MSMF Summer Boot Camp Lecture Notes

TABLE OF CONTENTS

Chapter 1

Calculus review

1.1 Techniques of integration

1.1.1 Theory

Integration by parts:

$$\int udv = uv - \int vdu.$$

Integration by substitution:

$$\int f'(g(x))g'(x)dx = f(g(x)) + c.$$

Switching the order of integration and differentiation : under certain conditions of the function $f(x)$ we have

$$\frac{d}{dx}\int f(u,x)du = \int \frac{\partial}{\partial x} f(u,x)du.$$

1.1.2 Problems

1. a) $\int xe^x dx$
   b) $\int_0^\infty e^{-2\sqrt{x}}dx$
   c) $\int \log(x)dx$
   d) $\int e^x \sin x dx$
   e) $\int \tan^{-1} x dx$
   f) $\int_0^\infty xe^{-x^2}dx$
   g) $\int_{-\infty}^\infty e^{-x^2}dx$

   2. $\int_{\mathbb{R}} e^{tx}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$. This is the moment generating function of a Normal distribution.

3*. $\int_{\mathbb{R}} e^{itx} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$. This is the characteristic function of a Normal distribution.

4. $\int_0^\infty x \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} dx$ and $\int_0^\infty x^2 \frac{1}{\lambda} e^{-\frac{1}{\lambda}x} dx$. Derive the expectation and variance of an exponential (λ) distribution.

5. Let T, K, and σ be positive constants, let r be a nonnegative constant, and let $x > 0$ and $t \in [0, T)$ be given. Define $\tau = T - t$, which is positive. Let Y be a standard normal random variable. Compute

$$c(t,x) = \mathbb{E}\left[e^{-r\tau} \left(x \exp\left\{ \sigma\sqrt{\tau}\, Y + \left(r - \frac{1}{2}\sigma^2 \right) \tau \right\} - K \right)^+ \right].$$

In other words, let

$$\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$$

be the standard normal density. Compute

$$c(t,x) = \int_{-\infty}^{\infty} e^{-r\tau} \left(x \exp\left\{ \sigma\sqrt{\tau}\, y + \left(r - \frac{1}{2}\sigma^2 \right) \tau \right\} - K \right)^+ \varphi(y)\, dy.$$

You should obtain the Black-Scholes formula

$$c(t,x) = xN\big(d_+(\tau,x)\big) - Ke^{-r\tau} N\big(d_-(\tau,x)\big),$$

where

$$d_\pm(\tau,x) = \frac{1}{\sigma\sqrt{\tau}} \left[\log\frac{x}{K} + \left(r \pm \frac{1}{2}\sigma^2 \right) \tau \right]$$

and $N(d)$ is the cumulative standard normal distribution $N(d) = \int_{-\infty}^{y} \varphi(y)\, dy$.

6. Evaluate the integral

$$F(\alpha) = \int_0^1 \frac{x^\alpha - 1}{\log x} dx \quad (\alpha \geq 0)$$

by first finding $F'(\alpha)$.

7. Evaluate the integral

$$F(\alpha) = \int_0^\infty e^{-x} \frac{\sin \alpha x}{x} dx$$

by first finding $F'(\alpha)$.

8*. Show that

$$F(\alpha) = \int_0^\infty e^{-x^2} \cos(\alpha x) dx = \frac{\sqrt{\pi}}{2} e^{-\frac{\alpha^2}{4}}$$

by first finding a first order ODE that $F(\alpha)$ satisfies.

1.2 Fundamental theorem of Calculus

1.2.1 Theory

1. $\int_a^b f'(x)dx = f(b) - f(a)$.
2. $\frac{d}{dx}\int_a^x f(u)du = f(x)$.
3. Leibniz integral formula:

$$\frac{d}{dx}\int_a^x f(u,x)du = f(x,x) + \int_a^x \frac{\partial}{\partial x} f(u,x)du.$$

1.2.2 Problems

1. Generalize the FTC to

$$\frac{d}{dx}\int_{f(x)}^{g(x)} F(u)du.$$

2. Generalize the Leibniz integral formula to

$$\frac{d}{dx}\int_{f(x)}^{g(x)} F(u,x)du.$$

3. Find $F'(x)$ for the function $F : \mathbb{R} \to \mathbb{R}$ defined by
a)

$$F(x) \triangleq \int_{x^3}^{e^{x^2}} \sin(t^2)dt, \quad x \in \mathbb{R}.$$

b*)

$$F(x) \triangleq \int_{\frac{1}{x}}^{\frac{2}{x}} \frac{\sin(xt)}{t}dt.$$

4. Define the function

$$f(x) \triangleq \int_{-\infty}^{e^{2x}} e^{-\frac{\left(y-\sin(x)\right)^2}{2}}dy$$

Compute f's derivative at the point $x = 0$. There can be no integrals in your answer.
5*. Write down the formula for $F(x) := \int_{-\infty}^{x} f(u)du$ if

$$\begin{aligned} f(x) &= 0, x < 0 \\ &= x, 0 \leq x \leq 1 \\ &= 2 - x, 1 < x \leq 2 \\ &= 0, x > 2. \end{aligned}$$

Where is f differentiable / not differentiable?

6. The function $N(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}} du$ is well defined (it is the cdf of the Normal distribution).

a) Verify that $N(x)$ is differentiable and increasing everywhere.

b) Analyze the concavity of $N(x)$.

c) Compute $-\int_{-\infty}^{0} N(x)dx + \int_{0}^{\infty}[1 - N(x)]dx$

d) Generalize the result in part c to the cdf of other distributions.

1.3 Taylor formula and series

1.3.1 Theory

Let $f(x)$ be an infinitely differentiable fucction. The nth order Taylor series expansion for a function $f(x)$ with remainder around $x = x_0$ is:

$$f(x) = \sum_{k=0}^{n} f^{(k)}(x_0) \frac{(x - x_0)^k}{k!} + f^{(n+1)}(\bar{x}) \frac{(x - x_0)^{n+1}}{(n+1)!},$$

for some $x_0 < \bar{x} < x$. The term $R_n := f^{(n+1)}(\bar{x}) \frac{(x-x_0)^{n+1}}{(n+1)!}$ is the remainder of the nth order expansion. If there is $L > 0$ such that $R_n \to 0$ as $n \to \infty$ for $x \in (x_0 - L, x_0 + L)$ then we say $f(x)$ is analytic at x_0 (or simply $f(x)$ has a Taylor series expansion around x_0) and we write

$$f(x) = \sum_{k=0}^{\infty} f^{(k)}(x_0) \frac{(x - x_0)^k}{k!}.$$

One should keep in mind that not all functions have a Taylor series expansion around a point x_0 (even if $f(x)$ is infinitely differentiable at x_0). A standard example is the function

$$\begin{aligned} f(x) &= e^{-1/x^2}, x > 0 \\ &= 0, x \leq 0. \end{aligned}$$

It can be verified that $f^{(n)}(0) = 0, n \geq 0$ and thus the Taylor series expansion of f around 0 is identically 0. On the other hand, $f(x)$ is not identically 0 and thus $f(x)$ is not analytic at $x = 0$.

On the other hand, the nth order Taylor series expansion of $f(x)$ is always valid, as long as f has $n+1$ continuous derivatives.The point to consider when using the nth order Taylor expansion then is how large the remainder R_n is (how close the finite series is to the actual function).

Taylor series and integration / differentiation: If the function $f(x)$ is analytic around x_0 :

$$f(x) = \sum_{n} a_n (x - x_0)^n$$

then for any x within the radius of convergence

$$f'(x) = \sum_n n a_n (x - x_0)^{n-1}$$

and

$$\int_{x_0}^{x} f(s)ds = \sum_n \frac{a_n (x - x_0)^{n+1}}{n+1}.$$

1.3.2 Problems

1. Write the Taylor series expansion around $x = 0$ for $\cos(x), \sin(x), e^x, \log(1+x), \sqrt{1+x}$. Use Taylor series to find

$$\begin{aligned}
&\lim_{x\to 0} && \frac{\sin x}{x} \\
&\lim_{x\to 0} && \frac{1-\cos x}{x^2} \\
&\lim_{x\to \infty} && \frac{e^x}{x^n} \\
&\lim_{x\to 0} && \frac{x - \log(1+x)}{x^2} \\
&\lim_{x\to 0} && \frac{x - \log(1+x)}{x^2} \\
&\lim_{x\to 0} && \frac{\sqrt{1+x} - 1 - \frac{1}{2}x}{x^2}.
\end{aligned}$$

2. Prove the Euler's formula: $e^{ix} = \cos x + i \sin x$.

3. The distribution function of a Poisson(λ) random variable is

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, k = 0, 1, \cdots$$

Show that this is a true distribution i.e. $\sum_k P(X = k) = 1$.

4. Compute $\sum_k k e^{-\lambda} \frac{\lambda^k}{k!}$. This is the expectation of a Poisson random variable. $\sum_k k^2 e^{-\lambda} \frac{\lambda^k}{k!}$. Derive the variance of a Poisson RV.

5. Compute $\sum_{k\geq 1} k(1-p)^{k-1}p$. This is the expectation of a geometric RV. Compute $\sum_{k\geq 1} k^2(1-p)^{k-1}p$. Derive the variance of a geometric RV*.

6. Show that $(1+x)^n \geq 1 + nx$ for all $x > -1$ and $n \geq 2$.

7. Use Taylor series to show that the solution the ODE

$$y' = y, y(0) = 1$$

is $y = e^x$.

8*. Use Taylor series to show that the solution the ODE

$$y'' = -y, y(0) = 0, y'(0) = 1$$

is $y = \sin x$.

9. Let $p > 0$ be given and consider the function $a : (0, \infty) \to \mathbb{R}$ defined by

$$a(t) \triangleq \sum_{n=1}^{\infty} e^{-n^p t}, \quad t > 0.$$

You may take it for granted that there exist constants C and α (possibly depending on p) such that

$$\lim_{t \downarrow 0} \frac{a(t)}{Ct^{\alpha}} = 1.$$

a) Prove that C and α are unique.

b) Determine α and C. *Hint: Compare the sum with and integral from* 0 *to* ∞ *and change the variables in the integral.*

c) For which values of p does the integral $\int_0^1 a(t)dt$ converge?

10*. Define the function

$$f(x) \triangleq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \sin(xy) e^{-\frac{y^2}{2}} dy$$

Provide the first order Taylor expansion of $f(x)$ around the point $x = 0$. There can be no integrals in your answer.

11. Let

$$N(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{y^2}{2}} dy$$

be the cumulative standard normal distribution. Work out the first four terms in the Taylor expansion of $N(x)$ around the point $x = 0$.

1.4 Partial derivatives

1.4.1 Theory

See a standard calculus text.

1.4.2 Problems

1. For a European call expiring at time T with strike price K, the Black-Scholes price at time $t \in [0, T)$, if the time-t stock price is x, is (see also the problem in the integration section)

$$c(t, x) = xN\big(d_+(T-t, x)\big) - Ke^{-r(T-t)} N\big(d_-(T-t, x)\big),$$

where

$$d_\pm(\tau, x) = \frac{1}{\sigma\sqrt{\tau}}\left[\log\frac{x}{K} + \left(r \pm \frac{1}{2}\sigma^2\right)\tau\right]$$

and $N(y)$ is the cumulative standard normal distribution

$$N(d) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{d} e^{-y^2/2}dy.$$

Here T, K and σ are positive constants and r is a nonnegative constant. The purpose of this problem is to show that the function $c(t, x)$ satisfies the Black-Scholes partial differential equation

$$c_t(t, x) + rxc_x(t, x) + \frac{1}{2}\sigma^2x^2c_{xx}(t, x) = rc(t, x),\ 0 \le t < T, x > 0, \tag{1.1}$$

the *terminal condition*

$$\lim_{t\uparrow T} c(t, x) = (x - K)^+, \quad x > 0, x \neq K, \tag{1.2}$$

and the *boundary conditions*

$$\lim_{x\downarrow 0} c(t, x) = 0, \quad \lim_{x\to\infty}\left[c(t, x) - \left(x - e^{-r(T-t)}K\right)\right] = 0, \quad 0 \le t < T. \tag{1.3}$$

Equations (1.2) and the first part of (1.3) are usually written more simply but less precisely as

$$c(T, x) = (x - K)^+,\ x \ge 0$$

and

$$c(t, 0) = 0,\ 0 \le t \le T.$$

For this exercise, we abbreviate $c(t, x)$ as simply c and $d_\pm(x, T - t)$ as simply $d_\pm$.

Verify first the equation

$$Ke^{-r(T-t)}N'(d_-) = xN'(d_+). \tag{1.4}$$

a. Show that $c_x = N(d_+)$. This is the *delta* of the option. (Be careful! Remember that d_+ is a function of x.)

b. Show that

$$c_t = -rKe^{-r(T-t)}N(d_-) - \frac{\sigma x}{2\sqrt{T-t}}N'(d_+).$$

This is the *theta* of the option.

c. Use the above formulas to show that c satisfies (1.1).

d. Show that for $x > K$, $\lim_{t\uparrow T} d_\pm = \infty$, but for $0 < x < K$, $\lim_{t\uparrow T} d_\pm = -\infty$. Use these equalities to derive the terminal condition (1.2).

e. Show that for $0 \le t < T$, $\lim_{x\downarrow 0} d_\pm = -\infty$. Use this fact to verify the first part of boundary condition (1.3) as $x \downarrow 0$.

f. Show that for $0 \leq t < T$, $\lim_{x\to\infty} d_{\pm} = \infty$. Use this fact to verify the second part of boundary condition (1.3) as $x \to \infty$. In this verification, you will need to show that

$$\lim_{x\to\infty} x\big(N(d_+) - 1\big) = 0.$$

2. Itô's formula: Suppose that

$$dS_t = rS_t dt + \sigma S_t dW_t$$

(a formal expression) and it holds that

$$df(t, S_t) = f_t dt + f_x dS_t + \frac{1}{2} f_{xx}\sigma^2 S_t^2 dt.$$

Find $d(S_t)^2, d(tS_t), d(\log S_t), de^{S_t}$.

1.5 Multiple integrals

1.5.1 Theory

Change of variable formula: The following is the chain rule in multi-dimensinal setting. Let $\mathbf{F}$ be a mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. For example, when $n = 2, \mathbf{F}(r, \theta) = (r\cos\theta, r\sin\theta)$ is the change from polar coordinates to rectangular coordinates. Let $\mathbf{u} = \mathbf{F}(\mathbf{x})$. We have

$$\mathbf{du} = \mathbf{F}'(\mathbf{x})\mathbf{dx}.$$

Here $\mathbf{du}, \mathbf{dx}$ are 2 vectors of n differentials in $\mathbb{R}^n$ and $\mathbf{F}'(\mathbf{x})$ is a $n \times n$ matrix of partial derivatives, also known as the Jacobian. $\mathbf{F}'(\mathbf{x})$ (for a fixed $\mathbf{x}$) can be thought of as a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$. It is a well known result in linear algebra that if A is a linear mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ and $V_{\mathbf{x}}$ is the volume of a parallelpiped generated by $\mathbf{x}$ then

$$V_{\mathbf{Ax}} = |det(A)|V_{\mathbf{x}}.$$

Thus in terms of the "volume" generated by the the differentials $\mathbf{du}, \mathbf{dx}$ we have

$$du_1 du_2 \cdots du_n = |det(\mathbf{F}'(\mathbf{x}))| dx_1 dx_2 \cdots dx_n.$$

Let f be a mapping from $\mathbb{R}^n$ to $\mathbb{R}$. The change of variable formula is

$$\iint_{S_1} f(\mathbf{F}(\mathbf{x}))|det(\mathbf{F}'(\mathbf{x}))| dx_1 dx_2 \cdots dx_n = \iint_{S_2} f(\mathbf{u}) du_1 du_2 \cdots du_n,$$

where S_1, S_2 are regions of $\mathbb{R}^n$ and $S_2 := \{\mathbf{F(x)}, \mathbf{x} \in S_1\}$ and we require that $det(\mathbf{F}'(\mathbf{x})) \neq \mathbf{0}$ over S_1 (or equivalently $\mathbf{F}'(\mathbf{x})$ is invertible over S_1).

Remark: The change of variable formula is usually used in the reverse direction where we start from

$$\iint_{S_2} f(\mathbf{u}) du_1 du_2 \cdots du_n$$

and want to convert to a nicer region of integration S_1 via a change of variable **u = F(x)**. For example, the integral

$$\iint_{x^2+y^2\leq 1} dxdy = 1$$

over the unit disc (S_2) in the x, y coordinates is converted to the "rectangular region" $0 \leq r \leq 1, 0 \leq \theta \leq 2\pi$ in the r, θ coordinates.

1.5.2 Problems

1. a) Find c such that

$$\iint_{x^2+y^2\leq 1} cdxdy = 1.$$

b) Compute

$$\iint_{x^2+y^2\leq 1} xydxdy$$
$$\iint_{x^2+y^2\leq 1} xdxdy$$
$$\iint_{x^2+y^2\leq 1} ydxdy.$$

Remark: This computes the covariance of X, Y which has uniform distribution on the unit circle. We shall see that the covariance is zero but X, Y are not independent.

