-Muller algorithm)

We will discuss a method that generates two independent r.v.'s $\sim N(0,1)$. The idea is to first uniformly choosing an angle in $\mathbb{R}^2$ and then generating the square distance from an exponential distribution $E\left(\frac{1}{2}\right)$. Hence is the algorithm is as follows

1) Sample random r.v. U_1, $U_2 \sim U(0,1)$.

2) Let $\theta = 2\pi U_1$ and $r = \sqrt{-2\ln U_2}$.

3) $X_1 = r\cos(\theta)$, $X_2 = r\sin(\theta)$.

Claim: X_1 and X_2 are independent identical distribution (i.i.d) and $X_i \sim N(0,1)$.
Proof: Denote the joint pdf of X_1 and X_2 by $f_{(X_1,X_2)}$. All we need to show is $f_{(X_1,X_2)}(x_1,x_2) = \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}$. In another word, we want to prove

$$\begin{aligned} P\left((X_1,X_2)\in A\right) &= \int_A f_{(X_1,X_2)}(x_1,x_2)dx_1dx_2 \\ &= \int_A \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}dx_1dx_2 \end{aligned} \tag{2}$$

for any region $A \subset \mathbb{R}^2$.

Consider the mapping $(u_1,u_2) \rightarrow (x_1,x_2)$ defined by $x_1 = \cos(2\pi u_1)\sqrt{-2\ln u_2}$ and $x_1 = \sin(2\pi u_1)\sqrt{-2\ln u_2}$. Suppose the mapping $(u_1,u_2) \rightarrow (x_1,x_2)$ maps region B to region A.

If we want to change the integration $\int_A \cdots dx_1dx_2$ on the right hand side of the above equation to $\int_B \cdots du_1du_2$, we need to add the Jacobian $\det\left(\frac{\partial x_i}{\partial u_j}\right)$. Straightforward computation shows $\det\left(\frac{\partial x_i}{\partial u_j}\right) = \frac{2\pi}{u_2}$. Note that $\frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}$ becomes $\frac{1}{2\pi}u_2$. Hence

$$\begin{aligned} \int_A \frac{1}{\sqrt{2\pi}}e^{-x_1^2/2}\frac{1}{\sqrt{2\pi}}e^{-x_2^2/2}dx_1dx_2 &= \int_B \frac{1}{2\pi}u_2\frac{2\pi}{u_2}du_1du_2 = \int_B du_1du_2 \\ &= P\left((U_1,U_2)\in B\right). \end{aligned}$$

Hence (2) follows from $P\left((X_1,X_2)\in A\right) = P\left((U_1,U_2)\in B\right)$.

3.3.3 Rejection sampling

This is due to von Neumann (1951). See page 84 of the book of Shonkwiler and Mendivil.

The aim is to sample r.v. with given pdf $f(x)$. Suppose we have a formula for $\ell(x) = cf(x)$ for some constant c; suppose we can easily generate samples with pdf $g(x)$ (e.g. $U(a,b)$, $E(\lambda)$, $N(\mu, \sigma^2)$); and suppose we can find an M so that $Mg(x) \geq \ell(x)$. The algorithm of rejection sampling for generating r.v $X \sim f(x)$ is as follows

(1) Draw a sample Y from $g(\cdot)$ and compute the ratio $h(Y) = \frac{\ell(Y)}{Mg(Y)}$.

(2) Generate $U \sim U(0,1)$. If $U \leq h(Y)$, let $X = Y$ and return. Otherwise, turn to step 1).

Proof: Let $P_0 = P(\text{proposed } Y \text{ is accepted})$. Since $P(U \leq h(y)|Y = y) = h(y)$, $P_0 = \int_y P(U \leq h(y)|Y = y)g(y)dy = \int_y h(y)g(y)dy = \frac{1}{M}\int_y \ell(y)dy = c/M$. By the formula $P(A|B) = \frac{P(A \cap B)}{P(B)}$

$$
\begin{aligned}
&P(X \in [y, y+dy]) = P(Y \in [y, y+dy] | \text{ proposed } Y \text{ is accepted}) \\
=&\frac{P(Y \in [y, y+dy] \text{ and proposed } Y \text{ is accepted})}{P(\text{proposed } Y \text{ is accepted})} \\
=&\frac{P((1) \text{ happens with } Y \in [y, y+dy] \text{ and then (2) happens with } U \leq h(Y))}{P(\text{proposed } Y \text{ is accepted})} \\
=&\frac{g(y)dy \times h(y)}{\frac{c}{M}} = f(y)dy. \qquad \square
\end{aligned}
$$

Remark: Why do we need this ℓ. For example, we may know

$$
f(x) \propto \exp\left(\mu \sum_{i=1:n-1, j=1:n-1} x_{i,j}(x_{i-1,j} + x_{i+1,j} + x_{i,j-1} + x_{i,j+1})\right)
$$

where $i, j = 0, ..., n$ and $x = (x_{i,j}) \in \mathbb{R}^{(n+1)\times(n+1)}$). x_{ij} can be viewed as a function defined on the grid points of $\mathbb{Z}^2$. μ is a constant. So, in this case, we have a formula for $\ell(x)$, but the exact formula for $f(x)$ is very difficult to compute as we need to determine the normalization constant.

Example: Prove that the expected number of samplings from $Y \sim g$ for obtaining one accepted sample is M/c. So, we shall keep M/c as small as we can.
Solution: Let Z be the number of samplings from $Y \sim g$. Then $P(Z = k) = (1 - P_0)^{k-1}P_0$ and $EZ = \sum_{k=1}^{\infty}(1 - P_0)^{k-1}P_0 k = -P_0 \frac{d}{dP_0}\sum_{k=1}^{\infty}(1 - P_0)^k = \frac{1}{P_0} = M/c$. $\square$

3.4 Monte Carlo integration

Suppose we want to calculate

$$
I = \int_0^1 h(x)dx
$$

with $h(x) = (\cos(50x) + \sin(20x))^2$. Although it is possible to integrate this function analytically, it is a good first test case. To calculate the integral, we generate $U_1, U_2, \ldots U_n$ independent identically distributed (iid) $U(0,1)$ random variables and approximate $\int_0^1 h(x)dx$ with $\hat{I}_n = \frac{1}{n}\sum_{i=1}^n h(U_i)$.

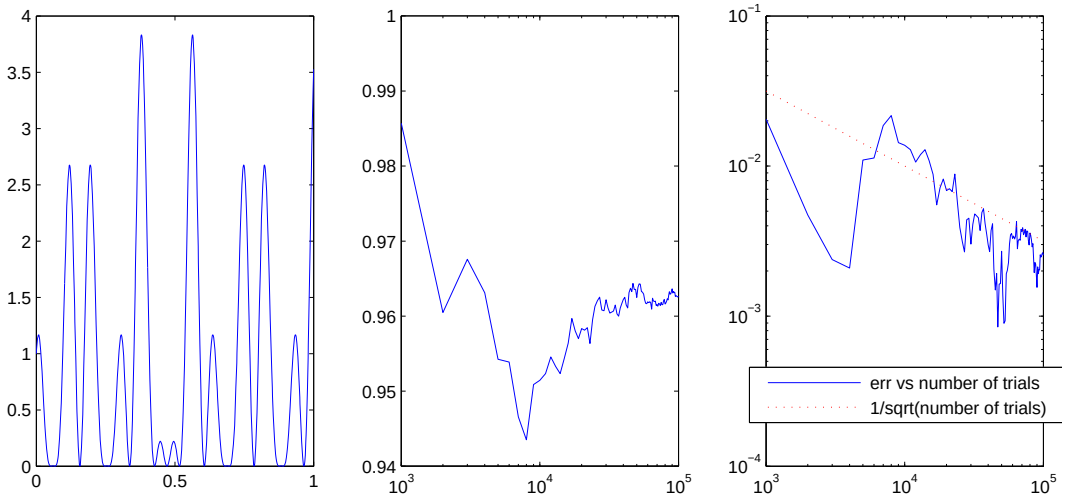

Figure 9: Calculation of $\int_0^1 (\cos(50x) + \sin(20x))^2 dx$. Left: plot of $h(x) = (\cos(50x) + \sin(20x))^2$. Middle: $\hat{I}_n$ vs n. Right: $|\hat{I}_n - I|$ vs n.

```
% syms x; I=int((cos(50*x)+sin(20*x)).^2,0,1)
I=1/200*sin(100)+103/105-1/70*cos(70)+1/30*cos(30)-1/80*sin(40);

x=linspace(0,1,1000);
h=(cos(50*x)+sin(20*x)).^2;
figure; subplot(1,3,1); plot(x,h);

rand('state',1);
N=100; L=1000;
n=zeros(N,1); vEst=zeros(N,1);
r=rand(N*L,1);
for k=1:N
    n(k)=L*k;
    x=r(1:n(k));
    h=(cos(50*x)+sin(20*x)).^2;
    vEst(k)=sum(h)/n(k);
end
err=abs(vEst-I);

subplot(1,3,2); semilogx(n,vEst);
```

```
subplot(1,3,3); loglog(n,err); hold on;
loglog(n,1./sqrt(n),'r:');
legend('err vs number of trials','1/sqrt(number of trials)');
```

For more examples, see the Lab tutorial problems and Question 2 of the homework.

3.5 Error estimates for Monte Carlo integration (simulation)

Here are some different types of convergence:

1) We say $Y_i \to Y$ in distribution if $\lim_{i\to\infty} F_i(y) = F(y)$ where $F_i(y)$ is the cdf of Y_i and $F(y)$ is the cdf of Y.

2) We say $Y_i \to Y$ in probability if $\forall \varepsilon > 0$, $\lim_{i\to\infty} P(|Y_i - Y| > \varepsilon) = 0$.

3) We say $Y_i \to Y$ almost surely (a.s.) if $P(\lim_{i\to\infty} |Y_i - Y| = 0) = 1$.

The following are theorems about the sum of random variables

1) Weak law of large numbers. If $X_1, ..., X_i, ...$ are i.i.d. random variables with finite mean μ, then $\frac{\sum_{i=1}^n X_i}{n} \to \mu$ in probability.

2) Strong law of large numbers. If $X_1, ..., X_i, ...$ are i.i.d. random variables with finite mean μ and finite variance, then $\frac{\sum_{i=1}^n X_i}{n} \to \mu$ a.s.

3) Central limit theorem. For mutually independent random variable $X_1, ..., X_i, ...$ with mean μ and variance σ^2, $\frac{\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n X_i - \mu\right)}{\sigma} \to N(0,1)$ in distribution. If $\mu = 0$, then $\frac{\sum_{i=1}^n X_i}{\sqrt{n}} \to N(0, \sigma^2)$ in distribution.

Let us now discuss the error related to Monte Carlo integration. Suppose we are trying to use Monte Carlo to estimate some value, call it θ. It could be the waiting time of a discrete event simulation or the neutron flux through a thickness of shielding or any scalar value that results from an instance of a simulation. Let X_1, X_2, ..., X_n be n estimates of θ as derived from the outcome of the simulation. If the X_i are independent and identically distributed with mean θ, then by the central limit theorem their sample average $\bar{X} = \frac{1}{n}\sum_{i=1}^n X_i$ is approximately normally distributed with mean θ and variance σ_X^2/n, where σ_X^2 is the (unknown) variance of the X_i. In this case $Y = \frac{\bar{X}-\theta}{\sqrt{\sigma_X^2/n}}$ is approximately $N(0,1)$ distributed. From a $N(0,1)$ table we notice that with probability 0.954 a normal sample lies within two standard deviations of the mean; hence

$$P\left(-2 < \frac{\bar{X}-\theta}{\sqrt{\sigma_X^2/n}} < 2\right) = 0.954.$$

In other words, with probability 0.954, θ lies in the interval

$$\bar{X} - 2\sqrt{\sigma_X^2/n} < \theta < \bar{X} + 2\sqrt{\sigma_X^2/n}.$$

Now, given a value for σ_X, we may calculate probabilistic error bounds for θ, or confidence intervals as they are called.

For example, if $\theta = \int_{\mathbb{R}^d} f(y)dy$ is some high dimensional integration, we can interpret it as $\theta = Eh(Y) = \int_{\mathbb{R}^d} h(y)g(y)dy$. Here $h = f/g$ and g is the pdf of r.v. Y. If we sample Y_i $(i = 1, ..., n)$ with pdf $g(y)$ and construct $\bar{\theta} = \frac{1}{n}\sum_{i=1}^n h(Y_i)$ (so $X_i = h(Y_i)$), then with probability 0.954, θ lies in the interval

$$\bar{\theta} - 2\sqrt{\sigma_X^2/n} < \theta < \bar{\theta} + 2\sqrt{\sigma_X^2/n}.$$

where $\sigma_X^2 = E\left(h(Y) - \theta\right)^2 = \int_{\mathbb{R}^d} \left(h(y) - \theta\right)^2 g(y)dy$.

In practice there are three problems with this program. First, usually σ_X must itself be estimated from the data. In that case Y above will not be normal. Second, the X_i may not be identically distributed; the simulation may suffer start-up effects, for example. And third, the X_i may be correlated. Both of these issues are, in fact, a common difficulty in queuing simulations.

Related to the first issue that σ_X can be unknown, to estimate the error, the value of σ_X can be estimated by the sample variance

$$s_n^2 = \frac{1}{n-1}\sum_{i=1}^n \left(X_i - \bar{X}\right)^2.$$

It can be shown that $E(s_n^2) = \sigma^2 = E(X - EX)^2$ when X_i are i.i.d.

An attempt to deal with the second and third issues is by batching. Divide the n trials into m batches each of size J:

$$X_1, ..., X_J \,|\, X_{J+1}, ..., X_{2J} \,|\, ... \,|\, X_{(m-1)J+1}, ..., X_{mJ}.$$

Thus there are $m = n/J$ batches. The batch random variables $B_i = \frac{1}{J}\sum_{i=(i-1)J+1}^{iJ} X_i$ tend to be independent and identically distributed. Now we may apply the development above to the B_i in place of the X_i. The sample variance for σ_B^2 is given by

$$s_B^2 = \frac{1}{m-1}\sum_{i=1}^m \left(B_i - \hat{\theta}\right)^2 \qquad \text{with} \quad \hat{\theta} = \frac{1}{m}\sum_{j=1}^m B_i.$$

3.6 Variance reduction methods

As we have see above, the error is of the size $\frac{\sigma_X}{\sqrt{m}}$. So we want to reduce σ_X and increase m. We now want to discuss how to reduce σ_X. This part is not discussed in the book since the book is intended to be introductory. So we only discuss it very briefly.

3.6.1 Importance sampling

Importance sampling can improve the efficiency by orders of magnitude in some problem, but it requires caution: an inappropriate implementation can reduce efficiency by orders of magnitude!

Suppose we want to estimate $I = \int_0^1 f(x)dx$. We can rewrite it as $E_U f(U)$ with $U \sim U(0,1)$ and E_U means taking expectation with respect to random variable U. On the other hand, choosing any random variable X with pdf $g(x)$, $I = \int_0^1 \frac{f(x)}{g(x)} g(x)dx = E_X\left(\frac{f(X)}{g(X)}\right)$. We want to choose g so that X is easy to generate and $\mathrm{var}_X \frac{f(X)}{g(X)} \leq \mathrm{var}_U f(U)$. If you look at the σ_X in the error estimates we discussed before, you see that we are trying to make it small. The X we used before in that section equals to to the $f(U)$ or the $\frac{f(X)}{g(X)}$ that we are talking about right now.

Since $\mathrm{var}_X \frac{f(X)}{g(X)} = E_X\left(\frac{f(X)}{g(X)}\right)^2 - \left(E_X \frac{f(X)}{g(X)}\right)^2 = E_X\left(\frac{f(X)}{g(X)}\right)^2 - I^2$, we want to choose $g \geq 0$ with $\int g = 1$ so that $E_X\left(\frac{f(X)}{g(X)}\right)^2 < E_U\left(f(U)\right)^2$, i.e., $\int \frac{f(x)^2}{g(x)}dx < \int f(x)^2 dx$.

Another way to look at it: $\mathrm{var}_X \frac{f(X)}{g(X)} = \int \left(\frac{f(x)}{g(x)} - I\right)^2 g(x)dx$. Observe that this integral is small if $f(x)/g(x) \approx I$ for most x. This leads to the rule of thumb: g should be large when f is large and this gives the name "importance sampling".

Example: Let $f(x) = 4\sqrt{1-x^2}$, $x \in [0,1]$. $I = \pi$. If we use simple sample, $\mathrm{var}_U f(U) = Ef(U)^2 - \left(Ef(U)\right)^2 = \int_0^1 16(1-x^2)dx - \pi^2 = 0.797$. Let $g(x) = \frac{4-2x}{3}$ which imitates the relative heights of the graph $f(x)$ moderately well. $\mathrm{var}_X \frac{f(X)}{g(X)} = E_X\left(\frac{f(X)}{g(X)}\right)^2 - \left(E_X \frac{f(X)}{g(X)}\right)^2 = \int_0^1 \frac{f(x)^2}{g(x)}dx - \pi^2 \approx 0.224$.

Example: Let $f(x) = 4\sqrt{1-x^2}$, $x \in [0,1]$. $I = \pi$. If we use simple sample, $\mathrm{var}_U f(U) = Ef(U)^2 - \left(Ef(U)\right)^2 = \int_0^1 16(1-x^2)dx - \pi^2 = 0.797$. Let $g(x) = 2 - 2x$, show that $\mathrm{var}_X \frac{f(X)}{g(X)} = 2.1$ which is worse than the simple sampling. (Draw the graph of $g(x)$ and $f(x)$ (See Figure 10) and you will see this g does seem to satisfy the rule of thumb.)