2. Compute

$$\int_0^z \int_0^{z-y} \mu\lambda e^{-\lambda x-\mu y} dxdy.$$

Remark : This computes the cdf of the random variable $Z = X + Y$ where X, Y have independent exponential distributions with parameters λ, μ.

3. Use Leibniz integral formula to compute

$$\frac{\partial}{\partial z}\big[\int_0^z \int_0^{z-y} \mu\lambda e^{-\lambda x-\mu y} dxdy\big].$$

This is the density of the variable Z described in part 2. Compare the result with the one obtained by differentiating the answer in part 2.

4. Use the given transformation to evalute the integrals:
a)

$$\iint_R (x-3y)dxdy, x = 2u+v, y = u+2v,$$

R is the triangular region with vertices $(0,0),(2,1),(1,2)$.
b)

$$\iint_R (4x+8y)dxdy, x = \frac{1}{4}(u+v), y = \frac{1}{4}(v-3u)$$

R is the parallelogram with vertices $(-1,3),(1,-3),(3,-1),(1,5)$.
c)

$$\iint_R x^2 dxdy, x = 2u, y = 3v,$$

R is the region bounded by the ellipse $9x^2+4y^2 = 36$.
d)

$$\iint_R x^2 - xy + y^2 dxdy, x = \sqrt{2}u - \sqrt{\frac{2}{3}}v, y = \sqrt{2}u + \sqrt{\frac{2}{3}}v.$$

R is the region bounded by the ellipse $x^2 - xy + y^2 = 2$.

1.6 Numerical differentiation

1.6.1 Theory

Finite difference : First derivative approximation:

$$\begin{aligned} f'(x) &\approx \frac{f(x+h)-f(x)}{h}, h > 0 \\ &\approx \frac{f(x)-f(x-h)}{h}, h > 0 \\ &\approx \frac{f(x+h)-f(x-h)}{2h}, h > 0. \end{aligned}$$

Second derivative approximation (Central difference formula) :

$$f''(x) \approx \frac{f(x+h)-2f(x)+f(x-h)}{h^2}, h > 0.$$

Partial derivative approximation: First order partial derivatives

$$\begin{aligned}\frac{\partial}{\partial x}f(x,y) &\approx \frac{f(x+h,y)-f(x,y)}{h}, h>0\\ &\approx \frac{f(x,y)-f(x-h,y)}{h}, h>0\\ &\approx \frac{f(x+h,y)-f(x-h,y)}{2h}, h>0.\end{aligned}$$

Second order partial derivatives:

$$\begin{aligned}\frac{\partial^2}{\partial x^2}f(x,y) &\approx \frac{f(x+h,y)-2f(x,y)+f(x-h,y)}{h^2}, h>0\\ \frac{\partial^2}{\partial x\partial y}f(x,y) &\approx \frac{1}{4hk}\Big[f(x+h,y+k)-f(x+h,y-k)\\ &- f(x-h,y+k)+f(x-h,y-k)\Big], h,k>0\end{aligned}$$

1.6.2 Problems

1. Explain the reasoning for the finite difference formulas for $f'(x)$. What is the difference in the three formulas? What are the order of the error terms ?

2. Use the finite difference formula to approximate $f'(1)$ where $f(x)=\sqrt{x}$ using $h=0.1, 0.01, 0.001$. Calculate the errors in each approximation.

3. Explain the reasoning for the central difference formula. What is the order of the error term in the central difference formula? ($\frac{h^2}{12}$ by Taylor's series).

4. Use the central difference formula to approximate $f''(1)$ where $f(x)=\sqrt{x}$ using $h=0.1, 0.01, 0.001$. Calculate the errors in each approximation.

5*. Develop an approximation for $f''(x)$ using $f(x), f(x+3h), f(x-h)$. (Expand $f(x+3h), f(x-h)$ around x and eliminate the $f'(x)$ terms).

6. Explain the reasoning for the mixed partial derivatives formula. What is the order of the error term?

7. Use numerical differentiation to find first and second order partial derivatives for $\sqrt{x+y}$ at $x=y=1$, with $h,k=0.1, 0.01, 0.001$.

1.7 Numerical Integration

1.7.1 Theory

Rectangular rule (Riemann sum approximation) :

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_i f(x_i)(x_{i+1}-x_i)\\ &\approx \sum_i f(x_{i+1})(x_{i+1}-x_i)\\ &\approx \sum_i f(\frac{x_i+x_{i+1}}{2})(x_{i+1}-x_i).\end{aligned}$$

These are referred to as the right point, left point and mid point approximation respectively.

Error term bound (midpoint rule) : $\frac{(b-a)(\Delta x)^2}{24}f^{(2)}(\xi), \xi \in [a,b]$.

Trapezoidal rule:

$$\int_a^b f(x)dx \approx \sum_i \frac{1}{2}\big[f(x_i)+f(x_{i+1}\big](x_{i+1}-x_i).$$

Error term bound: $\frac{(b-a)(\Delta x)^2}{12}f^{(2)}(\xi), \xi \in [a,b]$.

Simpson's rule :

$$\begin{aligned}\int_a^b f(x)dx &\approx \big[f(a)+4f(\frac{a+b}{2})+f(b)\big]\frac{(b-a)}{6}\\ &= \big[f(a)+4f(\frac{a+b}{2})+f(b)\big]\frac{\Delta x}{3}, \Delta x = \frac{b-a}{2}.\end{aligned}$$

Simpson's composite rule (equal spacing) : for $a = x_0 < x_1 < \cdots < x_n = b$ where n is even

$$\begin{aligned}\int_a^b f(x)dx &\approx \sum_i \frac{\Delta x}{3}\big[f(x_{2i-2})+4f(x_{2i-1})+f(x_{2i})\big]\\ &= \frac{\Delta x}{3}\big[f(a)+4f(x_1)+2f(x_2)+4f(x_3)+\cdots+4f(x_{n-1})+f(b)\big].\end{aligned}$$

Error term bound: $\frac{(b-a)(\Delta x)^4}{180}f^{(4)}(\xi), \xi \in [a,b]$.

Simpson's 3-8 rule :

$$\begin{aligned}\int_a^b f(x)dx &\approx \big[f(a)+3f(\frac{2a+b}{3})+3f(\frac{a+2b}{3})+f(b)\big]\frac{(b-a)}{8}\\ &= \big[f(a)+3f(\frac{2a+b}{3})+3f(\frac{a+2b}{3})+f(b)\big]\frac{3\Delta x}{8}, \Delta x = \frac{b-a}{3}.\end{aligned}$$

Simpson's 3-8 composite rule (equal spacing): for $a = x_0 < x_1 < \cdots < x_n = b$ where n os a multiple of 3

$$\begin{aligned}\int_a^b f(x)dx &\approx \frac{3\Delta x}{8}\big[f(a) + 3f(x_1) + 3f(x_2) + 2f(x_3) + \cdots \\ &\quad +3f(x_4) + 3f(x_5) + 2f(x_2) + \cdots + 3f(x_{n-2}) + 3f(x_{n-1}) + f(b)\big].\end{aligned}$$

Error term bound: $\frac{(b-a)(\Delta x)^4}{80} f^{(4)}(\xi), \xi \in [a, b]$.

1.7.2 Problems

1. Compute the following limits

a)

$$\lim_{n\to\infty} \sum_{k=1}^{n} \frac{1}{n}\sqrt{5 + \frac{2k}{n}}.$$

b)

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{i^4}{n^5} + \frac{i}{n^2}.$$

c)*

$$\lim_{n\to\infty} \sum_{i=1}^{n} \frac{\sqrt{i}}{n\sqrt{n}}.$$

2. Use all methods of integration described above to find up to 6 digit precision a)

$$\int_0^{0.4} \sqrt{1 + x^4}dx.$$

b)

$$\int_0^2 \frac{e^x}{1 + x^2}dx.$$

c)

$$\int_0^{\frac{\pi}{2}} \sqrt[3]{t}1 + \cos x dx.$$

d)

$$\int_0^4 x^3 \sin x dx.$$

e)

$$\int_0^1 \frac{x^2}{1+x^4}dx.$$

3. Use Taylor's series expansion to find up to 6 digit precision

$$\int_0^{0.4} \sqrt{1+x^4}dx.$$

Compare this approach to the above.

4. Find examples of functions for whose integrals the rectangular rule, trapezoidal rule, Simpson's rule and Simpson's 5-8 rule are exact. Explain the reason why they are exact in these cases.

5. Compare the errors in all the methods.

6. Relate the midpoint rule with the rectangular left and right hand rule.

7. We refer to Trapezoidal rule (Midpoint rule, Simpson's rule) with n step approximation as T_n (M_n, S_n respectively).

a) If f is a positive function and $f''(x) < 0$ on $[a, b]$ show that

$$T_n < \int_a^b f(x)dx < M_n$$

b) If f is a polynomial of degree 3 or lower then Simpson's rule is exact.

c*) Show that $\frac{1}{2}(T_n + M_n) = T_{2n}$.

d*) Show that $\frac{1}{3}T_n + \frac{2}{3}M_n = S_{2n}$.

1.8 Numerical solution to nonlinear equation

1.8.1 Theory

Bisection method: Consider the equation $f(x) = 0$. We start out with a_0, b_0 such that $f(a_0) < 0$ and $f(b_0) > 0$. Denote $m_{i+1} = \frac{a_i+b_i}{2}, i = 0, 1, \cdots$. The algorithm continues with $a_{i+1} = m_{i+1}, b_{i+1} = b_i$ if $f(m_{i+1}) < 0$ and $a_{i+1} = a_i, b_{i+1} = m_{i+1}$ if $f(m_{i+1}) > 0$. The underlying idea is by the intermediate value theorem, the value x_0 such that $f(x_0) = 0$ is always within (a_i, b_i) for any i. Since $b_i - a_i$ decreases by half after each iteration, they converge to x_0 with a linear rate of convergence.

Newton's method: Consider the equation $f(x) = 0$. We choose an initial value x_0. The algorithm continues with

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

Newton's method is not guaranteed to converge. It relies on the first guess x_0 to be sufficiently close to the root. It also requires $f(x)$ to be differentiable around the root. If

Newton's method converges it convereges quadratically. Newton's method comes from the Taylor approximation around the root $x = a$:

$$f(x) \approx f(a) + f'(a)(x - a) = f'(a)(x - a).$$

Thus

$$a \approx x - \frac{f(x)}{f'(a)} \approx x - \frac{f(x)}{f'(x)}.$$

1.8.2 Problems

1. Find $\sqrt{2}$ using Bisection method and Newton's method. Write a code to implement the algorithm to 6 digit precision. Compare the rate of convergence of the two methods.

2. Find the root of the equation $\tan^{-1}(x) = x - 1$ to 6 digit precision using Newton's method. Use $x_0 = 2$ as the initial guess.

3. Consider the equation $x^5 - x^3 + 2x^2 - 1 = 0$. Use Newton's method to find the roots of the equation to 6 digit precision. Try different initial points and remark on the results. For the first initial point, try $x_0 = 1$.

4*. Apply Newton's method to the equation $x^2 - a = 0$ to derive the following square-root algorithm :

$$x_{n+1} = \frac{1}{2}(x_n + \frac{a}{x_n}).$$

(This algorithm was known to the ancient Babylonians).

5. Apply Newton's method to the equation $\frac{1}{x} - a = 0$ to derive the following reciprocal algorithm :

$$x_{n+1} = 2x_n - ax_n^2.$$

This algorithm enables a computer to find reciprocals without actually dividing. Apply this to compute $\frac{1}{1.15}$ correct to six decimal places.

6.

a) Explain why Newton's method doesn't work for finding the root of

$$x^3 - 3x + 6 = 0$$

if the initial approximation is $x_0 = 1$.

b) Explain why Newton's method doesn't work for finding the root of

$$\sqrt[3]{x} = 0$$

for initial approximation $x_0 \neq 0$.

8. Recall that the Black-Scholes formula for a European call option with parameters S_0, r, T, K, σ is

$$\begin{aligned} c^{BS}(S_0, r, T, K, \sigma) &= S_0 N(d+) - K e^{-rT} N(d-) \\ N(d\pm) &= \frac{r \pm \frac{1}{2}\sigma^2 T - \log \frac{K}{S_0}}{\sigma\sqrt{T}}. \end{aligned}$$

The impplied volatility $\sigma^{implied}$ of a particular call option is defined such that for this option market's price c^{market},

$$c^{market} = c^{BS}(S_0, r, T, K, \sigma^{implied}).$$

In other words, $\sigma^{implied}$ is the solution to the equation

$$c^{BS}(S_0, r, T, K, \sigma) = c^{market}.$$

For $S_0 = 60, r = 0.05, T = 1, K = 50, c^{market} = 12.54$ use the bisection method to find $\sigma^{implied}$, starting with $\sigma_L = 0.05$ and $\sigma_R = 0.15$ (This would be an inefficient way to solve for the implied volatility. It is simply to demonstrate an application of numerical solution method in math finance.)

Chapter 2

Linear algebra review

2.1 Gaussian elimination, linear independence

2.1.1 Theory

See a standard linear algebra text.

2.1.2 Problems

1. Solve the linear system

$$\begin{aligned} x_1 - x_2 - 2x_3 &= 2 \\ 2x_1 + 4x_2 + 5x_3 &= 1 \\ 6x_1 - 3x_3 &= 3. \end{aligned}$$

2. Let

$$S = \left\{ \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}; \begin{bmatrix} 2 \\ 1 \\ 1 \end{bmatrix}; \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \right\}$$

Is span(S) = $\mathbb{R}^3$? Explain.

3.

$$S = \left\{ \begin{bmatrix} 1 \\ 2 \\ 1 \end{bmatrix}; \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix}; \begin{bmatrix} 1 \\ -2 \\ -1 \end{bmatrix} \right\}$$

Is S linearly independent ? Explain.

4. Let

$$A = \begin{bmatrix} 1 & 3 & -5 \\ 1 & 4 & -8 \\ -3 & -7 & 9 \end{bmatrix}; \mathbf{b} = \begin{bmatrix} 4 \\ 7 \\ -6 \end{bmatrix}$$

a) Does $A\mathbf{x} = \mathbf{b}$ have a unique solution?

b) If possible, write the solution to $A\mathbf{x} = \mathbf{b}$ in parametric vector form. If not possible, just stay not possible.

5*. Let

$$S = \left\{ \begin{bmatrix} 2 \\ 1 \\ -2 \end{bmatrix}; \begin{bmatrix} -3 \\ 1 \\ 4 \end{bmatrix} \right\}$$

and $\mathbf{y} = \begin{bmatrix} h \\ 2 \\ 1 \end{bmatrix}$. For what value of h is $\mathbf{y}$ in span(S)?

2.2 Eigenvalues and eigenvectors

2.2.1 Theory

Let A be a n $\times$ n matrix. $\lambda \in \mathbb{R}$ is called an eigenvalue and $\mathbf{v} \neq \mathbf{0} \in \mathbb{R}^n$ is called an eigenvector (corresponding to λ) if

$$A\mathbf{v} = \lambda\mathbf{v}.$$

The eigenvalues of A are the roots of the equation $det(A - \lambda I) = 0$. The eigenvectors corresponding to a specific eigenvalue λ can be found by plugging in the particular value of λ and solve for the nonzero solutions of the linear system $(A - \lambda I)\mathbf{v} = 0$. In other words, they are the basis vectors of the Null space of $A - \lambda I$).

General properties of eigenvalues and eigenvectors: A $n \times n$ matrix has exactly n eigenvalues (possibly complex) counting repetitions. There is at least one and at most m_λ eigenvectors corresponding to the eigenvalue λ where m_λ is the multiplicity of λ. Eigenvectors corresponding to distinct eigenvalues are linearly independent.

Diagonalization of a square matrix: A matrix A has a diagonal form if there exists a diagonal matrix D and invertible matrix P such that $A = PDP^{-1}$. We say that A is the diagonal matrix D. In this case the diagonal of D consists of the eigenvalues of A and the columns of P consist of the eigenvectors of A. A matrix A has a diagonal form if and only if it has n independent eigenvectors.