Solution: $X \sim g(x) = 2 - 2x$. $E_X\left(\frac{f(X)}{g(X)}\right) = \int_0^1 f(x)/g(x)g(x)dx = \pi$. $E_X\left(\frac{f(X)}{g(X)}\right)^2 = \int_0^1 f^2/g dx = 12$. Hence $\mathrm{var}_X\left(\frac{f(X)}{g(X)}\right) = 12 - \pi^2 \approx 2.1304$.

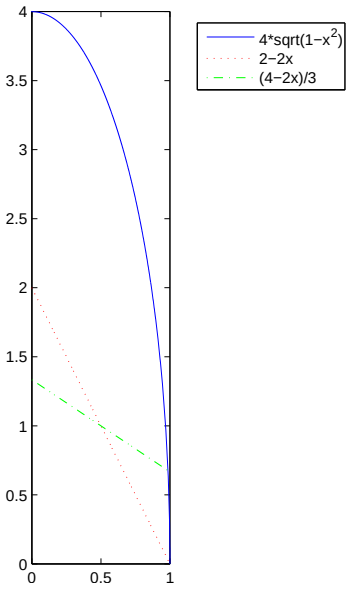

Figure 10: A wrong g is chosen for important sampling.

3.6.2 Stratified sampling

Suppose we want to estimate $\mu = \int_D f(x)dx$. If possible, we want to break the region D into the union of k disjoint subregions, D_1,...,D_k so that within each subregion, the function $f(x)$ is relatively "homogeneous" (i.e, close to being a constant). Then we can spend m_i random samples, $X^{(i,1)}$,...,X^{i,m_i} (each is a uniform distribution random variable) in the subregion D_i. The subregional integral $\mu_i = \int_{D_i} f(x)dx = E_{g_i} \frac{f(X^{(i)})}{g_i(X^{(i)})}$ with $g_i = 1/|D_i|$ can be approximated by

$$\hat{\mu}_i = \frac{1}{m_i} \left(|D_i| f(X^{(i,1)}) + \cdots + |D_i| f(X^{(i,m_i)}) \right).$$

The overall integral μ can be approximated by $\hat{\mu} = \hat{\mu}_1 + \cdots \hat{\mu}_k$ whose variance is now

$$\mathrm{var}(\hat{\mu}) = \frac{\sigma_1^2}{m_1} + \cdots + \frac{\sigma_k^2}{m_k}$$

where $\sigma_i^2 = \mathrm{var}_{g_i}\left(\frac{f}{g_i}\right)$ is the variance of f/g_i in region D_i. In contrast, if we use all the $m = \sum_{i=1}^k m_i$ samples to do a plain uniform sampling in the region D, the variance of the estimate would be σ^2/m with $\sigma^2 = \mathrm{var}_g\left(\frac{f}{g}\right)$ being the overall variance of f/g in D. Here $g = 1/|D|$.

See Problem 3 of Tutorial VII for further related discussions.

3.6.3 Control variate method

In this method, one uses a control variate C which is correlated with the sample X, to produce a better estimate. Suppose the estimation of $\mu = EX$ is of interest and $\mu_C = EC$

is known. Then we can construct Monte Carlo samples of the form

$$X(b) = X - b(C - \mu_C)$$

which has the same mean as X.

If the computation of $cov(X, C)$ and $var(C)$ is easy, then we can let $b = \frac{cov(X,C)}{var(C)}$. In the Tutorial problem, we will prove that $var(X(b)) < var(X)$.

3.6.4 Antithetic Variate Method

Suppose $U \sim U(0, 1)$ is a random number used in the production of a sample X that follows a distribution with cdf F (i.e. $X = F^{-1}(U)$). Then $X' = F^{-1}(1-U)$ also follows distribution F.

In the Tutorial problem, we will prove that $\text{var}(X + X') < 2\text{var}(X)$. This means using the pair X and X' is better than using two independent Monte Carlo draws for estimating EX.

3.6.5 Rao-Blackwellization

This method reflects a basic principle (or rule of thumb) in Monte Carlo computation: One should carry out analytical computations as much as possible. The problem can be formulated as follows: Suppose we are drawn independent samples $X^{(1)}, ..., X^{(m)}$ from the target distribution $f(x)$ and are interested in evaluating $I = E(h(X))$. A straight forward estimator is

$$\hat{I} = \frac{1}{m}\left(h(X^{(1)}) + \cdots + h(X^{(m)})\right).$$

Suppose in addition that X can be decomposed into two parts (X_1, X_2) and that the condition expectation $E(h(X)|X_2)$ can be carried out analytically. An alternative estimator of I is

$$\tilde{I} = \frac{1}{m}\left(E[h(X)|X_2^{(1)}] + \cdots + E[h(X)|X_2^{(m)}]\right).$$

Clearly $E(\hat{I}) = E(\tilde{I}) = I$ since $E(h(X)) = E[E\{h(X)|X_2\}]$. However, $\text{var}(\hat{I}) \geq \text{var}(\tilde{I})$ because

$$\text{var}\{h(X)\} \geq \text{var}\{E[h(X)|X_2]\}. \tag{3}$$

Intuitively, the above inequality makes sense, for a proof of it, see the book "probability theory" by Varadhan, page 80; Or the book "Probability and measures" by Billingsley, page 454. The above procedure is usually called "Rao-Blackwellization" because (3) follows from the Rao-Blackwell theorem.

In the following, we will give a simple proof of (3) assuming the pdf of x exists. Note that $E[h(X)|X_2 = x_2] = \frac{\int_{\mathbb{R}} h(x_1,x_2)f(x_1,x_2)dx_1}{\int_{\mathbb{R}} f(x_1,x_2)dx_1}$ since for any $A \subset \mathbb{R}$

$$\int_A E[h(X)|X_2 = x_2]\left(\int_{\mathbb{R}} f(x_1, x_2)dx_1\right)dx_2 = \int_{A\times\mathbb{R}} h(x_1, x_2)f(x_1, x_2)dx_2dx_1.$$

(The last equation is indeed the definition of $E[h(X)|X_2 = x_2]$.) So by Cauchy-Schwaz, $\left(\int h\sqrt{f}\sqrt{f}dx\right)^2 \leq \left(\int h^2 f\right)\left(\int f\right)$,

$$\begin{aligned}(E[h(X)|X_2 = x_2])^2 &= \left(\frac{\int_{\mathbb{R}} h(x_1, x_2)f(x_1, x_2)dx_1}{\int_{\mathbb{R}} f(x_1, x_2)dx_1}\right)^2 \\ &\leq \frac{\int_{\mathbb{R}} h^2(x_1, x_2)f(x_1, x_2)dx_1}{\int_{\mathbb{R}} f(x_1, x_2)dx_1} = E[(h(X))^2 |X_2 = x_2].\end{aligned}$$

Taking E on both sides, and using $E(E(Y|X)) = E(Y)$, we obtain $E\left(E[h(X)|X_2 = x_2]\right)^2 \leq E[(h(X))^2]$. Then, because $\text{var}(Y) = EY^2 - (EY)^2$, $\text{var}\left(E[h(X)|X_2 = x_2]\right) = E\left(E[h(X)|X_2 = x_2]\right)^2 - \left(E\left(E[h(X)|X_2 = x_2]\right)\right)^2 \leq E[(h(X))^2] - \left(E[h(X)]\right)^2 = \text{var}\{h(X)\}$. $\square$

3.7 The MCMC principle and basic properties of a Markov chain

So far, we have discussed the Monte Carlo integration, which simulates a sample from the distribution f to approximate the integral

$$I = \int_A h(x)f(x)dx.$$

If we can generate an ergodic Markov chain $\{X_t\}$ with stationary distribution f, then we can also approximate the above integral by $\frac{1}{T}\sum_{j=1}^{T} h(X_j)$.

Definition 7 *A Markov chain Monte Carlo (MCMC) method for the simulation of a distribution f is any method producing an ergodic Markov chain $\{X_t\}$ whose stationary distribution is f.*

We will elaborate these terminologies in the following discussion.

Consider a sequence of random variables $X_0, X_1, ...$, defined on a finite state space $\mathfrak{X}$. This sequence is called a Markov chain if it satisfies the Markov property:

$$P(X_{t+1} = y|X_t = x, ..., X_0 = z) = P(X_{t+1} = y|X_t = x);$$

that is, the value of X_{t+1} is dependent on its history only through its nearest past, X_t. If the form of the transition probability $P(X_{t+1} = y|X_t = x)$ does not change with t, then it is often expressed as a transition function, p_{xy}. Note that

$$\sum_y p_{xy} = 1 \qquad \forall\, x \in \mathfrak{X}. \tag{4}$$

Example 1: Figure 11 representation of a four-state Markov chain with corresponding matrix

$$\begin{pmatrix} p_{11} & p_{12} & 0 & p_{14} \\ p_{21} & 0 & p_{23} & 0 \\ p_{31} & 0 & 0 & p_{34} \\ 0 & 0 & p_{43} & p_{44} \end{pmatrix} \tag{5}$$

Example 2: Let $X_0 = (-1, ..., -1) \in \mathbb{R}^N$. We let X_{t+1} be generated recursively as follows: Randomly pick a coordinate of X_t and negate its current value. Then the sequence $X_0, X_1, ...$ forms a Markov chain. If we write two consecutive states as $x = (x_1, ..., x_N)$ and $y = (y_1, ..., y_N)$, then the transition function for this chain is

$$p_{xy} = 1/N \qquad \text{if } x_i = y_i \text{ for all but one component.}$$

This chain is often referred to as the simple random walk on a N-dimensional cube. For example, when $N = 3$, all the possible configuration this chain is allowed to visit correspond to the eight vertices of a three-dimensional cube.

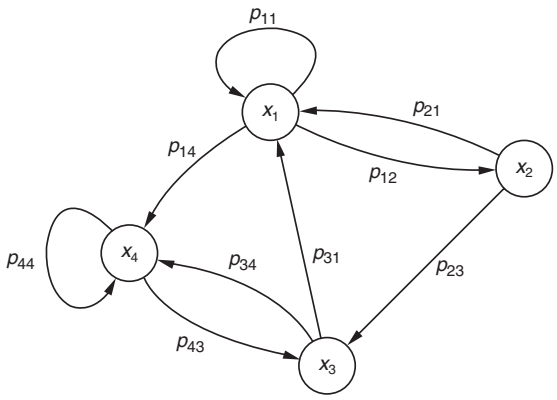

Figure 11: The graph representation of a four-state Markov chain.

Example 3: Simple random walk on a line. Suppose the Z_n are i.i.d. Bernoulli random variables (coin toss) with $P(Z_i = 1) = p$ and $P(Z_i = -1) = 1 - p$. Let $S_0 = 0$ and $S_t = Z_1 + \cdots + Z_t$. Then S_t, $t = 0, 1, ...$ forms a Markov chain.

If the state space is countable, say, $\mathfrak{X} = \{1, ..., N\}$ and the transition probability $p_{ij} = P(X_{t+1} = j | X_t = i)$ is independent of t, we can introduce the transition matrix $P = (p_{ij})$. Given a Markov chain X_t, we can define the state probability vector, or in brief, the probability vector x_t, which is the probability distribution of X_t for any $t \geq 0$, i.e, $x_t(i) = P(X_t = i)$. x_t is a column vector. Understanding the behavior of x_t is a major goal in understanding a Markov chain; in particular, the behavior of x_t for $t \to \infty$ is of interest. Using the transition matrix P we have just introduced, we have

$$x_{t+1} = x_t P. \tag{6}$$

The above equation is called Kolmogorov-Chapman equation. It is clearly true since $x_{t+1}(i) = P(X_{t+1} = i) = \sum_j P(X_t = j) P(X_{t+1} = i | X_t = j) = \sum_j x_t(j) p_{ji}$.

It is easy to see that since the sum of each row of P is 1, P has an eigenvalue 1 with right eigenvector $(1, 1, ..., 1)^\top$. It also has a left eigenvector, which we call it π:

$$\pi = \pi P. \tag{7}$$

Comparing with (6), a probabilistic interpretation of this π is that it is the invariant distribution (also called stationary distribution): Imagine that there is 1 unit of probability "mass", something like sand, and on each repetition of the chain some of the probability mass of a state flows out to another state and some flows in. The amount that flows out is calculated as the amount of probability mass $x_t(i)$ in state i at time t times the fraction that goes to j from i, summed over all j:

$$\text{mass out of state } i \text{ when time changes from } t \text{ to } t+1 = \sum_{j \neq i} x_t(i) p_{ij}.$$

And the amount that flows in from other states is

$$\text{mass into state } i \text{ when time changes from } t \text{ to } t+1 = \sum_{j \neq i} x_t(j) p_{ji}.$$

Invariant distribution is π means $x_t = \pi$ which is independent of time and the mass-out and mass-in are equal for any state i:

$$\pi(i) = \sum_j \pi(i) p_{ij} = \sum_j \pi(j) p_{ji} = [\pi P]_i \qquad \forall i \tag{8}$$

which means exactly $\pi = \pi P$. If π satisfies

$$\pi(i) p_{ij} = \pi(j) p_{ji} \qquad \forall i, j \tag{9}$$

then automatically (8) is true and π is the invariant distribution. Condition (9) is called detailed balance as it means the probability mass from state i to state j equals to the probability mass from state j to state i

Back to the transition matrix P: one can prove (it is called the Perron-Frobenius Theorem. See the book of Jun Liu or the Appendix B of the book of Shonkwiler and Mendivil) *if there is a $\delta > 0$ so that $P_{ij} \geq \delta$ for any i, j*, then (i) all the eigenvalues of P are real; (ii) the largest eigenvalue equals to 1, all the others are strictly smaller than 1 in absolute value; (iii) and P is diagonalizable. So,

$$P = B^{-1} \Lambda B \tag{10}$$

where $\Lambda = \text{diag}(1, \lambda_2, ..., \lambda_N)$ with $1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. From (10), we know that the row vectors of B are the left eigenvectors of P and the column vectors of B^{-1} are the right eigenvectors of P. So the first row of B is π and the the first column of B^{-1} is vector $(1, 1, ..., 1)^\top$.

Obviously, as $n \to \infty$,

$$P^n \to B^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 0 \end{pmatrix} B \tag{11}$$

if and only if $|\lambda_2| < 1$. In this case, direct calculation shows that (See Tutorial Problem Set VI) every row of the limiting matrix P^∞ is the same as π. So, for any column vector α with $[\alpha]_i \geq 0$ and $\sum_i [\alpha]_i = 1$, $\alpha P^\infty = \pi$.

The above argument shows that

$$x_0 P^n \to \pi \qquad \text{as} \qquad n \to \infty \tag{12}$$

where π is the unique invariant distribution if and only if matrix P's second largest eigenvalue in modular is strictly less than 1. It also shows that in the finite state space case, the

convergence rate is geometric (i.e., the "distance" between the distribution of X_t and the target distribution decreases geometrically).

However, it is not so easy to show algebraically that P's second largest eigenvalue in modular is strictly less than 1. Classical Markov chain theory says a Markov chain satisfies the following two conditions at the same time will has (12):

a) *irreducible*: This means, there is non-zero probability of moving from any state to any other state within finite number of steps.

b) *aperiodic*: This means there is no integer $d > 1$ such that the number of iterations between the return from any state to itself must be a multiplier of d.

What can go wrong if the above two conditions are not satisfied? The first condition is a type of mixing condition on the chain, implying that the only nontrivial invariant region is the entire state space. Clearly, if this condition is not satisfies, the state space is disconnected into several components (islands). Then if x_0 is supported on one "island", whenever $x_0 P^t$ converge, it will converge to a vector supported only on this specific "island". In other word, the limit $x_0 P^t$ can be different for different x_0. If the second condition is violated, for example, the chain has period 2. Then $P(x_{i+2k+1} = a | x_i = a) = 0$ which means the diagonal entries on P^{2k+1} must zero. But P^{2k} will not have this restriction. So (11) will not happen.

For a random walk on a N-dimensional cubic, to return to its initial position, the walker must walk even steps. So the Example 2 of Markov chain that we have discussed before is not aperiodic, and $d = 2$.

One can prove that if a finite state Markov chain is irreducible and aperiodic, then it converges to the invariant distribution exponentially fast.

3.8 MCMC and its error: independent or dependent random variables

The most critical step in developing an efficient Monte Carlo algorithm is the simulation (sampling) from an appropriate probability distribution $\pi(x)$. When directly generating independent samples from $\pi(x)$ is not possible, we have learnt or will learn two typical choices: (I) the rejection sampling, in which i.i.d. random samples are generated from a trial distribution different from (but close to) the target one before we decide if the proposed sample will be accepted or not; (II) produce statistically dependent samples based on the idea of Markov chain. In the latter approach, as the samples are not independent, the law of large number or the central limit theorem cannot be applied. However, we can use instead the following ergodic theorem (see page 114 of Shonkwiler and Mendivil. It is like the law of large number.)

Theorem 2 *Let $\{X_n\}$ be an irreducible aperiodic Markov chain with finite state space $\mathfrak{X}$ and stationary distribution π. Let $r : \mathfrak{X} \to \mathbb{R}$ be any function. Then*

$$\frac{1}{N} \sum_{i=1}^{N} r(X_i) \to E_\pi(r(X)) \qquad a.s.$$

for any initial distribution on X_0.

This theorem justifies the use of MCMC in computing expectation values which is the standard use of MCMC. But it is not enough to tell us about the size of the error. For that purpose, we have the following theorem (see the book of Jun Liu, section 12.7. It is like the central limit theorem.)