Jordan form: Suppose that A does not have a full set of eigenvectors. This means that A is not similar to a diagonal matrix. In this case, A is still similar to a near diagonal matrix, the so-called Jordan form of A. This is a matrix that has the eigenvalues of A on the diagonal and ones in certain positions on the diagonal above the main diagonal and zero elsewhere. In particular, let λ be an eigenvalue with multiplicity $m \geq 2$ and suppose the number of eigenvectors corresponding to λ is $k < m$. Let $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_k$ be k eigenvectors corresponding to the eigenvalue λ. It is a fact that the null space of $(A - \lambda I)^m$ has dimension m. That is there are m linearly independent vectors satisfying

$(A-\lambda I)^m\mathbf{v}=0$. We already know k of them: $\mathbf{v}_1,\mathbf{v}_2,\cdots,\mathbf{v}_k$. This means that there are $k-m$ other independent vectors satisfying $(A-\lambda I)^m=0$. These are referred to as the generalized eigenvectors corresponding to λ. We need another concept : $\mathbf{v}$ is a generalized eigenvector of rank $d\geq 1$ corresponding to λ if

$$\begin{aligned}(A-\lambda I)^d\mathbf{v}&=0\\(A-\lambda I)^{d-1}\mathbf{v}&\neq 0.\end{aligned}$$

Now if $\mathbf{v}$ is a generalized eigenvector of rank d then $(A-\lambda I)\mathbf{v}$ is a generalized eigenvector of rank $d-1$ and $(A-\lambda I)^2\mathbf{v}$ a generalized eigenvector of rank $d-2$ etc. These form the so called generalized eigenvector chain. In the above scenario, we only need to find a generalized eigenvector of rank $m-k+1$ and generate the chain from there to obtain $k-m$ generalized eigenvectors. Thus we see that a $n\times n$ matrix has n indpendent generalized eigenvectors. Let $P=[\mathbf{v}_1\quad\mathbf{v}_2\quad\cdots\mathbf{v}_n]$ be the matrix formed by these chains of generalized eigenvectors. We see that

$$AP=PJ$$

where J is the Jordan form described above. In particular, if $\mathbf{v}_i,\mathbf{v}_{i+1}$ are consecutive members of the generalized eigenvector chain we have

$$(A-\lambda I)\mathbf{v}_i=\mathbf{v}_{i+1}.$$

That is $A\mathbf{v}_i=\lambda\mathbf{v}_i+1\mathbf{v}_{i+1}$. This is why the Jordan form has 1 above the main diagonal. These correspond to the positions of the generalized eigenvectors. Note that the last generalized eigenvector in a chain must be an actual eigenvector. This is also reflected in the fact that the last column of $\mathbf{v}_n$ of P must be an actual eigenvector since it must satisfy $A\mathbf{v}_n=\lambda_n\mathbf{v}_n$. The Jordan decomposition of A is $A=PJP^{-1}$.

Conditions for invertibility of a matrix: from the definition of eigenvalues, it is easy to see that a matrix is invertible if and only if all of its eigenvalues are non-zero. On the other hand, computation of eigenvalues (and eigenvectors) can be time consuming (it is an important topic in numerical linear algebra). The following is a more easily verified (although rather strong) condition of invertibility : A matrix A is strictly diagonally row (column) dominant if the absolute value of its diagonal entries are strictly greater than the sum of the absolute values of the corresponding row (column) :

$$\begin{aligned}|A_{ii}| &> \sum_{j\neq i}|A_{ij}| \text{ (strict row dominant)}\\ |A_{ii}| &> \sum_{j\neq i}|A_{ji}| \text{ (strict column dominant) .}\end{aligned}$$

Row (column) strict dominance implies that the eigenvalues of A are nonzero via the so called Gershgorin's theorem.

Tridiagonal symmetric matrices: These matrices arise in discretization of partial differential equations (PDEs). In particular, let A be an $n \times n$ matrix such that $A(i,i) = a, A(i,i+1) = A(i,i-1) = -b, \forall i, A(i,j) = 0$ otherwise then the eigenvalues of A are

$$\lambda_i = a - 2b\cos\left(\frac{\pi i}{n+1}\right), i = 1, \cdots, n.$$

The corresponding eigenvectors $\mathbf{v}_i$ has the entries

$$\mathbf{v}_i(j) = \sin\left(\frac{\pi ij}{n+1}\right), j = 1, \cdots, n.$$

2.2.2 Problems

1. Let $\mathbf{v}_1, \mathbf{v}_2$ be two eigenvectors corresponding to distinct eigenvalues λ_1, λ_2. Show that $\mathbf{v}_1 \neq c\mathbf{v}_2$ for some constant c. Generalize this to the case of $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ be eigenvectors corresponding to distinct eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_n$. Show that $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_n$ are independent.

2. Verify that if A is strictly row dominant then A^T is strictly column dominant and vice versa.

3. Check invertibility of the following matrices. Explain your reasoning.

a)

$$\begin{bmatrix} 4 & -2 & -1 & 0 \\ 2 & -5 & 1 & 1 \\ -2 & 1 & 5 & 1 \\ 1 & 1 & 0 & 3 \end{bmatrix}$$

b)

$$\begin{bmatrix} 6 & -3 & 2 & 1 \\ 2 & 4 & 1 & -3 \\ 1 & 0.5 & 5 & 2 \\ -2 & 0 & -1 & 7 \end{bmatrix}$$

c)*

$$\begin{bmatrix} -4 & 3 & -1 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 0 & 2 & 2 \\ 0 & 0 & 0 & 5 \end{bmatrix}$$

4. Find the eigenvalues and eigenvectors of the following matrices:

a)

$$\begin{bmatrix} 2 & -1 \\ 1 & 4 \end{bmatrix}$$

b)

$$\begin{bmatrix} 2 & 0 & 0 \\ 1 & -3 & 0 \\ -1 & 2 & 1 \end{bmatrix}$$

c)*

$$\begin{bmatrix} -2 & -1 & 3 \\ 0 & 1 & 2 \\ 0 & 0 & 3 \end{bmatrix}$$

5. Let A be a matrix with a given set of eigenpairs.

a) Compare the eigenpairs of A with the eigenpairs of $A + \varepsilon I$ for some $\varepsilon > 0$.

b) Show that if A is singular (non-invertible) we can find an $\varepsilon > 0$ as small as desired such that $A + \varepsilon I$ is nonsingular (invertible) (this method of perturbing the diagonal is very useful in practical situations).

6*. Let A, B be two square matrices. Show that AB and BA have the same eigenvalues by :

a) Assuming that either A or B is non-singular and show that $det(AB-\lambda I) = det(BA-\lambda I)$.

b) Remove the assumption in part a by using question 5.

7. The strictly dominant assumption for invertibility is necessary. A weaker version of the concept is : A matrix A is weakly diagonally row (column) dominant if the absolute value of its diagonal entries are greater than or equal to the sum of the absolute values of the corresponding row (column) :

$$\begin{aligned} |A_{ii}| &\geq \sum_{j \neq i} |A_{ij}| \text{ (weak row dominance)} \\ |A_{ii}| &\geq \sum_{j \neq i} |A_{ji}| \text{ (weak column dominance) .} \end{aligned}$$

a) Show that

$$A = \begin{bmatrix} 4 & 2 & -1.5 \\ 0 & 2 & -2 \\ 0 & -1.5 & 1.5 \end{bmatrix}$$

is a weakly diagonally dominant singular matrix.

b) Verify that we can perturb the diagonal to make A strictly dominant, hence nonsingular.

8. Let A be a 4×4 tridiagonal matrix such that $A(i,i) = 2, A(i,i+1) = A(i,i-1) = -1$. Investigate the norm and the orthogonality of the eigenvectors of A.

9. Let J be a $n \times n$ matrix such that $J(i,i) = a, J(i,i+1) = b, J(i,j) = 0$ otherwise (J is NOT tridiagonal). J is called a Jordan block if $b = 1$. Find the eigenvalues and eigenvectors of J.

10. The forward Euler finite difference scheme for the heat PDE corresponds to the tridiagonal matrix $A(i,i) = 1 - 2h, A(i,i+1) = A(i,i-1) = h, h > 0$. The scheme is convergent if and only if

$$\|A\|_2 := \max |\lambda_i| < 1.$$

Show that this condition is satisfied if and only if $0 < h \leq \frac{1}{2}$.

11. The backward Euler finite difference scheme for the heat PDE corresponds to the tridiagonal matrix $A(i,i) = 1 + 2h, A(i,i+1) = A(i,i-1) = -h, h > 0$. The scheme is convergent if and only if

$$\|A^{-1}\|_2 := \max |\lambda_i| < 1,$$

where A^{-1} is the inverse of A.

a) Show that the discretization matrix A is indeed invertible.

b) Show that the convergence condition is satisfied for any $h > 0$.

12. Find the Jordan form of A and express $A = PJP^{-1}$ where A is

a)

$$\begin{bmatrix} 0 & 4 & 2 \\ -3 & 8 & 3 \\ 4 & -8 & -2 \end{bmatrix}$$

b)

$$\begin{bmatrix} 5 & 1 & -2 & 4 \\ 0 & 5 & 2 & 2 \\ 0 & 0 & 5 & 3 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

2.3 Symmetric matrices, symmetric positive definite matrices and Covariance matrices

2.3.1 Theory

A matrix A is symmetric if $A^T = A$. A symmetric positive definite (spd) matrix (or simply positive definite) is a symmetric matrix with the additional condition that

$$\mathbf{v}^T A \mathbf{v} > 0, \forall \mathbf{v}.$$

(Some authors include nonsymmetric matrix in the definition of positive definite, but most often positive definite matrix implies symmetry). If we replace $>$ with $\geq$ in the above equation then A is positive semi-definite. Similarly one can define negative definite and

negative semidefinite matrices. Positive definite matrices arise as covariance matrices of a multivariate distribution or in the context of least square regression.

The following are some conditions for checking positive-definiteness (positive semi-definiteness):

a) All eigenvalues of A are > 0 (≥ 0)

b) Sylvester's criterion: all leading principal minors of A are > 0 (≥ 0) (Principle minors are the determinant of the square matrices obtained by going down the diagonal of A)

c) A has a Cholesky decomposition (see next section).

Of all these conditions, the Cholesky decomposition is the most practical computationally (via the Cholesky decomposition algorithm with cost $n^3 + O(n^2)$). Eigenvalues computation is more expensive and can be imprecise (for matrix with small eigenvalues). Sylvester criterion is more of a theoretical result to apply for matrix of small dimension without using computational methods.

Some properties of symmetric and positive-definite (semi-definite) matrix:

a) The eigenvalues of symmetric matrices are real.

b) Eigenvectors corresponding to distinct eigenvalues of a symmetric matrix are orthogonal.

c) A symmetric matrix A has a diagonal form : $A = QDQ^T$ where Q is an orthogonal matrix (composed of eigenvectors of A),

d) If A is a $m \times n$ matrix (not necessarily square) then A^TA is a positive semidefinite matrix. A^TA is positive definite if the columns of A are linearly independent (the matrix A^TA arises in least square solution of $Ax = b$).

e) A positive definite matrix is invertible. Its inverse is also positive definite.

f) A strictly (weakly) diagonally dominant symmetric matrix with positive diagonal entries is positive definite (semi-definite).

g) Let A be a tri-diagonal symmetric matrix $A(i,i) = d, A(i,i+1) = A(i,i-1) = -a, \forall i, A(i,j) = 0$. A is positive definite if and only if

$$d > 2|a| \cos\left(\frac{\pi}{n+1}\right).$$

Least square solutions

The least square solution to $Ax = b$ is $\hat{x}$ such that $\|A\hat{x} - b\| \leq \|Ax - b\|$ for any $x \in \mathbb{R}^n$, where

$$\|x\|^2 = x^Tx, x \in \mathbb{R}^n.$$

One can see that $A\hat{x}$ is the orthogonal projection of b onto the column space of A (draw a picture where A is one single column). That is $b - A\hat{x}$ is orthogonal to $c_i, i = 1, \cdots, n$ where c_i is the ith column of A. That is

$$c_i^T(b - A\hat{x}) = 0, i = 1, \cdots, n.$$

The above equations are equivalent to $A^T(b - A\hat{x}) = 0$ or $\hat{x}$ is the solution to the normal equation

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

Here we see again the symmetric matrix $A^T A$ discussed above. The point of the least square solution is that while there may not exist a solution to $Ax = b$, there exists a point $\hat{b}$ in the column space of A that is closest to b (again $\hat{b}$ is the orthogonal projection of b onto the column space of A). Thus the normal equation is always consistent : there exists at least one $\hat{x}$ such that $A\hat{x} = \hat{b}$. The issue may be whether this $\hat{x}$ is unique. If the columns of A are linearly independent then $A^T A$ is nonsingular and $\hat{x}$ is unique.

Example: Least square linear regression. Given a set of points $(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)$ we want to find a line $y = ax + b$ that minimizes

$$\sum_n (ax_i + b - y_i)^2.$$

This can easily be seen as a least square problem with $A = [\mathbf{x} \quad \mathbf{1}]$ and $b = \mathbf{y}$.

Covariance and correlation matrices

Let $\mathbf{X} := (X_1, \cdots, X_n)$ be n given random variables. A is the covariance matrix of $\mathbf{X}$ if $A_{ij} = cov(X_i, X_j)$ for any (i,j) and B is the correlation matrix of $\mathbf{X}$ if $B_{ij} = corr(X_i, X_j)$ for any (i,j). Since $cov(X_i, X_j) = cov(X_j, X_i)$ and $corr(X_i, X_j) = corr(X_j, X_i)$ it is clear that covariance and correlation matrices are symmetric. Moreover, since $corr(X_i, X_j) = \frac{cov(X_i, X_j)}{\sqrt{Var(X_i)Var(X_j)}}$ a correlation matrix is a covariance matrix whose diagonal entries are 1's.
A covariance matrix must be positive semi-definite since for any constant vector $c \in \mathbb{R}^n$

$$Var(c^T \mathbf{X}) = c^T cov(\mathbf{X}) c \geq 0.$$

Conversely a positive semi-definite matrix can be viewed as a covariance matrix of some random vector. A positive semi-definite matrix with diagonal entries being 1's can be viewed as a correlation matrix.

Application: Principal component analysis in portfolio risk management

Given a covariance matrix Σ of a multivariate distribution $\mathbf{X}$ we see that $\Sigma = QDQ^T$ where Q is an orthogonal matrix. We make the additional assumption that D is arranged from the highest eigenvalues to the lowest. If we let $\mathbf{v_1}$ be the first column of Q then

$$Cov(\mathbf{v_1}^T \mathbf{X}) = \mathbf{v_1}^T \Sigma \mathbf{v} = \lambda_1.$$

Thus $\mathbf{v}_1$ is the direction that captures $\frac{\lambda_1}{\sum_i \lambda_i}$ percentage in variation of $\mathbf{X}$. If λ_1 is big compared to other λ's this is a high percentage. Usually one only need to use the first

few eigenvectors (referred to as factors) to capture the most variations of $\mathbf{X}$. Moreover, if $\lambda_1 \neq \lambda_2$ $\mathbf{v}_1$ is orthogonal to $\mathbf{v}_2$. Then

$$Cov(\mathbf{v}_1^T\mathbf{X}, \mathbf{v}_2^T\mathbf{X}) = \mathbf{v}_1^T\Sigma\mathbf{v}_2 = 0.$$

Thus $\mathbf{v}_i^T\mathbf{X}, \mathbf{v}_j^T\mathbf{X}$ are uncorrelated (and independent if the underlying multivariate distribution is normal) if $\mathbf{v}_i, \mathbf{v}_j$ corresponds to different eigenvalues. Thus the direction $\mathbf{v}_1^T + \mathbf{v}_2^T$ captures $\frac{\lambda_1+\lambda_2}{\sum_i \lambda_i}$ percentage in variation of $\mathbf{X}$. The eigenvectors are referred to as the factors and their entries the factor loadings.

An alternative way to look at the explanation power of the factors is to write X as a linear combination of the factors. Since Q is invertible, a particular realization $\mathbf{x}$ of $\mathbf{X}$ can be written as

$$\mathbf{x} = Q\mathbf{f}.$$

$\mathbf{f}$ is referred to as the vector of factor scores (for a particular instance of $\mathbf{x}$). We have $\mathbf{f} = Q^T\mathbf{x}$ and thus $f_i = \mathbf{v}_i^T\mathbf{x}$ as discussed above. Thus the variance of the ith factor score is just λ_i. Here we say that

$$\hat{\mathbf{x}} = \sum_{i=1}^{k} f_i\mathbf{v}_i$$

for a small k approximates $\mathbf{x}$ well in terms of explaining the variation of the original $\mathbf{x}$. Observe also the fact that the factor scores are uncorrelated.

In the financial context, suppose we have a portfolio of the stocks in an index(e.g. S&P 500). We can calculate the change of the portfolio value with respect to the movement of a particular stock, but this is not a good way to capture the portfolio's exposure. Using PCA, we estimate Σ, the covariance matrix of the log returns of the stocks in the index. Decompose $\Sigma = QDQ^T$ where the entries of D is arranged from the highest to lowest. The columns of Q are the factors discussed above. The portfolio exposure can be captured in a more efficient manner by looking at its exposure to the first few factors. In particular, let $\mathbf{\Delta\Pi}$ be the vector of the deltas of the portfolio with respect to individual names (more specifically with respect to the log returns of the individual names). Let $\mathbf{R}$ be the (random) vector of the log returns of the individual names. The total portfolio change is

$$\mathbf{\Delta\Pi}^T\mathbf{R}.$$

On the other hand, we can decompose $\mathbf{R}$ using the factors and factor scores. That is

$$\mathbf{R} = \sum_{i=1}^{n} f_i\mathbf{v}_i.$$

Then the portfolio change can be approximated by

$$\sum_{i=1}^{k} f_i\mathbf{\Delta\Pi}^T\mathbf{v}_i,$$

where k is a small number. We refer to this quantity as the exposure of the portfolio to the first k factors. One quantity that captures the risk is the variance (or standard deviation) of the exposure to the k factors. Since the factors are uncorrelated, the variance of the exposure to the first k factors are

$$\sum_{i=1}^{k} (\mathbf{\Delta \Pi}^T \mathbf{v}_i)^2 \lambda_i.$$

Note that PCA is not about efficiency of computation since one can easily calculate

$$Var(\mathbf{\Delta \Pi}^T \mathbf{R}) = \mathbf{\Delta \Pi}^T \Sigma \mathbf{\Delta \Pi}.$$

PCA gives the portfolio manager the information about the direction that the portfolio is "most exposed" to. The manager can use this information for hedging purpose, for example, to immunize the portfolio against the possible shocks in the short run.