Theorem 3 *Under the same condition as in the previous theorem,*

$$\sqrt{N}\left(\frac{1}{N}\sum_{i=1}^{N} r(X_i) - E_\pi(r(X))\right) \to N(0, \sigma(r)^2)$$

in distribution for any initial distribution on X_0. *Here* $\sigma(r)^2 = var_\pi r(X)$.

3.9 The Metropolis algorithm

The basic idea behind MCMC is to construct a Markov chain whose invariant distribution is the desired sampling distribution.

Here is the Metropolis algorithm to generate a Markov chain with invariant distribution $f(x)$: Select a state x^0 as the initial state of the chain. Then:
for $t = 0, 1, 2, ...$

- Propose a random "perturbation" of the current state x^t so as to generate a new state y. Mathematically, $x^t \to y$ can be viewed as one step of a Markov chain with transition probability function $P_{x^t,y}$. Here, we additionally require P is symmetric: $P_{x,y} = P_{y,x}$.
- Compute $h = \frac{f(y)}{f(x^t)}$.
- Generate $u \sim U(0,1)$. If $u \le h$, then $x^{t+1} = y$. Otherwise, $x^{t+1} = x^t$.

end

Note that in the book, they use $h = \min\left(1, \frac{f(y)}{f(x^t)}\right)$. But if $\frac{f(y)}{f(x^t)} \ge 1$, $h = 1$ or $\frac{f(y)}{f(x^t)}$ has the same consequence as the y is always accepted because $u \le 1$. But in the following analysis, we use $h = \min\left(1, \frac{f(y)}{f(x^t)}\right)$ because it is used as a probability which should not exceed 1.

Recall the rejection sampling of von Neumann. We see that in some sense, the new Markov chain (called Metropolis chain) is obtained by "rejection sampling" of a Markov chain with transition probability function $P_{x,y}$. We do not even need to know the formula for $P_{x,y}$. We just need it to be symmetric: $P_{x,y} = P_{y,x}$. We can choose the rejection rule which depends on f properly so that the invariant distribution of the new chain is f.

Proof that the Metropolis chain has invariant distribution $f(x)$: We are following the book of Madras. We will verify f satisfies the detailed balance condition (9). Note that

the Metropolis chain is obviously a Markov chain. Denote its transition probability function by Q. Then

$$Q_{x,y} = \begin{cases} \min\left(1, \frac{f(y)}{f(x)}\right) P_{x,y} & \text{if } y \neq x, \\ 1 - \sum_{z \neq x} \min\left(1, \frac{f(z)}{f(x)}\right) P_{x,z} & \text{if } y = x. \end{cases}$$

The precise formula for $Q_{x,y}$ when $x = y$ is not important. Because of the condition $P_{x,y} = P_{y,x}$, when $x \neq y$,

$$f(x)Q_{x,y} = \min\left(f(x), f(y)\right) P_{x,y} = \min\left(f(x), f(y)\right) P_{y,x} = f(y)Q_{y,x}.$$

When $x = y$, obviously $f(x)Q_{x,y} = f(y)Q_{y,x}$. $\square$

In order to apply Theorems 2 and 3, we are left to verify the Metropolis chain is irreducible and aperiodic. (I) If $f(y) > 0$ for all state y that we want to simulate, from the above formula for $Q_{x,y}$, we know if P is irreducible, so is Q. (II) As $\int_{\mathbb{R}^d} f dx = 1$, f goes to zero at infinity. Hence the maximum value of f is achieved at some point x. Then $Q_{x,x} > 0$ as $\frac{f(z)}{f(x)} < 1$ for many z where the chance $P_{x,z} > 0$. (For example, if $\frac{f(z_0)}{f(x)} < 1$ for some $P_{x,z_0} > 0$, then $1 - \sum_{z \neq x} \min\left(1, \frac{f(z)}{f(x)}\right) P_{x,z} \geq \left(1 - \frac{f(z_0)}{f(x)}\right) P_{x,z_0} > 0$.) Hence the chance for the Metropolis chain to stay at x is positive when the time goes from t to $t+1$. This implies the Metropolis chain cannot be periodic.

3.10 Homework III

Homework III, Part I

1) Suppose you want to estimate the volumn of a set B contained in Euclidean space $\mathbb{R}^d$. You know that B is a subset of A and you know the volume of A. The "hit-or-miss" method is to choose n independent points uniformly at random from A, and use the fraction of these points that land in B to get an unbiased estimation of B.

Suppose now that D is a subset of A and we know the volume of D and the volume of $D \cap B$. You decided to estimate the volume of B by choosing n points uniformly at random from $A \backslash D$ and count how many land in B. Show that the second method is better than the method of the preceding paragraph in the sense that it has smaller variance.

2) Let f be a function form the interval $[0,1]$ into the interval $(0,1)$. Here are two possible ways to estimate $I = \int_0^1 f(x)dx$:

a) Use the "hit-or-miss" with n random points. $A = [0,1] \times [0,1]$ and $B = \{(x,y) : 0 \leq x \leq 1, 0 \leq y \leq f(x)\}$.

b) Let U_1, ..., U_n be i.i.d. from the $U(0,1)$ distribution and use the estimator

$$\hat{I}_n = \frac{1}{n} \sum_{i=1}^{n} f(U_i).$$

Show that $\hat{I}_n$ has smaller variance than the estimator of a).

3) Suppose X_i $(i = 1, ..., n)$ are i.i.d. random variables. Let $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} \left(X_i - \bar{X}\right)^2.$$

Show that s_n^2 is an unbiased estimation of the variance of X_1, namely, $E(s_n^2) = \sigma^2 = E(X_1 - EX_1)^2$.

Homework III, Part II (1)

The "Metropolis-Hastings" algorithm is the following generalization of the Metropolis algorithm to the case where the matrix P is not symmetric. Prove that it generates a Markov chain with invariant distribution $f(x)$:

Metropolis-Hastings algorithm

Select a state x^0 as the initial state of the chain. Then:

for $t = 0, 1, 2, ...$

- Propose a random "perturbation" of the current state x^t so as to generate a new state y. Mathematically, $x^t \to y$ can be viewed as one step of a Markov chain with transition probability function $P_{x^t,y}$. Here, we do not require P is symmetric.
- Compute $h = \min\left(1, \frac{P_{y,x^t}}{P_{x^t,y}} \frac{f(y)}{f(x^t)}\right)$.
- Generate $u \sim U(0,1)$. If $u \leq h$, then $x^{t+1} = y$. Otherwise, $x^{t+1} = x^t$.

end

3.11 Computer project III

Part I: Monte Carlo integration : This project is related to Section 3.6.

let $f(x) = 4\sqrt{1-x^2}$, $x \in [0,1]$. We want to estimate $I = \int_0^1 f(x)dx = \pi$ with important sampling. We can choose $g(x) = 1$, $g(x) = \frac{4-2x}{3}$ and $g(x) = 2-2x$. Implement three different ways of sampling and numerically compare how their errors decay. The results handed-in should contain figures like Figure **??**.

Part II: Markov Chain Monte Carlo and the one dimensional Ising model: The Ising model serves to model the behavior of a magnet and is perhaps the best known and the most thoroughly studied model in statistical mechanics. The intuition behind the model is that the magnetism of a piece of material is the collective contribution of dipole moments of many atomic spins within the material. A simple 1d Ising model places these atomic spins on a $1 \times N$ lattice space $\mathfrak{L} = \{i, i = 1, ..., N\}$. In the model, each site $s \in \mathfrak{L}$ hosts a particle that has either a positive or a negative spin. Abstractly, the state of each particle can be represented by a random variable x_s which is either 1 or -1. A configuration of the whole system is then $x = \{x_s, s \in \mathfrak{L}\}$, whose potential energy is defined as

$$U(x) = -J \sum_{s=1}^{N-1} x_s x_{s+1}$$

assuming there is no external magnetic field. J is called the interaction strength. Several important quantities regarding a physical system are often of interest. For example, the internal energy is defined as

$$E_\pi U(x) = \int_D U(x)\pi(x)dx$$

where D is the set of all possible configuration of x and $\pi(x)$ is the pdf that system is in state x. According statistical mechanics (see page 120, eqn (3.16)), $\pi(x)$ is proportional to $\exp\left(-\frac{U(x)}{k_BT}\right)$ where T is the temperature and k_B is the Boltzmann constant. See Page 115 of the book for more details. For your information, we can also assign a stochastic differential equation

$$\frac{dx}{dt} = -\nabla U(x) + \varepsilon \dot{W}(t) \tag{13}$$

to the system. It is more or less the gradient decent algorithm to reach the equilibrium point x_* with $x_* = \text{argmin}U(x)$ but perturbed by a white noise due to the thermal fluctuation. One can prove that the process $x(t)$ that satisfies (13) has an invariant measure with pdf proportional to $\exp\left(-2\varepsilon^{-2}U(x)\right)$. If $\varepsilon = \sqrt{2k_BT}$, it becomes exactly the $\pi(x)$ we mentioned before. From the relation of ε and T, we see that the higher the temperature, the stronger the perturbation. For more details, see any textbook on stochastic differential equation, for example, see page 129 of the book "Random perturbations of dynamical systems" by Freidlin and Wentzel (2nd edition).

If we introduce $Z = \sum_x \exp\left(-\frac{U(x)}{k_BT}\right)$ which is the so called partition function, then $\pi(x) = \frac{\exp\left(-\frac{U(x)}{k_BT}\right)}{Z}$.

For your information, the partition function is perhaps the most important quantity in statistical mechanics. For example, let $\beta = \frac{1}{k_BT}$,

$$\frac{\partial \ln Z}{\partial \beta} = \sum_x -U(x)\frac{\exp\left(-\beta U(x)\right)}{Z} = -E_\pi U(x).$$

For this 1d Ising model, we can derive a closed form of $Z = 2(e^{-\mu} + e^{\mu})^{N-1}$ where $\mu = J\beta = J/(k_BT)$. Now we'd like to derive this formula. First note that for any function $f(x)$ with $x = (x_1, ..., x_N)$, $\sum_x f(x) = \sum_{x_N=1,(x_1,...,x_{N-1})} f(x) + \sum_{x_N=-1,(x_1,...,x_{N-1})} f(x) = \sum_{x_N=\pm 1}\left(\sum_{(x_1,...,x_{N-1})} f(x)\right) = \cdots$
$= \sum_{x_N=\pm 1}\left(\sum_{x_{N-1}=\pm 1}\left(\cdots\left(\sum_{x_1=\pm 1} f(x)\right)\right)\right)$. Hence

$$\begin{aligned} Z &= \sum_x \exp\left(\mu \sum_{s=1}^{N-1} x_s x_{s+1}\right) = \sum_x \prod_{s=1}^{N-1} e^{\mu x_s x_{s+1}} \\ &= \sum_{x_N=\pm 1}\left(\sum_{x_{N-1}=\pm 1}\left(\cdots\left(\sum_{x_1=\pm 1}\prod_{s=1}^{N-1} e^{\mu x_s x_{s+1}}\right)\right)\right). \end{aligned}$$

Next, note that $e^{\mu x_2} + e^{-\mu x_2}$ always equal to $e^{\mu} + e^{-\mu}$ no matter $x_2 = 1$ or -1. Hence $\sum_{x_1=\pm 1}\prod_{s=1}^{N-1} e^{\mu x_s x_{s+1}} = (e^{\mu x_2} + e^{-\mu x_2})\prod_{s=2}^{N-1} e^{\mu x_s x_{s+1}} = (e^{\mu} + e^{-\mu})\left(\prod_{s=2}^{N-1} e^{\mu x_s x_{s+1}}\right)$. Hence by induction, after $N-1$ steps, $Z = \sum_{x_N=\pm 1}(e^{\mu} + e^{-\mu})^{N-1} = 2(e^{\mu} + e^{-\mu})^{N-1}$.

The computer project is to do the Metropolis simulate of the 1-d Ising model with $\mu = 1$ (large T) and then $\mu = 2$ (low T) with 20,000 samples chosen from 1,000,000 Metropolis steps for each μ. The chosen samples are 1 in every 50 lags. Take $N = 50$. Plot the histogram of the magnetic moment

$$\langle M \rangle = \sum_x \left(\sum_{s=1}^N x_s \right) \frac{1}{Z} e^{\mu \sum_{s=1}^{N-1} x_s x_{s+1}} = E_\pi \left(\sum_{s=1}^N x_s \right)$$

where $\pi(x) = \frac{1}{Z} e^{\mu \sum_{s=1}^{N-1} x_s x_{s+1}}$ is the pdf of random variable $x = (x_1, ..., x_N)$.

So the Metropolis algorithm is to keep repeating the following three steps:

- Choose a site, say, the jth site randomly.
- Compute the Metropolis ratio $h = e^{-2\mu x_j^t \left(x_{j-1}^t + x_{j+1}^t \right)}$. Here x_j^t means the jth component of the vector x^t. When $j = 1$ or N, exclude the x_{j-1}^t or x_{j+1}^t term when computing h.
- Generate $u \sim U(0, 1)$. If $u \leq h$, set the spin x_j^t to $-x_j^t$, otherwise, do not change x^t.

For the computer project, you record $M^t = \sum_{s=1}^N x_s^t$ for $t = \texttt{50:50:1e6}$. Then plot its histogram. The project you handed in should include the Matlab code as well as the histograms of M (use 101 bins) for $\mu = 1$ and $\mu = 2$. You can set $x^0 = (1, ..., 1)$.

Homework III, Part II (2): How is the above algorithm related to the Metropolis algorithm in the previous section? Answer this question by figuring out what is f, what is P and what is y for the Metropolis algorithm.

3.12 Tutorial questions set VI

1. Given the pdf's for the following distributions, compute their means and variances:

   (1) We say that X has Bernoulli distribution with parameter p if $P(X = 1) = p$ and $P(X = 0) = 1 - p$.

   (2) We say that X has a Poisson distribution with parameter λ if $P(X = j) = e^{-\lambda}\lambda^j/j!$ for $j = 0, 1, 2, \ldots$.

   (3) X is said to have a geometric distribution with success probability $p \in (0, 1)$ if $P(X = k) = p(1 - p)^{k-1}$ for $k = 1, 2, \ldots$. X is the number of independent trials needed to observe an event with probability p.

   (4) Uniform distribution on (a, b) (denoted by $U(a, b)$). $f(x) = \frac{1}{b-a}$ for $x \in (a, b)$, 0 otherwise.

   (5) Exponential distribution $E(\lambda)$. $f(x) = \lambda e^{-\lambda x}$ for $x \geq 0$, 0 otherwise.

   (6) Standard normal distribution. $f(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$.

2. Show that

   (1) $m = E(X)$ minimizes $E[(X - m)^2]$.

   (2) $E(X) \leq \sqrt{EX^2}$.

   (3) If X, Y are random variables, then $E(XY) \leq \sqrt{E(X^2)E(Y^2)}$. [Hint: use the fact that $E(X + tY)^2 \geq 0$ for all t.]

3. Recall the change of variable formula: Let T be a continuous differentiable map of the open set U onto V. Suppose that T is one-to-one and that $J(x) = \det(\partial T(x)/\partial x) \neq 0$ for all x. Then

$$\int_U f(T(x))|J(x)|dx = \int_V f(y)dy. \tag{14}$$

   Suppose that (X_1, X_2) has pdf

$$f(x_1, x_2) = \frac{1}{2\pi} e^{-\frac{1}{2}(x_1^2+x_2^2)},$$

   and let g be the transformation from $(X_1, X_2) \mapsto (R, \Theta)$ with $R = \sqrt{X_1^2 + X_2^2}$ and Θ is the angle from the positive x-axis to (X_1, X_2). What is the pdf of R and Θ?

4. A random variable X is said to satisfy Cauchy distribution $\mathcal{C}(0, 1)$ if it pdf is $\frac{1}{\pi(1+x^2)}$. Suppose we want to determine the probability that $X > 2$ or equivalently, suppose we want to compute the integral

$$p = \int_2^\infty \frac{1}{\pi(1 + x^2)} dx.$$

One way of estimating p is to generate an iid (independent identically distributed) sample $X_1, X_2, \ldots X_m \sim \mathcal{C}(0,1)$ (we can use the rejection sampling to do that), then an estimation of p is

$$\hat{p}_1 = \frac{1}{m}\sum_{j=1}^{n} \mathbb{I}_{X_j>2}.$$

Here, $\mathbb{I}_{X_j>2} = 1$ if $X_j > 2$ and 0 otherwise. Prove that $E(\hat{p}_1) = p$. By the law of large number, this implies $\hat{p}_1 \to p$ when $n \to \infty$. Let

$$\hat{p}_2 = \frac{1}{2m}\sum_{j=1}^{n} \mathbb{I}_{|X_j|>2}.$$

Prove that $E(\hat{p}_2) = p$. We can also rewrite p as

$$p = \frac{1}{2} - \int_0^2 \frac{1}{\pi(1+x^2)}dx.$$

The integral above can be considered to be the expectation of $h(X) = \frac{2}{\pi(1+X^2)}$ where $X \sim U(0,2)$. Hence an alternative method of evaluation for p is

$$\hat{p}_3 = \frac{1}{2} - \frac{1}{m}\sum_{j=1}^{m} h(U_j)$$

for $U_j \sim U(0,2)$. Moreover, since p can be rewritten as

$$p = \int_0^{\frac{1}{2}} \frac{y^{-2}}{\pi(1+y^{-2})}dy,$$

this integral can be seen as the expectation of $\frac{1}{4}h(Y) = \frac{1}{2\pi(1+Y^2)} = \tilde{h}$ against the uniform distribution on $[0, \frac{1}{2}]$ and another evaluation of p is

$$\hat{p}_4 = \frac{1}{m}\sum_{j=1}^{m} \tilde{h}(Y_j)$$

when $Y_j \sim U(0, \frac{1}{2})$. Since $\frac{d}{dx}\text{atan}x = \frac{1}{1+x^2}$, we indeed can compute p exactly. Represent $\text{var}(\hat{p}_i)$ $(i = 1, 2, 3, 4)$ in terms of m. Which $\hat{p}_i$ has the largest variance and which $\hat{p}_i$ has the smallest variance?