2.3.2 Problems

1. Prove (or find out about the proofs) of the statements in the theory section above.

2. Diagonalize:

a)

$$A = \begin{bmatrix} 1 & 2 \\ 2 & 4 \end{bmatrix}.$$

b)*

$$A = \begin{bmatrix} 1 & 0 \\ 6 & -1 \end{bmatrix}.$$

c)

$$A = \begin{bmatrix} 4 & 2 & 2 \\ 2 & 4 & 2 \\ 2 & 2 & 4 \end{bmatrix}.$$

3. Find the least square solution to $Ax = b$ where

a)

$$A = \begin{bmatrix} -1 & 2 \\ 2 & -3 \\ -1 & 3 \end{bmatrix}$$

$$b = \begin{bmatrix} 4 \\ 1 \\ 2 \end{bmatrix}$$

b)*

$$A = \begin{bmatrix} 2 & 1 \\ -2 & 0 \\ 2 & 3 \end{bmatrix}$$

$$b = \begin{bmatrix} -5 \\ 8 \\ 1 \end{bmatrix}$$

c)

$$A = \begin{bmatrix} 1 & 3 \\ 1 & -1 \\ 1 & 1 \end{bmatrix}$$

$$b = \begin{bmatrix} 5 \\ 1 \\ 0 \end{bmatrix}$$

4.

a) Let A be a symmetric positive definite matrix. Use the diagonal form of A to find a matrice B such that $B^2 = A$ (B can be thought of as the "square-root" of A).

b) Show that a matrix A is symmetric positive semidefinite if and only if there exists a symmetric matrix B such that $B^2 = A$.

c) Let A be a positive semi-definite matrix and B such that $B^2 = A$ as in part b). Show that

$$(Ax, y) = (Bx, By), x, y \in \mathbb{R}^n$$

(Thus B really acts like a square root of A when it comes to inner product).

d) Show that

$$(Ax, y)^2 \leq (Ax, x)(Ay, y), x, y \in \mathbb{R}^n.$$

5. Show that for any constant vector $c \in \mathbb{R}^n$ and a random vector $\mathbf{X} \in \mathbb{R}^n$

$$Var(c^T\mathbf{X}) = c^T cov(\mathbf{X})c \geq 0.$$

6. Let

$$A = \begin{bmatrix} 4 & 2 & 1 \\ 2 & 2 & 1.1 \\ 1 & 1.1 & 1 \end{bmatrix}.$$

Is A a covraince matrix? If yes, what is the correlation matrix corresponding to A?

7. Let

$$A = \begin{bmatrix} 1 & 0.5 & 0.2 \\ 0.5 & 1 & \rho \\ 0.2 & \rho & 1 \end{bmatrix}.$$

Find the range of ρ such that A is a correlation matrix.

8. Computational errors can cause the result in the computation of a covariance matrice to be not positive semi-definite. Give an example of an "almost" positive semi-definite matrix. Suggest a (reasonable) modification so that if A is an "almost" positive semi-definite matrix then the modification of A is positive semi-definite.

2.4 LU decomposition, QR decomposition, Cholesky decomposition

A big area of numerical linear algebra is finding numerical solution to $Ax = b$ efficiently. The majority of the techniques rely on iterative methods, which is beyond the scope of this review. On the other hand, these iterative techniques involve the idea of decomposing A into "simpler" components such as triangular or diagonal matrices via a sum structures. This section covers fundamental decompositions of the form $A = B_1B_2$ where B_1, B_2 have nice structure. While these types of decomposition may or may not be applicable in a numerical procedure, their ideas are fundamental and can be useful in both theoretical and pratical contexts.

We first mention some basic facts in solving a linear system: The systems $Lx = b$ or $Ux = b$ where L is lower triangular and U is upper triangular are solved by backward or forward substitution. They take $O(n^2)$ operations. The system $Dx = b$ where D is diagonal is trivial. It takes exactly n operations to solve. The system $Qx = b$ where Q is an orthogonal matrix has solution $x = Q^Tb$ as $Q^{-1} = Q^T$. In general, one avoids finding the inverse of a matrix directly when solving a system.

2.4.1 LU decomposition

The LU decomposition of a square matrix A has the form $PA = LU$ where P is a permutation matrix, L is lower triangular and U is upper triangular. This is referred to as LU decomposition with partial pivoting. A permutation matrix is a matrix obtained by permuting the rows of the identity matrix. The idea of LU decomposition is that if we want to solve $Ax = b$ this is equivalent to solving $LUx = Pb$. We first solve for $Ly = Pb$ and then $Ux = y$ each of which takes $O(n^2)$ operation. One can also consider LU decomposition without pivoting. That is the decomposition of the form $A = LU$. This decomposition may not always be possible. For example, consider A invertible where $A_{11} = 0$. If $A = LU$ then $A_{11} = L_{11}U_{11}$. This means either $L_{11} = 0$ or $U_{11} = 0$. That is either L is not invertible or U is not invertible. But that contradicts the assumption that A is invertible.

The idea behind LU decomposition is Gaussian elimination. The row echelon form of a Gaussian elimination is an upper triangular matrix. At step i in Gaussian elimination the goal is to zero out all the entries below $A_{ii}^{(i)}$, where $A^{(i)}$ denotes the resulting matrix at the ith step. If $A_{ii}^{(i)} \neq 0$ this can be accomplished by left multiplying $A^{(i)}$ with a lower triangular matrix L_i. If $Aii = 0$ we can swap row i with another row j, $j > i$. This is why the general form of LU decomposition is $PA = LU$. In the case of no pivoting required, the Gaussian elimination steps can be described as

$$L_n L_{n-1} \cdots L_2 L_1 A = U.$$

Since the product of lower triangular matrices is lower triangular, this can be written as $\tilde{L}A = U$. Thus $A = LU$ where $L = (\tilde{L})^{-1}$.

The LU factorization is the cheapest factorization algorithm. Its operations count can be verified to be O($\frac{2}{3}n^3$). In general, one would need to do partial pivoting to make sure LU factorization is stable.

Uniqeness of LU factorization: It is easy to see that LU factorization is not unique. The additional assumption that the diagonal entries of L are 1 is conventionally applied to have uniqueness of LU decomposition.

Algorithms for the factorization of an $n \times n$ matrix A :

LU Factorization without pivoting by hand

1. Reduce A to echelon form U without row permutation, if possible.

2. Place entries in L such that the same sequence of row operations reduces L to I.

Step 1 and 2 are done in tandem. That is we reduce the first column of A and place entries in the first column of L and move to the second columns of A and L. For example, if the first column of A is $\begin{bmatrix} 2 \\ -4 \\ 2 \\ -6 \end{bmatrix}$ then the first column of L is $\begin{bmatrix} 1 \\ -2 \\ 1 \\ 3 \end{bmatrix}$. After this reduction, if the second column of A becomes $\begin{bmatrix} 4 \\ 3 \\ -9 \\ 12 \end{bmatrix}$ (which should not be the original second column of A) then the second column of L is $\begin{bmatrix} 0 \\ 1 \\ -3 \\ 4 \end{bmatrix}$.

Pseudo-code Algorithm (LU Factorization)

```
Initialize U = A, L = I
    for k = 1 : n - 1
        for j = k + 1 : n
        L(j, k) = U(j, k)/U(k, k)
```

U(j, k : n) = U(j, k : n) - L(j, k)U(k, k : n)
end
end

Pseudo-code Algorithm (LU Factorization with Partial Pivoting)
Initialize U = A, L = I, P = I
for k = 1 : n - 1
find $i \geq k$ to maximize $|U(i, k)|$
Swap $U(k, k : n)$ with $U(i, k : n)$
Swap $L(k, 1 : k - 1)$ with $L(i, 1 : k - 1)$
Swap $P(k, :)$ with $P(i, :)$
for j = k + 1 : n
L(j, k) = U(j, k)/U(k, k)
U(j, k : n) = U(j, k : n) - L(j, k)U(k, k : n)
end
end

The operations count for this algorithm is also O($\frac{2}{3}n^2$). In practice one would usually not physically swap rows. Instead one would use pointers to the swapped rows and store the permutation operations instead.

2.4.2 QR decomposition

Given a matrix $A_{m\times n}$, the QR decomposition of A is $A = QR$ where $Q_{m\times n}$ satisfies $Q^TQ = I$ (Q is orthogonal if A is square) and $R_{n\times n}$ is upper triangular (this is referred to as the reduced QR decomposition). Any matrix A has a QR decomposition. If A has independent columns then R is invertible. QR decomposition is often used to solve the linear least squares problem and is the basis for a particular eigenvalue algorithm, the QR algorithm.

QR decomposition is connected with the Gram-Schmidt procedure of transforming a set of vectors into an orthonormal set of vectors with the same span. Indeed, given $A_{n\times n}$ and suppose that A is full rank we can find an orthonormal set of of n vectors whose span is the same as the span of the columns of A. These vectors form the columns of Q. From the Gram-Schmidt procedure, something stronger is true: the span of the first k columns of A is equal to the span of the first k columns of Q. Thus if we solve the matrix equation $A = QR$, we see that R must be upper triangular. Note that the size of the orthonormal set in Gramm-Schmidt procedure may be less than n if A is not invertible. In fact, if A has k linearly independent columns, then the first k columns of Q form an orthonormal basis for the column space of A. More generally, the first k columns of Q form an orthonormal basis for the span of the first k columns of A for any $1 \leq k \leq n$.The fact that any column k of A only depends on the first k columns of Q is responsible for the triangular form of R.

QR algorithm to find eigenvalue: Let A be a square matrix with real entries which we want to compute the eigenvalues of, and let $A_0 := A$. At the k-th step (starting with k = 0), we compute the QR decomposition $A_k = Q_k R_k$. We then form $A_{k+1} = R_k Q_k$. Note that

$$A_{k+1} = R_k Q_k = Q_k^{-1} Q_k R_k Q_k = Q_k^{-1} A_k Q_k = Q_k^{\mathsf{T}} A_k Q_k.$$

so all the A_k are similar and hence they have the same eigenvalues. Under certain conditions, the matrices A_k converge to a triangular matrix, the Schur form of A. The eigenvalues of a triangular matrix are listed on the diagonal, and the eigenvalue problem is solved.

Pseudo inverse in the least squared problem: Recall that the solution to the least squared problem $Ax = b$ is $\hat{x}$ that satisfies

$$A^T A\hat{x} = A^T b.$$

If the columns of A are linearly independent, $A^T A$ is invertible and $\hat{x} = (A^T A)^{-1} A^T b$. The matrix $A^\dagger := (A^T A)^{-1} A^T$ is referred to as the pseudo inverse of A. The psedo-inverse can be expressed as

$$\begin{aligned} A^\dagger &= ((QR)^T (QR))^{-1} (QR)^T \\ &= (R^T Q^T QR)^{-1} R^T Q^T \\ &= (R^T R)^{-1} R^T Q^T (Q^T Q = I) \\ &= R^{-1} (R^T)^{-1} R^T Q^T (R \text{ is nonsingular }) \\ &= R^{-1} Q^T. \end{aligned}$$

Thus $\hat{x} = R^{-1} Q^T b$ is the solution to the least square problem $Ax = b$. One can easily verify that if A is invertible then $A^{-1} = A^\dagger$.

QR decomposition algorithm: The stable numerical procedure for QR decomposition is the Householder algorithm, which utilizes the Householder reflection matrix. Here we present the algorithm based on Gram-Schmidt decomposition (which is more intuitive but not stable in certain cases)and under the assumption that A has independent columns. In the following, $\mathbf{a}_i$ denotes the ith column of A.

Define the projection:

$$\text{proj}_{\mathbf{u}} \mathbf{a} = \frac{\langle \mathbf{u}, \mathbf{a} \rangle}{\langle \mathbf{u}, \mathbf{u} \rangle} \mathbf{u}$$

then:

$$
\begin{aligned}
\mathbf{u}_1 &= \mathbf{a}_1, & \mathbf{e}_1 &= \frac{\mathbf{u}_1}{\|\mathbf{u}_1\|} \\
\mathbf{u}_2 &= \mathbf{a}_2 - \operatorname{proj}_{\mathbf{u}_1} \mathbf{a}_2, & \mathbf{e}_2 &= \frac{\mathbf{u}_2}{\|\mathbf{u}_2\|} \\
\mathbf{u}_3 &= \mathbf{a}_3 - \operatorname{proj}_{\mathbf{u}_1} \mathbf{a}_3 - \operatorname{proj}_{\mathbf{u}_2} \mathbf{a}_3, & \mathbf{e}_3 &= \frac{\mathbf{u}_3}{\|\mathbf{u}_3\|} \\
&\vdots & &\vdots \\
\mathbf{u}_k &= \mathbf{a}_k - \sum_{j=1}^{k-1} \operatorname{proj}_{\mathbf{u}_j} \mathbf{a}_k, & \mathbf{e}_k &= \frac{\mathbf{u}_k}{\|\mathbf{u}_k\|}
\end{aligned}
$$

We can now express the $\mathbf{a}_i$s over our newly computed orthonormal basis:

$$
\begin{aligned}
\mathbf{a}_1 &= \langle \mathbf{e}_1, \mathbf{a}_1 \rangle \mathbf{e}_1 \\
\mathbf{a}_2 &= \langle \mathbf{e}_1, \mathbf{a}_2 \rangle \mathbf{e}_1 + \langle \mathbf{e}_2, \mathbf{a}_2 \rangle \mathbf{e}_2 \\
\mathbf{a}_3 &= \langle \mathbf{e}_1, \mathbf{a}_3 \rangle \mathbf{e}_1 + \langle \mathbf{e}_2, \mathbf{a}_3 \rangle \mathbf{e}_2 + \langle \mathbf{e}_3, \mathbf{a}_3 \rangle \mathbf{e}_3 \\
&\vdots \\
\mathbf{a}_k &= \sum_{j=1}^{k} \langle \mathbf{e}_j, \mathbf{a}_k \rangle \mathbf{e}_j
\end{aligned}
$$

where

$$
\langle \mathbf{e}_i, \mathbf{a}_i \rangle = \|\mathbf{u}_i\|.
$$

This can be written in matrix form:

$$
A = QR
$$

where:

$$
Q = [\mathbf{e}_1, \cdots, \mathbf{e}_n]
$$

and

$$
R = \begin{pmatrix}
\langle \mathbf{e}_1, \mathbf{a}_1 \rangle & \langle \mathbf{e}_1, \mathbf{a}_2 \rangle & \langle \mathbf{e}_1, \mathbf{a}_3 \rangle & \dots \\
0 & \langle \mathbf{e}_2, \mathbf{a}_2 \rangle & \langle \mathbf{e}_2, \mathbf{a}_3 \rangle & \dots \\
0 & 0 & \langle \mathbf{e}_3, \mathbf{a}_3 \rangle & \dots \\
\vdots & \vdots & \vdots & \ddots
\end{pmatrix}.
$$

2.4.3 Cholesky decomposition

The Cholesky decomposition of a symmetric positive definite matrix A is $A = LL^T$ where L is a lower triangular matrix with positive diagonal entries (which implies that L is non-singular). Some author defines the Cholesky decomposition as $A = U^T U$ where U is upper triangular. The two definitions are obviously equivalent.

First note that simply because we can write $A = LL^T$, A must be symmetric. Moreover, for any vector x

$$x^T Ax = x^T LL^T x = \|L^T x\|^2 > 0,$$

unless $x = 0$ or L is singular. Thus A must be positive definite. Thus A has a Cholesky decomposition if and only if A is symmetric positive definite. The Cholesky decomposition can be seen as a stronger version of the LU decomposition (with stronger conditions on A). Cholesky decomposition typically has half the cost of LU decomposition (due to the symmetry). The Cholesky decomposition is unique as a consequence of the requirement that L has positive diagonal entries (otherwise if $A = LL^T$ then it also follows that $A = (-L)(-L)^T$) .

The Cholesky decomposition obviously has applications where LU decomposition is appropriate, especially when the matrix involved is positive definte. For example again the least square solution to $Ax = b$ is $A^T A\hat{x} = A^T b$. If A has linearly independent column then $A^T A$ is symmetric positive definite, so Cholesky decomposition is appropriate here.

Another application of the Cholesky decomposition is in generating multivariate Normal distribution with a desired covariance matrix Σ. It can be shown that if X is a vector of random variables with covariance matrix Σ then for any constant matrix A, the covariance matrix of Ax is $A\Sigma A^T$. Monte Carlo simulation begins with generating a vector X of independent standard Normal distributions (thus the covariance matrix of X is the identity matrix). If our target covariance matrix is Σ then we need to find a matrix A such that $AA^T = \Sigma$. Thus A can be chosen to be L where L is from the Cholesky decomposition of Σ. The random vector $Y = LX$ has the desired distribution since the Normal distribution is closed under affine transformation. For example, since

$$\begin{bmatrix} 1 & 0 \\ \rho & \sqrt{1-\rho^2} \end{bmatrix} \begin{bmatrix} 1 & \rho \\ 0 & \sqrt{1-\rho^2} \end{bmatrix} = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix},$$

to generate two Normal random variables with correlation ρ and unit variances we can generate two independent random Normal variables X_1, X_2 and let $Y_1 = X_1, Y_2 = \rho X_1 + \sqrt{1-\rho^2} X_2$. (Y_1, Y_2) has unit variances and correlation ρ as desired.

Algorithm for Cholesky Decomposition

Input: an $n \times n$ SPD matrix A

Output: the Cholesky factor, a lower triangular matrix L such that $A = LL^T$

Theorem:(proof omitted) For a symmetric matrix A, the Cholesky algorithm will succeed with non-zero diagonal entries in L if and only if A is SPD. If A is not SPD then the

algorithm will either have a zero entry in the diagonal of some L_k (making L_k singular) or will require complex numbers in step 0 or step 1.2.

Notation: L_{k-1}: the $k-1 \times k-1$ upper left corner of L

a_k: the first $k-1$ entries in column k of A

l_k: the first $k-1$ entries in column k of L^T (or equivalently l_k^T is the first $k-1$ entries in row k of L)

a_{kk} and l_{kk}: the kk entries of A and L, respectively.

0) Initialize $L_1 = \sqrt{a_{11}}$.

1) For $k = 2; \cdots ; n$

1.1) Solve $L_{k-1}l_k = a_k$ for l_k (L_{k-1} is $k-1 \times k-1$: for $k = 2$ this is a scalar equation)

1.2) $l_{kk} = \sqrt{a_{kk} - (l^k)^T l_k}$.

1.3)

$$L_k = \begin{bmatrix} L_{k-1} & 0 \\ l_k^T & l_{kk} \end{bmatrix}$$

Example : For

$$A = \begin{bmatrix} 16 & 4 & 4 & -4 \\ 4 & 10 & 4 & 2 \\ 4 & 4 & 6 & -2 \\ -4 & 2 & -2 & 4 \end{bmatrix}$$

construct a Cholesky decompostion of A.

Solution:

$k = 1 : L_1 = \sqrt{16} = 4$

$k = 2 :$ $L_1 = 4; a_2 = 4; a_{22} = 10$. Solve the 1×1 system $L_1 l_2 = a_2$ or $4l_2 = 4$ so $l_2 = 1. l_{22} = \sqrt{10-1} = 3$. Therefore

$$L_2 = \begin{bmatrix} 4 & 0 \\ 1 & 3 \end{bmatrix}$$

$k = 3 :$ $L_2 = \begin{bmatrix} 4 & 0 \\ 1 & 3 \end{bmatrix}, a_3 = \begin{bmatrix} 4 \\ 4 \end{bmatrix}; a_{33} = 6$. Solve the 2×2 system $L_2 l_3 = a_3$ so $l_3 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$. $l_{33} = \sqrt{6 - \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}} = 2$. Therefore

$$L_{33} = \begin{bmatrix} 4 & 0 & 0 \\ 1 & 3 & 0 \\ 1 & 1 & 2 \end{bmatrix}.$$

$k = 4$: $L_3 = \begin{bmatrix} 4 & 0 & 0 \\ 1 & 3 & 0 \\ 1 & 1 & 2 \end{bmatrix}, a_4 = \begin{bmatrix} -4 \\ 2 \\ -2 \end{bmatrix}; a_{44} = 4$. Solve the 3×3 system $L_3 l_4 = a_4$

so $l_4 = \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}$. $l_{44} = \sqrt{4 - [-1 \quad 1 \quad -1] \begin{bmatrix} -1 \\ 1 \\ -1 \end{bmatrix}} = 1$. Therefore

$$L = L_{44} = \begin{bmatrix} 4 & 0 & 0 & 0 \\ 1 & 3 & 0 & 0 \\ 1 & 1 & 2 & 0 \\ -1 & 1 & -1 & 1 \end{bmatrix}.$$

2.4.4 Problems

1. Find the LU factorization of a)

$$A = \begin{bmatrix} 3 & -7 & -2 \\ -3 & 5 & 1 \\ 6 & -4 & 0 \end{bmatrix}.$$

b)

$$A = \begin{bmatrix} 4 & 3 & -5 \\ -4 & -5 & 7 \\ 8 & 6 & -8 \end{bmatrix}.$$

c)*

$$A = \begin{bmatrix} 2 & -1 & 2 \\ -6 & 0 & -2 \\ 8 & -1 & 5 \end{bmatrix}.$$

2. When A is invertible, Matlab find A^{-1} by factoring $A = LU$ and then compute $U^{-1}L^{-1}$. Use this method to compute A^{-1} of the previous problem (where applicable).

3.

a) Find the QR factorization of

$$A = \begin{bmatrix} -1 & -1 & 1 \\ 1 & 3 & 3 \\ -1 & -1 & 5 \\ 1 & 3 & 7 \end{bmatrix}.$$

b) Find the pseddo inverse of A.

c) Find the least square solution to the problem $Ax = b$ where $b = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}$.

4. Find the eigenvalues of A using the QR algorithm where

a)

$$A = \begin{bmatrix} 7 & 4 \\ -3 & -1 \end{bmatrix}.$$

b)

$$A = \begin{bmatrix} 4 & 0 & 1 \\ -2 & 1 & 0 \\ -2 & 0 & 1 \end{bmatrix}.$$

5. Let

$$A = \begin{bmatrix} 9 & -3 & 6 & -3 \\ -3 & 5 & -4 & 7 \\ 6 & -4 & 21 & 3 \\ -3 & 7 & 3 & 15 \end{bmatrix}.$$

Find the Cholesky decomposition of A.

Chapter 3

Differential equations review

3.1 First order ODE

3.1.1 Theory

Linear equations

Consider the equation

$$\begin{aligned} y' + g(t)y &= f(t) \\ y(0) &= y_0. \end{aligned}$$

Under certain conditions, this equation has a unique solution. The solution also has an explicit formula, derived as followed:

$$e^{\int g(t)dt} y' + e^{\int g(t)dt} g(t) y = e^{\int g(t)dt} f(t).$$

That is

$$\frac{d}{dt}\left(e^{\int g(t)dt} y\right) = e^{\int g(t)dt} f(t).$$

Hence

$$y(t) = y_0 + e^{-\int_0^t g(u)du} \int_0^t e^{\int_0^u g(s)ds} f(u)du.$$

Remark: If the iinital condition is $y(t_0) = y_0$ for some $t_0 \neq 0$ the above approach can easily be adapted to fit the new initial condition.

Separable equations

Consider the equation

$$M(x) + N(y)\frac{dy}{dx} = 0.$$

It can be rewriten as

$$M(x)dx = -N(y)dy.$$

Integrating both sides lead to an equation connecting y and x. Note that this does not directly give y as a function of x.

General form of first order equation

First order ODE generally has the form

$$\frac{dy}{dt} = f(t, y).$$

There is no general method for finding explicit solution for general first order ODE (a simple example is the equation $y' = e^{-x^2}$). In these cases, we use numerical methods to find approximation for the solution.

3.1.2 Problems

1. Solve a) $y' - y = 2te^{2t}, y(0) = 1$.
   b) $y' + 2y = 2te^{-2t}, y(1) = 0$.
   c) $ty' + 2y = t^2 - t + 1, y(1) = \frac{1}{2}, t > 0$.
   d)* $y' - 2y = e^{2t}, y(0) = 2$.
   2. Solve
   a) $y' + y^2 \sin x = 0$
   b) $xy' = \sqrt{1 - y^2}$.
   c) $y' = \frac{x - e^{-x}}{y + e^y}$.
   d)* $y' = \frac{x^2}{1+y^2}$.

3.2 Second order ODE

3.2.1 Theory

Here we only consider second order ODE with constant coefficients. Consider the IVP:

$$\begin{aligned} ay'' + by' + cy &= f(t) \\ y(0) &= c_1, y'(0) = c_2 \end{aligned}$$

where a, b, c, c_1, c_2 are constants. We first focus on finding the general solution $y(t)$ to

$$ay'' + by' + cy = f(t). \tag{3.1}$$

The solution to the IVP is found by plugging the initial conditions into $y(t)$. The general solution $y(t)$ has the form

$$y(t) = y_h(t) + y_p(t)$$

where $y_h(t)$ satisfies the homogenous equation

$$ay'' + by' + cy = 0$$

($y_h(t)$ has two undetermined coefficients) and $y_p(t)$ is a particular solution to (3.1). $y_h(t)$ has the form

$$y_h(t) = C_1 e^{\lambda_1 t} + C_2 e^{\lambda_2 t}$$

where λ_1, λ_2 are roots of the quadratic equation

$$ax^2 + bx + c = 0.$$

Note that λ_1, λ_2 can be complex numbers and thus $y_h(t)$ involves the natural exponential, cosine and sine in its form.

The standard approach to finding $y_p(t)$ when $f(t)$ has the form of $P_n(t)e^{ct}$ (where c can be a complex number and $P_n(t)$ is a polynomial of degree n) is to guess $y_p(t) = t^s Q_n(t)e^c(t)$ where $Q_n(t)$ is a polynomial of degree n with undetermined coefficients and s is the smallest integer such that $y_p(t)$ does not correpsond to a homogeneous solution. Not surprisingly, this method is referred to as the method of undertermined coefficients.

In general, consider the linear ODE

$$y'' + p(t)y' + q(t)y = f(t). \tag{3.2}$$

A particular solution for this ODE is

$$y_p(t) = -y_1(t) \int_{t_0}^{t} \frac{y_2(s)f(s)}{W(y_1, y_2)(s)} ds + y_2(t) \int_{t_0}^{t} \frac{y_1(s)f(s)}{W(y_1, y_2)(s)} ds,$$

where t_0 y_1, y_2 are two fundamental homogenous solutions and

$$W(y_1, y_2)(t) = y_1 y_2'(t) - y_2 y_1'(t)$$

is the Wronskian of y_1, y_2. This formula is valid in any open interval I over which p, q, f are continuous. Note that the coefficient of y'' in (3.2) is 1 which is slightly different from the form

$$ay'' + by' + c = f(t)$$

that we were considering so far.

The formula comes from the guess of

$$y_p(y) = u_1(t)y_1(t) + u_2(t)y_2(t)$$

and after plugging in to obtain

$$\begin{aligned} u_1'(t) &= -\frac{y_2(t)f(t)}{W(y_1,y_2)(t)} \\ u_2'(t) &= \frac{y_1(t)f(t)}{W(y_1,y_2)(t)}. \end{aligned}$$

This method is due to Lagrange and referred to as variation of parameters.

3.2.2 Problems

1. Solve using undetermined coefficients
   a) $y'' - 2y' - 3y = -3te^{-t}$
   b) $y'' + 4y = 3\sin(2t)$
   c)* $y'' + 2y' + 5y = 4e^{-t}\cos(2t), y(0) = 1, y'(0) = 0.$
   2. Solve using variation of parameters
   a) $y'' + 4y' + 4y = t^{-2}e^{-2t}, t > 0.$
   b) $4y'' + y = 2\sec(\frac{t}{2}), -\pi < t < \pi.$
   c) $y'' - 2y' + y = \frac{e^t}{1+t^2}, t > 0.$
   3. Show that

$$y_p(t) = -y_1(t)\int_{t_0}^{t} \frac{y_2(s)f(s)}{W(y_1,y_2)(s)}ds + y_2(t)\int_{t_0}^{t} \frac{y_1(s)f(s)}{W(y_1,y_2)(s)}ds,$$

is a particular solution to

$$y'' + p(t)y' + q(t)y = g(t).$$

   4. The Euler equation is of the form:

$$x^2y'' + \alpha xy' + \beta y = 0.$$

This is a particular instance of the general form

$$P(x)y'' + Q(x)y' + R(x)y = 0,$$

where $P(x_0) = 0$ and we seek the solution around a neighborhood of x_0.

a) Consider the ansatz $x(t) = t^r$ and derive a quadratic equation $F(r) = 0$ that r has to satisfy.

b) In the case that the equation $F(r) = 0$ has two real roots, write down the general solutions to the Euler equation.

c) In the case that the equation $F(r) = 0$ has one repeated (real) root r_1, $F(r)$ has the form $F(r) = (r - r_1)^2$. Let L be the differential operator in the Euler equation, we have

$$L(x^r) = x^rF(r) = x^r(r - r_1)^2.$$

r_1 is the repeated root of the quadratic $F(r) = 0$ if and only if $F(r_1) = F'(r_1) = 0$. We can use this as a suggestion to differentiate the RHS of $L(x^r)$ with respect to r and set it equal to 0. This action can be captured as

$$\frac{\partial}{\partial r} L(x^r) = \frac{\partial}{\partial r}[x^r (r - r_1)^2] = 0$$

On the other hand

$$\frac{\partial}{\partial r} L(x^r) = L(\frac{\partial}{\partial r}[x^r]) = L(x^r \ln x).$$

Thus if r_2 is the solution to $\frac{\partial}{\partial r}[x^r (r - r_1)^2] = 0$ then $x^{r_2} \ln x$ is the other solution to the Euler equation. Find out what this value r_2 is.

d) When $F(r) = 0$ has two complex roots we can use the relation $x^r = e^{r \ln x}$ to write down the solution using real coefficients. Suppose $r = \lambda \pm i\mu$. Write down the general solution to the Euler equation.

e) Solve the following Euler equations:

$$\begin{aligned} x^2 y'' + 4xy' + 2y &= 0 \\ x^2 y'' - 3xy' + 4y &= 0 \\ x^2 y'' - xy' + y &= 0 * . \end{aligned}$$

5. Method of convolution:

a)* Consider the IVP

$$y'' + y = f(t), y(0) = y'(0) = 0.$$

Show that the solution is $y(t) = \int_0^t \sin(t - s) f(s) ds$.

b) Consider the general second order IVP with constant coefficients:

$$L[y] = (D^2 + bD + c)y = f(t), y(0) = 0, y'(0) = 0.$$

Here $D := \frac{\partial}{\partial t}$ is the first order differential operator and L represents the differential operator associated with the second order ODE. Assume the solution has the form

$$y(t) = \int_0^t K(t - s) f(s) ds.$$

Find the equation that K has to satisfy. Conclude that K can be constructed from the solutions to the homogeneous ODE $L[y] = 0$. What initial conditions does K have to satisfy?

3.3 Linear system of first order ODEs

3.3.1 Theory

We consider the first order system

$$\mathbf{x}' = A\mathbf{x} + \varphi(t),$$

where $\mathbf{x}(t) := (x_1(t), x_2(t), \cdots, c_n(t))$ and similarly for $\varphi(t)$ are two vectors in $\mathbb{R}^n$.A is a $n \times n$ matrix of constant entries. We note that any n-th order linear ODEs with constant coefficients where $y(t)$ is the unknown can be converted to a first order system by the change of variables $x_i(t) = y^{(i-1)}(t), i = 1, \cdots, n$.

We first address the homogeneous system

$$\mathbf{x}' = A\mathbf{x},$$

and then return to the solution of the non-homogenous system $\mathbf{x}' = A\mathbf{x} + \varphi(t)$ later.

Homogeneous system

Consider the homogeneous sytem

$$\begin{aligned} \mathbf{x}' &= A\mathbf{x} \\ \mathbf{x}(0) &= \mathbf{c}. \end{aligned}$$

The solution for this system can be written generally as

$$\begin{aligned} \mathbf{x}(t) &= e^{At}\mathbf{c} \\ e^{At} &= \sum_{n=0}^{\infty} \frac{(At)^n}{n!}. \end{aligned}$$

It is possible to show that e^{At} is well defined for any square matrix A. One possible drawback of this solution is e^{At} may not be easily computable (via the definition). One special case is when A is diagonalizable (recall that this is equivalent to A having n independent eigenvectors). In this case if $A = PDP^{-1}$

$$e^{At} = Pe^{Dt}P^{-1},$$

where e^{Dt} is the matrix with entries $e^{D_{ii}t}$ on the diagonal. Furthermore, $\mathbf{x}' = A\mathbf{x}$ is

$$PDe^{Dt}P^{-1}\mathbf{c} = APe^{Dt}P^{-1}\mathbf{c}.$$

Thus if we call $\mathbf{c}_0 = P^{-1}\mathbf{c}$ we see that $\mathbf{y} = Pe^{Dt}\mathbf{c}_0$ also satisfies

$$\mathbf{y}'(t) = A\mathbf{y}(t)$$

(but with initial condition $\mathbf{y}(0) = P\mathbf{c}_0$). This is how the solution to the homogenous system is usually presented : as $\mathbf{v}e^{\lambda t}$ where λ is an eigenvalue of A and $\mathbf{v}$ the corresponding eigenvector.