5. Recall the Central Limit Theorem: Let $X_1, X_2, \ldots,$ be independent identically distributed random variables having mean μ and finite variance σ^2. Let $S_n = X_1 + \cdots + X_n$. Then

$$\lim_{n\to\infty} P\left(a \le \frac{S_n - n\mu}{\sigma\sqrt{n}} \le b\right) = P(a \le X \le b)$$

where $X \sim N(0,1)$.

Let S be the number of heads in 1,000,000 tosses of a fair coin. Use (a) Chebyshev's inequality, and (b) the Central Limit Theorem, to estimate the probability that S lies between 499,500 and 500,500. Use the same two methods to estimate the probability that S lies between 499,000 and 501,000, and the probability that S lies between 498,500 and 501,500.

3.13 Tutorial questions set VII

1. Consider the problem of sampling a Bernoulli trial with probability $p = 1/3$ using coin flips. Here is one solution: flip the coin twice; if the outcome is HH, return "success", if HT or TH, return "failure", and if TT, reject the trial and go again. Prove that this gives the right probabilities. (The moral we learned from this problem is that by adding the acceptance/rejection discipline, only the relative values of the target probabilities matter.)

2. (Cauchy from quotient of Normals) Understand the proof in Section 3.3.2 and then prove that if X_1, $X_2 \sim N(0,1)$, then the probability density function of $Y = X_1/X_2$ is $\frac{1}{\pi(1+y^2)}$. Y is called Cauchy distribution random variable. (Note that Y does not have finite variance since $\int_{-\infty}^{\infty} \frac{1}{\pi(1+y^2)} y^2 dy = +\infty$.) [Hint: Let $Y_1 = X_1/X_2$ and $Y_2 = X_2$. Define the mapping $(y_1, y_2) \to (x_1, x_2) = (y_1 y_2, y_2)$. It mapps Ω in the (y_1, y_2) palne to Δ in the (x_1, x_2) plane. We have $P((Y_1, Y_2) \in \Omega) = \int_\Omega f_{Y_1,Y_2}(y_1, y_2) dy_1 dy_2 = P((X_1, X_2) \in \Delta) = \int_\Delta f_{X_1,X_2}(x_1, x_2) dx_1 dx_2 = \int_\Omega f_{X_1,X_2}(y_1 y_2, y_2)|J| dy_1 dy_2$. In the last step, we have used the change of variables and $|J| = |\det \frac{\partial(x_1,x_2)}{\partial(y_1,y_2)}| = |y_2|$. Then integrate Y_2 out of the joint pdf.]

3. Will stratified sampling always reduce the variance no matter how we partition the D? Find an example that stratified sampling reduce the variance. [The key is how to choose m_i. For a related answer, see Theorem 3.2 and 3.3 of "Lectures on Monte Carlo methods" by Madras.]

4. (Control variate method) Suppose the estimation of $\mu = EX$ is of interest and $\mu_C = EC$ is known. Then we can construct Monte Carlo samples of the form

$$X(b) = X - b(C - \mu_C)$$

   which has the same mean as X. If the computation of $\mathrm{cov}(X, C)$ and $\mathrm{var}(C)$ is easy, then we can let $b = \frac{\mathrm{cov}(X,C)}{\mathrm{var}(C)}$, show that $\mathrm{var}(X(b)) < \mathrm{var}(X)$.

5. For the Antithetic Variate Method, show that $\mathrm{var}(X + X') < 2\mathrm{var}(X)$. [Hint: if g is monotonic function, then

$$\{g(u_1) - g(u_2)\}\{g(1 - u_1) - g(1 - u_2)\} \leq 0$$

   for u_1, $u_2 \in [0, 1]$. Show that for two independent uniform random variable U_1 and U_2, $2\mathrm{cov}(X_1, X_1') = E\left(\{g(U_1) - g(U_2)\}\{g(1 - U_1) - g(1 - U_2)\}\right) \leq 0$ where $X_1 = F^{-1}(U_1)$ and $X_1' = F^{-1}(1 - U_1)$.]

3.14 Tutorial questions set VIII

1. Suppose we want to use rejection sampling to sample from an unbounded beta distribution whose pdf is

$$f(x) = cx^{\alpha-1}(1-x)^{\alpha-1}, \qquad 0 \le x \le 1$$

for some given $\alpha < 1$. Here c is a constant so that $\int_0^1 f(x)dx = 1$. Prove that $f(x) = f(1-x)$, which means f is symmetric about $x = 1/2$. So, we need only sample X from the interval $[0, 1/2)$ and, with probability $1/2$, return either X or $1 - X$. Therefore, as the proposed density, we use

$$g(x) = \tilde{c}x^{\alpha-1}, \qquad 0 \le x \le \frac{1}{2}.$$

Determine $\tilde{c}$ by $\int_0^{1/2} g(x)dx = 1$. Recall Theorem 1, which says that if $U \sim U(0,1)$, then $X = F^{-1}(U)$ has cdf F. Propose a method which samples from $g(x)$.

2. (Continue with Question 1). Propose the rejection sampling to sample from $f(x) = cx^{\alpha-1}(1-x)^{\alpha-1}$ for $x \in [0,1]$.

3. (1) Given the transition matrix

$$P = \begin{pmatrix} 0 & 0.4 & 0.6 & 0 & 0 \\ 0.65 & 0 & 0.35 & 0 & 0 \\ 0.32 & 0.68 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0.12 & 0.88 \\ 0 & 0 & 0 & 0.56 & 0.44 \end{pmatrix},$$

examine whether the corresponding Markov chain is irreducible and aperiodic.

(2) Given the transition matrix

$$P = \begin{pmatrix} 0 & 0.4 & 0.6 & 0 & 0 \\ 0.65 & 0 & 0.35 & 0 & 0 \\ 0.32 & 0.68 & 0 & 0 & 0 \\ 0 & 0 & 0.12 & 0 & 0.88 \\ 0.14 & 0.3 & 0 & 0.56 & 0 \end{pmatrix},$$

show that the corresponding chain is aperiodic, despite the null diagonal.

4. A random walk on the non-negative integers $I = \{0, 1, 2, 3, \ldots\}$ can be constructed in the following way. For $0 < p < 1$, let $Y_0, Y_1, \ldots$ be iid random variables with $P(Y_i = 1) = p$ and $P(Y_i = -1) = 1 - p$, and $X_k = \sum_{i=0}^{k} Y_i$. Then $\{X_n\}$ is a Markov chain with transition probabilities

$$P(X_{i+1} = j + 1 | X_i = j) = p \qquad P(X_{i+1} = j - 1 | X_i = j) = 1 - p,$$

but we make the exception that $P(X_{i+1} = 1 | X_i = 0) = p$ and $P(X_{i+1} = 0 | X_i = 0) = 1 - p$

(1) Show that $\{X_n\}$ is a Markov chain.

(2) Show that $\{X_n\}$ is irreducible.

(3) Show that the invariant distribution of the chain is given by

$$a_k = \left(\frac{p}{1-p}\right)^k a_0, \qquad k = 1, 2, 3, ...,$$

where a_k is the probability that the chain is at k and a_0 is arbitrary. For what value of p and a_0 is this a probability distribution?

5. (This is taken from the book of Shonkwiler and Mendivil) Our description and motivation of the Metropolis algorithm start with the rejection sampling method. Recall that in the rejection sampling method we have a target density f, a proposal density g, and an acceptance discipline h. Assume, for convenience, that f is a discrete probability density with outcome space Ω. Samples from f will be approximated by the sequence of states $X_1, X_2, \ldots$ of a Markov chain over Ω. As in the rejection method, generating the next state is a two-part process: a proposal followed by acceptance or rejection. By regarding the process as a chain, transitions from state to state may depend on the value of the current state. Thus, both the proposal process and the acceptance process can depend on the current state.

   This means that $g_x(\cdot)$ must be defined for every $x \in \Omega$, but in fact, this is often easy to do, and is guided by the particular application (or distribution) in question. In the simplest case, $g_x = g$ of the rejection method for all x, but then we get nothing new (and the associated Markov chain will be very simple, with each iteration independent of the previous one).

   The other difference compared to the rejection method is that the acceptance discipline also depends (in general) on the current state x as well as the proposed state y. In fact, the acceptance discipline for the Metropolis algorithm is

$$h(x, y) = \min\left(1, \frac{f(y)}{f(x)}\right)$$

   This says that if the proposed state y has greater probability than the current state x, $f(y) \geq f(x)$, then accept it with certainty. If not, then accept it with probability equal to the ratio $f(y)/f(x)$. As before with the rejection method, only the relative values of f are needed.

   Here is the Metropolis algorithm (assuming $g_x(y) = g_y(x)$):

   Select a state x.

   for $t = 0, 1, 2, ...$

   - select y using density g_x
   - put $h = \min(1, f(y)/f(x))$

– if $U \sim U(0,1) < h$, then $x = y$

– else x unchanged

end

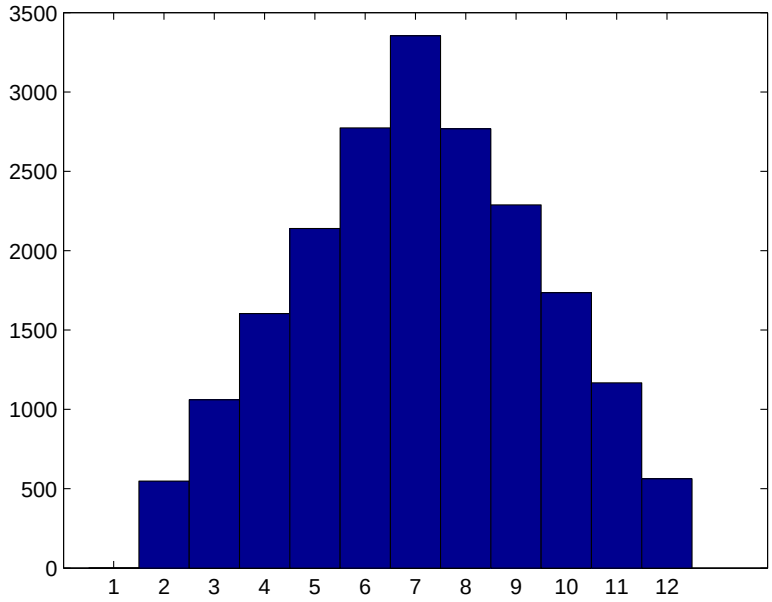

Figure 12: Computational result with $f = [0, 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1]$, 20,000 samples.

Here is a simple Matlab code with $\Omega = \{2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12\}$. What is the target distribution f? What is g_x? If the first line is changed to $f = [0, 5, 4, 3, 2, 1, 1, 1, 2, 3, 4, 5]$, how would the graph looks like?

```
f=[0, 1, 2, 3, 4, 5, 6, 5, 4, 3, 2, 1];
d=zeros(1,20000); % for histogramming
x=5;
for i = 1:20000
    U = rand;
    if x == 2
        if U < 0.5
            y = 3;
        else
            y = 2;
        end
    elseif x == 12
        if U < 0.5
            y = 11;
        else
```

```
            y = 12;
        end
    else
        if U < 0.5
            y = x-1;
        else
            y = x+1;
        end
    end
    h = min(1,f(y)/f(x));
    U = rand;
    if U < h
        x = y;
    end
    d(i) = x; % record the state
end
a = 1:1:12;
hist(d,a)
[N,h] = hist(d,a)
```

3.15 Tutorial questions set IX

1. Determine the stationary distribution of the Markov chain with transition probability matrix

$$\begin{pmatrix} c_0 & \frac{1}{2}e^{-3} & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2}e^{-1} & c_2 & \frac{1}{2} \\ 0 & 0 & \frac{1}{2}e^{-1} & c_3 \end{pmatrix}.$$

2. Given the transition matrix P, suppose there is an invertible matrix B so that

$$P = B^{-1}\Lambda B \tag{15}$$

where $\Lambda = \text{diag}(1, \lambda_2, ..., \lambda_N)$ with $|\lambda_i| < 1$ for $i = 2, ..., n$. Prove the row vectors of B are the left eigenvectors of P and the column vectors of B^{-1} are the right eigenvectors of P. Then prove the first row of B is the invariant distribution π and the the first column of B^{-1} is vector $(1, 1, ..., 1)^\top$. Prove that when $n \to \infty$,

$$P^n \to P^\infty \stackrel{def}{=} B^{-1} \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ & & \ddots & \\ 0 & 0 & \cdots & 0 \end{pmatrix} B. \tag{16}$$

Then prove that every row of P^∞ is π. So, for any column vector α with $[\alpha]_i \geq 0$ and $\sum_i [\alpha]_i = 1$, $\alpha P^\infty = \pi$.

3. Consider a probability density g on $[0, 1]$ and a function $0 < \rho < 1$ such that $\int_0^1 \frac{g(x)}{1-\rho(x)} dx < \infty$. Let δ_x is the Dirac mass at x, i.e., $\delta_x(x') = 1$ when $x' = x$ and $\delta_x(x') = 0$ otherwise. Prove that the Markov chain with transition probability

$$P_{x,x'} = \rho(x)\delta_x(x') + (1 - \rho(x))g(x')$$

has stationary distribution

$$f(x) \propto g(x)/(1 - \rho(x)).$$

4. How is the following algorithm related to Metropolis-Hastings algorithm and the rejection sampling of von Neumann and what is its stationary distribution?

   Select a state x^0 as the initial state of the chain. Then:

   for $t = 0, 1, 2, ...$

   - Generate a new random variable y with pdf $g(\cdot)$.
   - Compute $h = \frac{f(y)g(x^t)}{f(x^t)g(y)}$.
   - Generate $u \sim U(0, 1)$. If $u \leq h$, then $x^{t+1} = y$. Otherwise, $x^{t+1} = x^t$.

   end

   The above algorithm is called independent Metropolis-Hastings algorithm because the proposed Markov chain is independent of x^t.

5. (continue of 4.) [This problem is a little bit hard.] Prove that if there is an M such that $f(x) \leq Mg(x)$ for all x and if the chain x^t has reached its stationary distribution (means t large enough), then the acceptance probability of the above algorithm is at least $\frac{1}{M}$, i.e., $E\left(\min\left(\frac{f(y)g(x^t)}{f(x^t)g(y)}, 1\right)\right) \geq \frac{1}{M}$. How is this result compared with the rejection sampling?

3.16 Review problems

1. If $A = I - 2uu^\top$ with $u \in \mathbb{R}^{n\times 1}$ and $\|u\|_2 = 1$, what is $\|A\|_2$?

2. Consider the Ising model in the last homework and computer project. (It means the x_j^t's in the following can be either 1 or -1.)

   (a) If you start from some vector $x^t = (x_1^t, ..., x_n^t)$. Choose a site, say, the jth site randomly and then flip x_j^t to $-x_{j+1}^t$ to get a new vector x^{t+1}. If you keep doing this, what is the invariant distribution of the Markov chain x^t.

   (b) Now, suppose you only interested in those x^t which satisfies $\sum_{j=1}^n x_j^t \geq \frac{3}{5}n$. Denotes the set of those x^t's by S. If you start from some vector $x^t = (x_1^t, ..., x_n^t) \in S$. Choose a site, say, the jth site randomly and then flip x_j^t to $-x_{j+1}^t$ to get a new vector y. If $y \in S$, set $x^{t+1} = y$. Otherwise, set $x^{t+1} = y$. If you repeating this process, what is the invariant distribution of the Markov chain x^t?

3. Suppose you estimate $I = \int_0^1 f(x)dx$ by $\hat{I}_n = \frac{1}{n}\sum_{i=1}^n f(U_i)$ with $U_i \sim U(0,1)$. When n is large, how is $|I - \hat{I}_n|$ related to n?

4. In your computer project, how do you generate the random variable on $[0,1]$ with given pdf $g(x) = \frac{4-2x}{3}$?

5. What is the bisection method for finding the root of $f(x)$? (See Lab tutorial II)

3.17 Lab tutorial: III

Read the Hit-or-miss integration example in the book of Shonkwiler and Mendivil(from page 37) and do the following:

(1) Plot the surface $z = \sin(x)\sin(y)$ and then estimate the volume of the body below this surface but above $A = \{(x, y) : 0 \leq x \leq 1, 0 \leq y \leq 1\}$. Compare your result with the exact value $\int_A z dx dy = (1 - \cos(1))^2$. Plot how the error changes when you number of trials increasing according to $2^{1:20}$. Generate some kind of plots like the Fig.1.13 in the book.

(2) Use the Matlab function to generate uniform random numbers (`rand`) and random numbers with normal distribution (`randn`). Plot the histogram (`hist`) and understand the meaning of the x and y axis. How is histogram related to the pdf?

3.18 Lab tutorial: IV

Read section 2.7.1 of the book of Shonkwiler and Mendivil and implement a code that draw samples from $Be(2,3)$ with pdf $f(x) = 12x(1-x)^2$, $x \in [0,1]$. Plot the histogram and compare it with the "scaled" pdf. How should you choose the scaling factor? [Hint: note that $f(x)dx = \frac{\text{number of r.v. in } dx}{\text{total number of r.v.'s}}$].