In the case that A is not diagonalizable, we discussed that $A = PJP^{-1}$ where J is the Jordan form of A and P is the matrix formed by the chains of generalized eigenvectors of A. If we denote $\mathbf{y} = P^{-1}\mathbf{x}$ then $\mathbf{y}$ satisfies

$$\mathbf{y}' = J\mathbf{y}.$$

Since J is almost diagonal this system is much easier to solve than the original system. In particular, the individual equations are

$$y_i'(t) = \lambda_i y_i(t) + \varepsilon_i y_{i+1}(t)$$

where λ_i is an eigenvalue and ε_i is either 0 or 1. On the other hand, it always holds that

$$y_n'(t) = \lambda_n y_n(t).$$

Thus we can solve for $y_n(t) = e^{\lambda_n t}$ and back solve for the other values of $y_i(t)$. The solution then can be obtained from $\mathbf{x} = P\mathbf{y}$.

3.3.2 Nonhomogenous system

First consider the system

$$\mathbf{x}' = A\mathbf{x} + \varphi(t).$$

From our discussion in the previous section, the matrix A can be decomposed as PDP^{-1} or PJP^{-1}. Either way, letting $\mathbf{y} = P^{-1}\mathbf{x}$ and multiplying both sides of the above system by P^{-1} we have

$$\begin{aligned} \mathbf{y}' &= D\mathbf{y} + P^{-1}\varphi(t) \text{ or} \\ \mathbf{y}' &= J\mathbf{y} + P^{-1}\varphi(t). \end{aligned}$$

These systems are either diagonal so the equations are separated; or almost diagonal in the Jordan form so the last equation can be solved and then back-substitute to the previous ones.

Next we consider the general case where the coefficients of A can depend on t:

$$\mathbf{x}' = A(t)\mathbf{x} + \varphi(t).$$

We introduce the notion of a fundamental matrix $\Psi(t)$, which is a matrix whose columns consist of the independent solutions to the homogenous equation

$$\mathbf{x}' = A(t)\mathbf{x}.$$

That is

$$\Psi'(t) = A(t)\Psi(t).$$

From the general theory of ODE (and partially from our previous discussion), it can be seen that $\Psi(t)$ is an $n \times n$ invertible matrix. We now guess that the solution to the original non-homogenous system has the form

$$\mathbf{x}(t) = \Psi(t)\mathbf{u}(t),$$

and find the equation that $\mathbf{u}(t)$ has to satisfied. This is indeed the method of variation of parameters discussed in the 2nd order ODE section. Plugging in we have

$$\begin{aligned} \Psi'(t)\mathbf{u}(t) + \Psi(t)\mathbf{u}'(t) &= A(t)\Psi(t)\mathbf{u}(t) + \varphi(t) \\ \Psi'(t)\mathbf{u}(t) + \Psi(t)\mathbf{u}'(t) &= \Psi'(t)\mathbf{u}(t) + \varphi(t). \end{aligned}$$

Thus

$$\begin{aligned} \Psi(t)\mathbf{u}'(t) &= \varphi(t) \text{ or} \\ \mathbf{u}'(t) &= \Psi^{-1}(t)\varphi(t). \end{aligned}$$

Thus $\mathbf{u}(t) = \int \Psi^{-1}(t)\varphi(t)dt$ and we have

$$\mathbf{x}(t) = \Psi(t)\Psi^{-1}(0)\mathbf{x}^0 + \Psi(t)\int_0^t \Psi^{-1}(s)\varphi(s)ds$$

as the solution to the IVP

$$\begin{aligned} \mathbf{x}' &= A(t)\mathbf{x} + \varphi(t) \\ \mathbf{x}(0) &= \mathbf{x}^0. \end{aligned}$$

Note how this form is the generalization of the solution to the first order linear ODE (when $n = 1$ this is the form we found in that section).

Finally we have the solution form as a convolution:

$$\mathbf{x}(t) = e^{At}\mathbf{x}^0 + \int_0^t e^{A(t-s)}\varphi(s)ds.$$

3.3.3 Problems

1. Find the solution to $\mathbf{x}' = A\mathbf{x}$ where
   a)

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -8 & -5 & -3 \end{bmatrix}$$

b)*

$$A = \begin{bmatrix} 1 & 0 & 0 \\ 2 & 1 & -2 \\ 3 & 2 & 1 \end{bmatrix}$$

c)

$$A = \begin{bmatrix} 1 & 1 & 1 \\ 2 & 1 & -1 \\ -3 & 2 & 4 \end{bmatrix}.$$

(Here A has an eigenvalue $\lambda = 2$ with multiplicity 3)

2. Find the solution to $\mathbf{x}' = A\mathbf{x} + \varphi(t)$ where

a)

$$A = \begin{bmatrix} 2 & -1 \\ 3 & -2 \end{bmatrix}; \quad \varphi(t) = \begin{bmatrix} e^t \\ t \end{bmatrix}.$$

b)

$$A = \begin{bmatrix} 2 & -5 \\ 1 & -2 \end{bmatrix}; \quad \varphi(t) = \begin{bmatrix} -\cos t \\ \sin t \end{bmatrix}.$$

3. Euler equation : The system $t\mathbf{x}' = A\mathbf{x}$ is the analogy of the second order Euler equation discussed in the previous section.

a) Convert the Euler equation to the system $t\mathbf{x}' = A\mathbf{x}$ by identifying the matrix A.

b) Assume that the form of the solution is $\mathbf{x} = \mathbf{v}t^r$, show that $(A - rI)\mathbf{v} = 0$.

4. Solve the system

a)

$$t\mathbf{x}' = \begin{bmatrix} 2 & -1 \\ 3 & -2 \end{bmatrix}\mathbf{x}.$$

b)

$$t\mathbf{x}' = \begin{bmatrix} 3 & -2 \\ 2 & -2 \end{bmatrix}\mathbf{x}$$

5. Verify that the given vector is the general solution to the corresponding homogeneous sytem and solve the non-homogeneous system :

a)

$$t\mathbf{x}' = \begin{bmatrix} 2 & -1 \\ 3 & -2 \end{bmatrix}\mathbf{x} + \begin{bmatrix} 1 - t^2 \\ 2t \end{bmatrix}$$

$$\mathbf{x} = c_1 \begin{bmatrix} 1 \\ 1 \end{bmatrix} t + c_2 \begin{bmatrix} 1 \\ 3 \end{bmatrix} t^{-1}.$$

b)

$$\begin{aligned} t\mathbf{x}' &= \begin{bmatrix} 3 & -2 \\ 2 & -2 \end{bmatrix} \mathbf{x} + \begin{bmatrix} -2t \\ t^4 - 1 \end{bmatrix} \\ \mathbf{x} &= c_1 \begin{bmatrix} 1 \\ 2 \end{bmatrix} t^{-1} + c_2 \begin{bmatrix} 2 \\ 1 \end{bmatrix} t^2. \end{aligned}$$

3.4 Basic numerical techniques for ODEs

3.4.1 Theory

Consider the 1 dimensional IVP

$$\begin{aligned} y'(t) &= f(t, y) \\ y(0) &= y_0. \end{aligned}$$

We want to find approximate values of $y(t)$ on an interval $[0, T]$. A class of numerical techniques, called the finite difference methods, employ the idea of partitioning the interval $[0, T]$ into $0 = t_0 < t_1 < t_2 < \cdots < t_n = T$ and give the approximate values of $y(t_i), i = 0, \cdots, n$. Note that the partition does not have to be equi-distance, even though the more basic algorithms assume equal step size. When this is the case the step size is denoted by h. By the fundamental theorem of calculus:

$$y(t_{i+1}) = y_{t_i} + \int_{t_i}^{t_{i+1}} f(s, y_s) ds.$$

The basic idea of numerical techniques is to proceed from $i = 0$ where $y(0)$ is known and approximate $\int_{t_i}^{t_{i+1}} f(s, y_s) ds$ to obtain an approximation of $y(t_1)$ and then repeat the procedure. Note that aside from time $t = 0$ at any other time we start from an approximation of the previous point $y(t_i)$ as well as approximate the change $y(t_{i+1}) - y(t_i)$ via the integral. In this way the error accumulates as we progress in time. Thus there are two kinds of approximating errors one can talk about : local truncation error for approximating the integral at any local time t_i and global truncation error as the total accumulated error once we arrive at $t_n = T$. Different ways to approximate $\int_{t_i}^{t_{i+1}} f(s, y_s) ds$ and different ways to decide the partition on $[0, T]$ give rise to different numerical techniques with different precisions.

Forward Euler method: The simplest approach is to approximate

$$\int_{t_i}^{t_{i+1}} f(s, y_s) ds \approx f(t_i, y(t_i))(t_{i+1} - t_i).$$

This simply

$$y(t_{i+1}) = y(t_i) + f(t_i, y(t_i))(t_{i+1} - t_i).$$

Backward Euler method: Euler forward method approximates the integral with the left point. We can also approximate it with the right point leading to

$$y(t_{i+1}) = y(t_i) + f(t_{i+1}, y(t_{i+1}))(t_{i+1} - t_i).$$

It is important do recognize that in this formula, $y(t_i)$ is known and $y(t_{i+1})$ is unknown and to be solved for. Unlike the forward method, here $y(t_{i+1})$ involves solving an equation (non-linear if f is non-linear in y). For this reason, the forward Euler method is also referred to as the explicit method while the backward Euler method is called the implicit method. This naming convention applies in general to any method where the next grid point needs to be solved for rather than given explicitly. Even though this adds to the complexity of the algorithm in the implicit method, it is a rule of thumb that implicit methods are more stable than explicit methods (and for this reason should be paid attention to).

Improved Euler method: The improved Euler method uses both left and right points to approximate the integral. That is

$$y(t_{i+1}) = y(t_i) + \frac{1}{2}\big[f(t_i, y(t_i)) + f(t_{i+1}, y(t_{i+1}))\big](t_{i+1} - t_i).$$

Since $y(t_{i+1})$ appears on both sides of the equation (and thus needs to be solved for) this is also an implicit method.

Runge-Kutta method: The Runge-Kutta method uses a weighted average to approximate the integral $\int_{t_i}^{t_{i+1}} f(s, y_s)ds$. Here for ease of notation we denote $h = t_{i+1} - t_i$. The integral is approximated as

$$\begin{aligned}
\int_{t_i}^{t_{i+1}} f(s, y_s)ds &\approx \frac{h}{6}\big(k_{i1} + 2k_{i2} + 2k_{i3} + k_{i4}\big), \\
k_{i1} &= f(t_i, y(t_i)) \\
k_{i2} &= f(t_i + \frac{h}{2}, y(t_i) + \frac{h}{2}k_{i1}) \\
k_{i3} &= f(t_i + \frac{h}{2}, y(t_i) + \frac{h}{2}k_{i2}) \\
k_{i4} &= f(t_i + h, y(t_i) + hk_{i3}).
\end{aligned}$$

The various values of $k_{ij}, j = 1, \cdots 4$ can be seen as the approximation of the slope $y'(t)$ at either the left, mid or right point of the interval $[t_i, t_{i+1}]$. For example, k_{i1} is the slope at the left point, k_{i2} is an approximation of the slope at the mid point using k_{i1} to approximate the change $y(t_i + \frac{h}{2}) - y(t_i)$. k_{i3} is also an approximation of the slope at the mid point but now using k_{i2} to approximate the change $y(t_i + \frac{h}{2}) - y(t_i)$. Finally k_{i4} is an approximation of the slope at the right point using k_{i3} to approximate the change $y(t_{i+1}) - y(t_i)$. Runge-Kutta is an explicit method.

A remark on the step size: The methods presented above make no assumption on the step size $t_{i+1} - t_i$. The (optimal) choice of step size is a topic in numerical analysis by

itself. Here we only remark that in general to reduce the error, one would want to choose small step size if the slope is large and one can afford to choose bigger step size if the slope is small. As the slope is represented by the absolute value of $f(t, y)$ we can adapt the step size according to whether $f(t, y)$ is large or small in absolute value. This idea in general is referred to as adaptive step method.

Multi-step method: The previous methods only use the information of the previous step ($y(t_i)$) to calculate the value for the next step. The multi-step method uses the values of more than one previous step to calculate the value of the next one. Recall again that from the FTC:

$$y(t_{i+1}) = y_{t_i} + \int_{t_i}^{t_{i+1}} y'(s)ds.$$

This time we do not replace $y'(s)$ directly with $f(s, y(s))$ in the integral. The idea is once we have the values of $y(t_j), j \leq i$ we also have the values of $y'(t_j)$ given by $y'(t_j) = f(t_j, y(t_j))$. Thus we can approximate $y'(s), t_{i-k} \leq s \leq t_i$ by a interpolating polynomial P_k^i of degree k over the most recent k steps. We can then obtain

$$y(t_{i+1}) = y_{t_i} + \int_{t_i}^{t_{i+1}} P_k^i(s)ds.$$

Note that the interpolating polynomial depends on i. That is at each step i we need to update it to reflect the newest information that have been obtained on $y(t_i)$. These collection of methods of approximating $y'(t)$ using interpolating polynomials and then calculate $y(t_{i+1})$ are referred to as the Adams method. The Adams method are explicit methods.

Another class of multi-step method (which one may say employ the opposite idea of the Adams method) is referred to as the backward differentiation method. The idea is at step i we can approximate $y(s), t_{i-k+1} \leq s \leq t_{i+1}$ (note the interval) by an interpolating polynomial P_k of degree k so that $P'(t_{j+1}) = f(t_{j+1}, P(t_{j+1}))$ and $P(t_j) = y(t_j), t_{i-k+1} \leq t_i$. These provide a system of equations to solve for P_k. Since $y(t_{j+1})$ appears implicitly in the equation for $P'(t_{j+1})$ the backward differentiation methods are implicit methods.

Numerical solution for first order system: Consider the system of first order equations

$$\mathbf{x}' = f(t, \mathbf{x}), \mathbf{x}(0) = \mathbf{x}^0.$$

All of the one-step methods we discussed above can be extended straightforwardly to the multi-dimensional version. For example, the Euler forward method is

$$\mathbf{x}(t_{i+1}) \approx \mathbf{x}(t_i) + \varphi(t_i, \mathbf{x}(t_i))(t_{i+1} - t_i).$$

The multi-step method can also be extended, but in a less straightforward way so we will not present them here.

Remark on error estimates: It is important in numerical scheme to know the order of the error (in terms of the step size) we commit by different acts of approximations,

both locally (locally at a specific time t_i) and globally (from time t_0 to $t_n = T$). The global error estimate in general is more difficult than local error estimate. As we mentioned before, it involves both the sum of the local errors plus the accumulated error by using the approximated values of the previous steps into the next step. The global error estimate only gets more complex with the adaptive step method. Here we limit ourselves to discuss the local error estimate of the Euler forward method. Recall that the forward Euler method is

$$\begin{aligned} y(t_{i+1}) &\approx y(t_i) + f(t_i, y(t_i))(t_{i+1} - t_i) \\ &= y(t_i) + y'(t_i)(t_{i+1} - t_i). \end{aligned}$$

This is exactly the first order Taylor series expansion of y. Thus we know the error in this approximation is $y''(\xi)(t_{i+1} - t_i)^2$ for $\xi \in [t_i, t_{i+1}]$. Note that this error estimate relies on $y''(t)$ which we do not have direct information on. Nevertheless, since $y'(t) = f(t, y(t))$ we do know that

$$y''(t) = f_t(t, y(t)) + f_y(t, y(t))y'(t) = f_t(t, y(t)) + f_y(t, y(t))f(t, y(t)).$$

Thus assuming f has bounded partial derivatives and is bounded itself, we can have a bound on $y''(t)$ based on f. This analysis shows that the Euler forward method is of the order $O(h^2)$ where h is the step size.

3.4.2 Problems

1. Find the approximating value for $y(1)$ in the following problems, using all the numerical schemes discussed above. Compare the errors of different methods.

a) $y' = 3 + t - y, y(0) = 1$

b) $y' = 5t - 3\sqrt{y}, y(0) = 2$

c)* $y' = 2t + e^{-ty}, y(0) = 1$

2. Convert the following problem to a first order system and use the one-step numerical schemes to find the approximate value of $x(1)$:

$$x'' + t^2x' + 3x = t, x(0) = 1, x'(0) = 2.$$

3.5 Some PDEs overview

3.5.1 Theory

A partial differetial equation (PDE) is an equation that involes the partial derivatives of a multi-variate function. The variables can be purely spatial (x, y, z) such as the wave equation or both temporal and spatial (t, x) such as the heat equation. There is no universal technique to find explicit solution of a PDE (even a linear one). In fact, explicit solution is rather the exception than the norm when one investigates a PDE. It is also beyond the scope of this review to investigate into the techniques of explicit solutions of PDE. We present

in this section the classification of second order PDEs and an example of using the Fourier series technique to solve for the homogenous heat equation. The heat equation is chosen because of its connection to financial mathematics.

Classification of second order linear PDEs : A second order linear PDE has the form

$$Au_{xx} + Bu_{xy} + Cu_{yy} + Du_x + Eu_y + Fu = G,$$

where A, B, C, D, E, F, G are constants. It is *second order* because the highest partial derivative has second order. It is *linear* because a linear combination of solutions to such an equation is also a solution. We have the following classifications:

a. The equation is **elliptic** if $B^2 - 4AC < 0$

b. The equation is **parabolic** if $B^2 - 4AC = 0$

c. The equation is **hyperbolic** if $B^2 - 4AC > 0$.

Example 3.5.1. *The* **heat equation**

$$u_t - ku_{xx} = 0, k > 0$$

is parabolic. The **wave equation**

$$u_{tt} - \alpha^2 u_{xx} = 0$$

is hyperbolic. The **Laplace equation**

$$u_{xx} + u_{yy} = 0$$

is elliptic.

These classifications are important because the techniques to solve different types of PDEs are very different. Also, different types of PDEs model different physical phenomena. For example, the parabolic PDEs usually model the temperature of an object (thus it is called the **heat equation**) while the hyperbolic PDEs usually model the displacement of an object from its equilibrium (thus it is called the **wave equation**). The Black-Scholes equation is a heat equation.

A particularly useful technique when we look for an explicit solution of a PDE is to make an *ansatz*, that is a guess for the functional form of the solution. The guess should certainly be based on the structure of the equation, for example the wave equation mentioned above

$$u_{xx} + u_{yy} = 0$$

can be re-written as

$$u_{xx} = -u_{yy}.$$

So it is natural to guess that a form of $u(x, y)$ is

$$u(x, y) = a(x)b(y), \tag{3.3}$$

for some function a, b (since then the part of x is unaffacted by differentiation with respect to y and vice versa). Any solution of $u(x, y)$ in the form (3.3) is referred to as a *product solution* of the PDE, and the technique of finding a product solution is called *separation of variables*. Lastly the PDE is said to be *separable* if we can use separation of variables to find a solution for it.

Example: Find the product solution of

$$u_{xx} = 4u_y.$$

Sol:
Let $u(x, y) = a(x)b(y)$. Then the equation becomes

$$a_{xx}(x)b(y) = 4a(x)b_y(y).$$

That is

$$\frac{a_{xx}(x)}{4a(x)} = \frac{b_y(y)}{b(y)}.$$

Since the LHS only depends on x and the RHS only on y, it means that they both equal to a constant $-c$. Solving

$$\frac{b_y}{b} = -c$$

gives $b(y) = Ke^{-cy}$ for some arbitrary constant K. The second order ODE

$$a_{xx} + 4ca = 0$$

has solution

$$`a(x) = c_1 e^{\sqrt{2|c|}x} + c_2 e^{-\sqrt{2|c|}x}.$$

if $c < 0$. If $c > 0$ then it has solution

$$a(x) = c_1 \cos(\sqrt{2c}x) + c_2 \sin(\sqrt{2c}x).$$

Note: For the particular choice $c_1 = c_2 = \frac{1}{2}$ and $c_1 = -c_2 = \frac{1}{2}$ we have $\sinh(\sqrt{2c}x)$ and $\cosh(\sqrt{2c}x)$ as general solution. Thus it is also possible, and indeed common, to express the general solution when $c > 0$ in terms of $\sinh$ and $\cosh$. Lastly if $c = 0$ then it has solution

$$a(x) = c_1 + c_2 x.$$

Heat Equation and Boundary Value Problems (BVPs)

Consider an insulated rod with length L whose temperature at the two ends are always kept at 0:

$$u(t,0) = u(t,L) = 0, t > 0.$$

Suppose that its initial temperature profile is given by a function $f(x)$:

$$u(0,x) = f(x), 0 < x < L.$$

From our derivation of the heat equation above, the temperature $u(t,x)$ of the rod at any time t and position x is described by the solution to the BVP

$$\begin{aligned} u_t &= ku_{xx} && (3.4)\\ u(t,0) &= u(t,L) = 0, t > 0 && (3.5)\\ u(0,x) &= f(x), 0 < x < L. && (3.6) \end{aligned}$$

The general solution

We will now solve this BVP. First we look for the solution to the equation

$$u_t = ku_{xx}$$

without worrying about the boundary and initial conditions. This equation is separable, that is we make the ansatz

$$u(t,x) = A(t)B(x).$$

Note that we have considered this equation in the example of the last section. Plugging in to the equation we have

$$\frac{A_t}{kA} = \frac{B_{xx}}{B} = -\lambda,$$

for some constant λ. Thus A, B satisfy respectively the DEs

$$\begin{aligned} A_t + k\lambda A &= 0\\ B_{xx} + \lambda B &= 0. \end{aligned}$$

The solution for $A(t)$ is

$$A(t) = Ce^{-k\lambda t},$$

for some constant C to be determined. The solution for $B(x)$ depends on the sign of λ and it is

$$\begin{aligned} B(x) &= C_1 x + C_2, \lambda = 0\\ B(x) &= C_1 \cos(\alpha x) + C_2 \sin(\alpha x), \lambda = \alpha^2 > 0\\ B(x) &= C_1 \cosh(\alpha x) + C_2 \sinh(\alpha x), \lambda = -\alpha^2 < 0. \end{aligned}$$

We now consider the boundary condition. $u(t,0) = u(t,L) = 0$ implies

$$A(t)B(0) = A(t)B(L) = 0.$$

Since $A(t)$ is not the zero function, we conclude that $B(0) = B(L) = 0$ and note that $B(x)$ solves the regular Sturm-Liouville problem

$$\begin{aligned} B_{xx} + \alpha^2 B &= 0 \\ B(0) = B(L) &= 0. \end{aligned}$$

The condition $B(0) = B(L) = 0$ implies that $\lambda = \alpha^2 > 0$ and

$$B(x) = C_1 \cos(\alpha x) + C_2 \sin(\alpha x)$$

since the other two possibilities of λ also forces $B(x) = 0$, which again implies $u(t,x) = 0$.

The condition $B(0) = 0$ implies $C_1 = 0$. The condition $B(L) = 0$ implies

$$\sin(\alpha L) = 0.$$

Thus (recalling that we have the freedom to choose what $\lambda = \alpha^2$ is)

$$\alpha = \frac{n\pi}{L},$$

for some natural number n. For each choice of $\alpha = \frac{n\pi}{L}$, it is conceivable that we have a different corresponding constant C_2^n. That is we have a family of solution

$$B_n(x) = C_n \sin(\frac{n\pi}{L}x).$$

Since the boundary condition is homogeneous, the sum of solutions is a solution and thus the general solution to the DE:

$$\begin{aligned} B_{xx} + \alpha^2 B &= 0 \\ B(0) = B(L) &= 0 \end{aligned}$$

is

$$B(x) = \sum_{n=1}^{\infty} C_n \sin(\frac{n\pi}{L}x).$$

Thus the general solution to the heat equation, considering only the boundary condition is

$$u(t,x) = \sum_{n=1}^{\infty} C_n e^{-k\frac{n^2\pi^2}{L^2}t} \sin(\frac{n\pi}{L}x).$$

(The constant C in the general solution of $A(t)$ is absorbed into the C_n in the above representation).

Remark: As $t \to \infty$, each term of the series converges to 0. Thus we may believe that $u(t,x) \to 0$ as $t \to \infty$ (we need to justify exchanging the limit and the summation to make it rigorous). This agrees with our intuitition that the temperature of the rod converges to 0 everywhere due to the fact that the two ends are kept at constant 0 degree.

The condition $u(0,x) = f(x)$ implies

$$\sum_{n=1}^{\infty} C_n \sin(\frac{n\pi}{L}x) = f(x).$$

That is $C_n's$ are the Fourier coefficients in the half-range expansion of $f(x)$ on the interval $[0, L]$ using the sine series. A standard result from Fourier series gives :

$$C_n = \frac{2}{L}\int_0^L f(x)\sin(\frac{n\pi}{L}x)dx.$$

Thus the solution to the heat equation (3.4) is

$$u(t,x) = \sum_{n=1}^{\infty}\Big(\frac{2}{L}\int_0^L f(x)\sin(\frac{n\pi}{L}x)dx\Big)e^{-k\frac{n^2\pi^2}{L^2}t}\sin(\frac{n\pi}{L}x).$$

An example

Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \\ u(t,0) &= u(t,\pi) = 0, t > 0 \\ u(0,x) &= 1, 0 < x < \pi. \end{aligned}$$

The Fourier coefficient C_n is

$$C_n = \frac{2}{\pi}\int_0^{\pi}\sin(nx)dx = \frac{2}{\pi}\frac{-1-(-1)^n}{n}.$$

That is

$$u(t,x) = \frac{2}{\pi}\sum_{n=1}^{\infty}\frac{-1-(-1)^n}{n}e^{-n^2t}\sin(nx).$$

3.5.2 Problems

1. Find the product solutions of :
   a) The wave equation : $u_{tt} - \alpha^2 u_{xx} = 0$.
   b) The Laplace equation : $u_{xx} + u_{yy} = 0$.
   2. Use the Fourier techniques to solve for

a)

$$\begin{aligned} u_t &= u_{xx} \\ u(t,0) &= u(t,\pi) = 0, t > 0 \\ u(0,x) &= x, 0 < x < \pi. \end{aligned}$$

b)

$$\begin{aligned} u_t &= 4u_{xx} \\ u(t,0) &= u(t,\pi) = 0, t > 0 \\ u(0,x) &= \sin x, 0 < x < \pi. \end{aligned}$$

3.6 Finite difference method for the heat equation

In this section we discuss the finite difference methods for numerical solution of a heat equation. The idea behind finite difference methods is to replace the differential operators in a PDE with their finite difference approximations. Finite difference methods in PDE naturally involve matrix multiplication and / or solution of linear systems. This is where it connects back to the linear algebra theory and numerical linear algebra that we touched upon in the linear algebra review.

3.6.1 Theory

Forward difference method of the heat equation

Consider the heat equation

$$\begin{aligned} u_t &= \alpha^2 u_{xx} \\ u(t,0) &= u(t,L) = 0, t > 0 \\ u(0,x) &= f(x), 0 < x < \pi. \end{aligned}$$

Let $(\Delta t, \Delta x)$ be the step sizes in the (t,x) direction. At a point (t_i, x_j) we have

$$\begin{aligned} u_t(t_i,x_j) &\approx \frac{u(t_i+\Delta t, x_j) - u(t_i,x_j)}{\Delta t} \\ u_{xx}(t_i,x_j) &\approx \frac{u(t_i, x_j+\Delta x) - 2u(t_i,x_j) + u(t_i, x_j - \Delta x)}{\Delta x^2} \end{aligned}$$

These are the forward difference and central difference formulae we discussed before in numerical differentiation. Plugging these approximations into the heat equation we have:

$$\frac{u(t_i+\Delta t, x_j) - u(t_i,x_j)}{\Delta t} = \alpha^2 \frac{u(t_i, x_j+\Delta x) - 2u(t_i,x_j) + u(t_i, x_j - \Delta x)}{\Delta x^2}.$$

Because the initial condition $u(0,x) = f(x)$ is given, we can start at time $t_0 = 0$ and use the above formula to find $u(t_1, x_j), j = 1, \cdots, N-1$ and then repeat the procedure. Specifically :

$$\begin{aligned} u(t_i + \Delta t, x_j) &= u(t_i, x_j) + \frac{\alpha^2 \Delta t}{\Delta x^2}\big(u(t_i, x_j + \Delta x) - 2u(t_i, x_j) + u(t_i, x_j - \Delta x)\big), \\ & 1 \le j \le N-1. \end{aligned}$$

Note that if we define $x_0 = 0$ and $x_N = L$ then the above scheme is only applicable for $1 \le i \le N$. We impose $u(t_i, x_0) = u(t_i, x_N) = 0$ to satisfy the boundary conditions. We can organize the information of $u(t_i, x_j), 1 \le j \le N-1$ into a vector $\mathbf{u}_i$:

$$\mathbf{u}_i = \begin{bmatrix} u(t_i, x_1) \\ u(t_i, x_2) \\ \vdots \\ u(t_i, x_{N-2}) \\ u(t_i, x_{N-1}) \end{bmatrix}.$$

Note that the elements are $u(t_i, x_0), u(t_i, x_N)$ are not incorporated into the vector $\mathbf{u}_i$ because they are forced by the boundary conditions. The forward scheme becomes:

$$\mathbf{u}_{i+1} = A\mathbf{u}_i,$$

where

$$\begin{aligned} A &= \begin{bmatrix} a & b & 0 & \cdots & 0 \\ b & a & b & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & b & a \end{bmatrix}, \\ a &= 1 - 2\lambda, b = \lambda, \lambda = \frac{\alpha^2 \Delta t}{\Delta x^2}. \end{aligned}$$

Thus A is a tridiagonal matrix that we discussed before. Note how A incorporates the boundary conditions with the first and last row :

$$\begin{aligned} u(t_i + \Delta t, x_1) &= \lambda u(t_i, x_0) + (1 - 2\lambda)u(t_i, x_1) + \lambda u(t_i, x_2) \\ &= (1 - 2\lambda)u(t_i, x_1) + \lambda u(t_i, x_2) \end{aligned}$$

since $u(t_i, x_0) = 0$ and

$$\begin{aligned} u(t_i + \Delta t, x_{N-1}) &= \lambda u(t_i, x_{N-2}) + (1 - 2\lambda)u(t_i, x_{N-1}) + \lambda u(t_i, x_N) \\ &= \lambda u(t_i, x_{N-2}) + (1 - 2\lambda)u(t_i, x_{N-1}) \end{aligned}$$

since $u(t_i, x_N) = 0$.

Stability of the forward finite difference scheme: We have

$$\mathbf{u}_n = A^n \mathbf{u}_0,$$

and thus if A^n is not "well-behaved" for large n the scheme is not stable. The matrix A is diagonalizable, and thus we see if it has an eigenvalue with absolute value larger than 1 then A^n possibly has some large entries that can cause issue. In particular, recall problem 10 in section 2 of the Linear Algebra review:

The forward Euler finite difference scheme for the heat PDE corresponds to the tridiagonal matrix $A(i,i) = 1 - 2\lambda, A(i,i+1) = A(i,i-1) = \lambda, \lambda > 0$. The scheme is convergent if and only if

$$\|A\|_2 := \max |\lambda_i| < 1.$$

Show that this condition is satisfied if and only if $0 < \lambda \leq \frac{1}{2}$.

This is equivalent to

$$\frac{\alpha^2 \Delta t}{\Delta x^2} \leq \frac{1}{2}$$

or $\Delta t \leq \frac{\Delta x^2}{2\alpha^2}$. Suppose we choose $\Delta x = 0.01$ and $\alpha = 1$. The stability condition implies $\Delta t = 0.005$. This is a very small step size (and much smaller if Δx is even smaller). Thus it may take many iterations for the forward difference scheme to "arrive" at the desired time for the solution.

Backward finite difference method for the heat equation

Just as in the ODE case we can approximate u_t via a backward difference formula rather than a forward one:

$$u_t(t_i, x_j) \approx \frac{u(t_i, x_j) - u(t_i - \Delta t, x_j)}{\Delta t}.$$

The central difference formula for u_{xx} is the same. Plugging into the heat equation :

$$\frac{u(t_i, x_j) - u(t_i - \Delta t, x_j)}{\Delta t} = \alpha^2 \frac{u(t_i, x_j + \Delta x) - 2u(t_i, x_j) + u(t_i, x_j - \Delta x)}{\Delta x^2}.$$

That is :

$$\begin{aligned} u(t_i - \Delta t, x_j) &= u(t_i, x_j) + \frac{\alpha^2 \Delta t}{\Delta x^2}\big(- u(t_i, x_j + \Delta x) + 2u(t_i, x_j) - u(t_i, x_j - \Delta x)\big), \\ & \quad 1 \leq j \leq N-1. \end{aligned}$$

Or

$$\begin{aligned} u(t_i, x_j) &= u(t_i + \Delta t, x_j) + \frac{\alpha^2 \Delta t}{\Delta x^2}\big(- u(t_i + \Delta t, x_j + \Delta x) + 2u(t_i + \Delta t, x_j) \\ &- u(t_i + \Delta t, x_j - \Delta x)\big), 1 \leq j \leq N-1. \end{aligned}$$

Here $u(t_i, x_j)$ is **known** and $u(t_i + \Delta t, x_j)$ is **to be solved for.** In terms of the matrix representation we have:

$$\mathbf{u}_i = A\mathbf{u}_{i+1},$$

where

$$\begin{aligned} A &= \begin{bmatrix} a & b & 0 & \cdots & 0 \\ b & a & b & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & b & a \end{bmatrix}, \\ a &= 1 + 2\lambda, b = -\lambda, \lambda = \frac{\alpha^2 \Delta t}{\Delta x^2}. \end{aligned}$$

At each step we need to solve for $\mathbf{u}_{i+1}$ using the information of $\mathbf{u}_i$, starting with $\mathbf{u}_0$ as given by $f(x)$. Note also how A incorporates the boundary conditions with the first and last row :

$$\begin{aligned} u(t_i, x_1) &= -\lambda u(t_i + \Delta t, x_0) + (1 + 2\lambda)u(t_i + \Delta t, x_1) - \lambda u(t_i + \Delta t, x_2) \\ &= (1 + 2\lambda)u(t_i + \Delta t, x_1) - \lambda u(t_i + \Delta t, x_2) \end{aligned}$$

since $u(t_i, x_0) = 0$ and

$$\begin{aligned} u(t_i, x_{N-1}) &= -\lambda u(t_i + \Delta t, x_{N-2}) + (1 + 2\lambda)u(t_i + \Delta t, x_{N-1}) - \lambda u(t_i + \Delta t, x_N) \\ &= -\lambda u(t_i + \Delta t, x_{N-2}) + (1 + 2\lambda)u(t_i + \Delta t, x_{N-1}) \end{aligned}$$

since $u(t_i, x_N) = 0$.

Stability: We have

$$\mathbf{u}_n = (A^{-1})^n \mathbf{u}_0,$$

Similar to the discussion above, and recalling Problem 11 in Section 2 of the Linear algebra review: The backward Euler finite difference scheme for the heat PDE corresponds to the tridiagonal matrix $A(i, i) = 1 + 2\lambda, A(i, i + 1) = A(i, i - 1) = -\lambda, \lambda > 0$. The scheme is convergent if and only if

$$\|A^{-1}\|_2 := \max |\lambda_i| < 1,$$

where A^{-1} is the inverse of A.

a) Show that the discretization matrix A is indeed invertible.

b) Show that the convergence condition is satisfied for any $\lambda > 0$.

We see that the backward difference scheme is **always** stable. Thus in this way it is preferable to the forward scheme.

On other boundary conditions

The above forward and backward scheme incorporates the homogeneous Dirichlet boundary condition $u(t,0) = u(t,L) = 0$. One can incorporate nonhomogeneous Dirichlet boundary condition : $u(t,0) = g(t), u(t,L) = h(t)$, as well as Neumann boundary condition: $u_x(t,0) = g(t), u_x(t,L) = h(t)$. Here we discuss the Neumann condition implementation and leave the nonhomogeneous Dirichlet condition as an exercize.

One needs to use central difference approximation for Neumann boundary condition instead of forward or backward difference approximation. That is

$$u_x(t_i, x_i) \approx \frac{u_x(t_i, x_{i+1}) - u_x(t_i, x_{i-1})}{2\Delta x}.$$

The reason is because we used the central difference formula to approximat u_{xx}. This involves a second order Taylor approximation. If we use the first order Taylor approximation for u_x at the boundaries, this will be inconsistent and leads to inaccuracies in the solution (in fact it will lead to the wrong solution as we shall see). Put in another way, if we were to use

$$u_x(t_i, x_i) \approx \frac{u_x(t_i, x_{i+1}) - u_x(t_i, x_i)}{\Delta x},$$

the error term is of the form $\frac{1}{2}u_{xx}(t_i, \xi)\Delta x$. But u_{xx} is a term that appears in our equation and thus should not appear in the error term.

The central difference approximation naturaly involves the points $u(t_i, x_{-1})$ and $u(t_i, x_{n+1})$. We refer to these as ghost points. They are values on the imaginary grid points that help us compute the values of $u(t_{i+1}, x_0)$ and $u(t_{i+1}, x_n)$ at the next time step using the same three-point stencil scheme as elsewere in the grid. Here again we see why the central difference approximation must be used. Inspecting the scheme we see that the scaling factor is λ which is $O(\frac{1}{(\Delta x)^2})$. When λ is multiplied with a central difference approximation it leaves an error of order Δx. But if λ is multiplied with a forward or backward difference approximation it leaves error of a constant order. Since the ghost points are used to compute the values on the boundary in the next iteration, they are multiplied by λ and thus if the error is not of the right order, it will "contaminate" the error involved in the u_{xx} approximation.

We'll leave it for the reader to verify that for the homogeneous Neumann condition $u_x(t,0) = u_x(t,L) = 0$, the forward method is

$$\mathbf{u}_{i+1} = A\mathbf{u}_i,$$

where

$$A = \begin{bmatrix} 1-2\lambda & \mathbf{2}\lambda & 0 & \cdots & 0 \\ \lambda & 1-2\lambda & \lambda & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \mathbf{2}\lambda & 1-2\lambda \end{bmatrix},$$

For the homogeneous Neumann problem, the backward method is

$$\mathbf{u}_i = A\mathbf{u}_{i+1},$$

where

$$A = \begin{bmatrix} 1+2\lambda & -\mathbf{2}\lambda & 0 & \cdots & 0 \\ -\lambda & 1+2\lambda & -\lambda & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & -\mathbf{2}\lambda & 1+2\lambda \end{bmatrix}.$$

One can verify that the first order approximation for the homogeneous boundary condition leads to

$$A = \begin{bmatrix} \mathbf{1}-\lambda & \lambda & 0 & \cdots & 0 \\ \lambda & 1-2\lambda & \lambda & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda & \mathbf{1}-\lambda \end{bmatrix},$$

for the forward method and

$$A = \begin{bmatrix} \mathbf{1}+\lambda & \lambda & 0 & \cdots & 0 \\ \lambda & 1-2\lambda & \lambda & \cdots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \cdots & \cdots & \cdots & \vdots \\ 0 & 0 & \cdots & \lambda & \mathbf{1}+\lambda \end{bmatrix},$$

for the backward method. We can verify that the first order approximation leads to the wrong solution by checking that the numerical solution for the problem

$$\begin{aligned} u_t &= u_{xx} \\ u_x(t,0) &= 0, u_x(t,\pi) = 0, t > 0 \\ u(0,x) &= \sin x, 0 < x < \pi \end{aligned}$$

converges to $u(t,x) = 0$ as $t \to \infty$.

3.6.2 Problems

1. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \\ u(t,0) &= u(t,1) = 0, t > 0 \\ u(0,x) &= \cos(\frac{\pi x}{\Delta x}), 0 < x < \pi. \end{aligned}$$

where $\Delta x = 0.01$. (The initial condition is rather artificial because Δx has to be given apriori. Nevertheless it demonstrates the stability of the forward backward scheme).

a) Use the forward scheme with $\Delta t = 0.01$ to find $u(T, x_j)$ where $T = 2$. What is your observation?

b) Use the backward scheme with $\Delta t = 0.01$ to find $u(T, x_j)$ where $T = 2$. What is your observation?

c) Find an appropriate Δt to implement the forward scheme. What is your observation?

2. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \\ u(t,0) &= u(t,\pi) = 0, t > 0 \\ u(0,x) &= 1, 0 < x < \pi. \end{aligned}$$

We have given a Fourier series solution to this equation above. Use either the forward scheme and backward scheme and compare the numerical solution you found with the Fourier series solution.

3*. Consider the heat equation

$$\begin{aligned} u_t &= u_{xx} \\ u_x(t,0) &= a, u_x(t,L) = b, t > 0 \\ u(0,x) &= f(x), 0 < x < L. \end{aligned}$$

This equation is given with non-homogeneous Dirichlet condition. Discuss how you can adapt the above schemes to this problem. Apply the scheme(s) to solve for the PDE numerically

$$\begin{aligned} u_t &= u_{xx} \\ u(t,0) &= 1, u(t,\pi) = 0, t > 0 \\ u(0,x) &= \sin x, 0 < x < \pi. \end{aligned}$$

Chapter 4

Probability review

4.1 Theory

4.1.1 Probability and Events

Events and their properties

Consider an experiment where we toss a coin twice. All the possible outcomes are

$$\{TT\}, \{HH\}, \{TH\}, \{HT\}.$$

We call these (elementary) events. The events have the following properties:

a. The union of two events is an event:

$$\{TT\} \cup \{TH\} = \{\text{First toss is } T\}.$$

b. The intersubsection of two events is an event:

$$\{\text{First toss is } T\} \cap \{\text{Second toss is } T\} = \{TT\}.$$

c. The complement of an event is an event:

$$\{TT\}^c = \{ \text{ At least one of the toss is } H\}.$$

Note: In everyday language, union corresponds to OR, intersubsection corresponds to AND, complement corresponds to NOT.

Suppose we toss a coin n times. It is not difficult to see that that more generally we have the followings:

a'. The union of finitely many events is an event: The event $\{$First toss is $T\}$ is the union of finitely many events where each of them has the form $\{T \cdots\}$.

b'. The intersubsection of finitely many events is an event: The event $\{$All tosses are $T\}$ is the intersubsection of n events where each of them has the form $\{$ The nth toss is $T\}$.

Suppose we toss a coin indefinitely. Then we have the followings:

a". The union of (countably) infinitely many events is an event: The event $\{$We eventually see a $T\}$ is the (countable) union of events of the form $\{$ The nth toss is $T, n = 1, 2 \cdots\}$.

b". The intersubsection of (countably) infinitely many events is an event: The event $\{$All the even toss is $T\}$ is the (countable) intersubsection of events of the form $\{$ The nth toss is $T, n = 2, 4, 6 \cdots\}$.

Terminology: When two events have nothing in common (their intersubsection is $\emptyset$, the empty set) we say they are *mutually exclusive*. For example, the two events $\{$First toss is $H\}$ and $\{$First toss is $T\}$ are mutally exclusive.

Abstractly, we use capital letters at the beginning of the alphabet: A, B or $E_1, E_2 \cdots$ to denote an event. We also see that in the examples above, an outcome (or an elementary event) is an event that has no sub-event contained in it (in other words, a smallest possible event).

Probability

The union of all possible outcomes is an event, (the *universal* event, also called the *sample space*), which we denote by Ω. Then all events are subsets of Ω. We assign a probability, which is a number between 0 and 1, on each event. The probability then is nothing but a mapping from the set of events to the interval $[0, 1]$. Intuitively, this mapping should satisfy the following property:

a. The probability of the union of all outcomes is 1: $P(\Omega) = 1$.

b. The probability of the empty set is 0: $P(\emptyset) = 0$.

c. The probability of the union of two mutually exclusive events is the sum of the individual probability of each event: If $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

From c, we have the following inclusion - exclusion principle: For any events A, B (not necessarily mutually exclusive)

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

Exercise: Prove the inclusion - exclusion principle.

Using a,b,c one can come up with more probability identity, for example $P(A^c) = 1 - P(A)$ etc.

When assigning probability, besides a,b,c, we also use the following "commonsense" principle: outcomes that are equally likely have the same probability. For example, if a coin is *fair*, then all outcomes $\{TT\}, \{HH\}, \{TH\}, \{HT\}$ are equally likely. Now applying a and c, we see easily that each of them should have probability equals $1/4$.

Examples

Example 4.1.1. *We toss a coin twice. The probability that we get at least 1 tail is*

$$P(\{TT\} \cup \{TH\} \cup \{HT\}) = \frac{3}{4}.$$

The probability that we get no tail is

$$P(\{HH\}) = \frac{1}{4}.$$

Example 4.1.2. *Combinatorics Suppose an urn has 2 white balls and 3 red balls. We pick out (without replacement) 2 balls. What is the probability that the 2 balls are red?*

Ans: Here we need to see what the sample space is. It is all possible ways we can pick out 2 balls from the urn. What is the event of interest? It is all possible ways we can pick 2 red balls form the urn. Since each outcome from our pick is equally likely (by equally likely outcome here we mean suppose we number all the balls from 1 to 5, then the possibility we pick out balls 1,2 is the same as the possibiltiy we pick out balls 4,5), the probability of interest is just the ratio of the size of the event with the size of the sample space.

Concretely, the number of ways we can pick 2 balls out of 5 balls is $\binom{5}{2} = 10$*. The number of ways we can pick 2 red balls is* $\binom{3}{2} = 3$*. So the probability is* $\frac{3}{10}$.

4.1.2 Conditional probability and independent events

Conditional probability

Suppose we toss a coin twice. What is the probability that we get 2 tails? From the above, it's $\frac{1}{4}$. Suppose, however, that you know the additional information that the first toss is a tail. We ask the same question: what is the probability that we get 2 tails? Clearly it's no longer $\frac{1}{4}$, because for you, the set of *all possible events* have changed. Namely, the outcomes $\{HH\}, \{HT\}$ are no longer possible.

Concretely, the set of all possible outcomes now are:

$$\{TT\}, \{TH\}.$$

Thus the probability that you get 2 tails is $\frac{1}{2}$. We say: the probability that we get 2 tails, *conditioned on* the first toss being a tail, is $\frac{1}{2}$.

Definition 4.1.3. *Let* A, B *be events. If* $P(A) > 0$*, the probability of* B *conditioned on* A*, or* B *given* A*, denoted* $P(B|A)$*, is defined as:*

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

The interpretation is that we have already had the knowledge that A happened. So the probability of the event B happening, given that A has happened, should be calculated as given in the definition.

Remark 4.1.4. *If* $P(A) = 0$ *then we cannot use the above formula to define* $P(B|A)$*. There is a way around it, using the measure theoretic definition of conditional expectation, and the notion of regular conditional probability. We'll discuss this later on in Lecture 1b. See also the discussion on conditional density in Lecture 1b.*

Example 4.1.5. *We toss a die. What is the probability that we get a 6, given that we know the toss is even?*

Ans: Let A be the event that we get an even toss, B the event that we get a 6 (when you get used to this, you don't have to explicitly name out the events). Then $P(A) = 1/2$, $P(A \cap B) = P(B) = 1/6$. Thus $P(B|A) = 1/3$.

Bayes' rule

From the definition of conditional probability, we have

$$P(B|A)P(A) = P(B \cap A).$$

It is clear that

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

Therefore, we conclude

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

This formula is called the Baye's rule. At first glance this is pure mathematical manipulation. But it has an important implication: that of switching what we conditioned on. An example would illustrate what this means.

It is well-known that medical test is not 100% reliable. That is suppose you test for a disease, which has 1% chance of happening, then even if the test comes out negative, it doesn't mean you have 0% of contracting the disease. Instead, with a very small probability, it could be a false negative. Concretely, suppose that if you indeed have the disease, then there is 98% chance that the test comes out positive, and 2% negative. However, suppose you don't have the disease, there is 95% chance the test comes out negative, and 5% chance it comes out positive. Now you go for the test, and it comes out negative. What is the probability that you contract the disease?

Ans: Let A be the event that you contract the disease and B be the event that the test is positive. Then we have

$$P(B|A) = .98, P(B^c|A) = .02, P(B|A^c) = .05, P(B^c|A^c) = .95.$$

The question asks for $P(A|B^c)$. Thus you see how Bayes' rule is appropriate for the situation. Can you figure out what it is?

Independent Events

Definition 4.1.6. *Two events A and B, are said to be independent if $P(A|B) = P(A)$ and $P(B|A) = P(B)$.*

Remarks: If $P(A|B) = P(A)$ then $P(B|A) = \frac{P(B\cap A)}{P(A)} = \frac{P(A|B)P(B)}{P(A)} = P(B)$. Thus we actually need one of the two equalities given above for the definition of 2 independent events.

Interpretation: Intuitively, two events are independent if the knowledge of one event already happened does not influence the probability of the other happening, hence the definition.

Alternatively, one can define A and B to be independent if $P(A \cap B) = P(A)P(B)$. You should check that this is equivalent to the condition $P(A|B) = P(A)$ given in the definition. So in fact one have two possible ways to define what it means for 2 events to be independent. The interpretation of the equality $P(A \cap B) = P(A)P(B)$ is not very clear (at least to me) so I prefer to use the other equality for definition of independence.

4.1.3 Random variables

Definition

In an experiment, we have (random) outcomes. We can give them names (for example tossing a coin twice, we can get $HH, TT \cdots$). Each of these have some weight attached to them, i.e. their probability (in the coin toss example, 1/4 for each). However, we cannot do computations with these outcomes unless we give them some numerical values. A *random variable* is a way to *quantify* the random outcomes in a meaningful manner. We use capital letters at the end of the alphabet: X, Y, Z, to denote random variables.

Formally, a random variable (*from now on abbreviated as* RV) X is a mapping from the set of outcomes to the real line ($\mathbb{R}$) such that all sets of the form $\{X \in [a, b]\}$ are events. That is, we can assign probability to these sets.

Example 4.1.7. *Let X be a random variable corresponding to a coin toss. That is $X = 1$ is the coin turns up H and $X = 0$ if the coin turns up T. Then we can see that $P(X = 1) = P(X = 0) = 1/2$.*

Note: There is no reason why 1 has to be assigned to H and 0 assigned to T. One can assign a different value to these outcomes and get a different variable, as suited one's purpose. For example, the RV Y such that $Y = 1$ if the coin is H nd $Y = -1$ if the coin is T is also an example of a RV.

Example 4.1.8. *Let X be a random variable that corresponds to the time one has to wait at the Hill Center's bus stop before one can catch a bus to College Ave. Suppose that the bus arrives every 15 minutes, and they arrive uniformly during any time frame. The we see that $P(a < X < b) = \frac{b-a}{15}$, for $0 \leq a \leq b \leq 15$. Also one should observe that $P(X = a) = 0$ for any $a \in [0, 15]$ (the probability that one waits exactly 7 minutes before the bus arrives is 0).*

Discrete versus continuous RVs

In probability theory, one distinguishes between discrete and continuous RVs (note that these are not the only types of RVs there are. One can have a mixed RV as well). Roughly speaking, a discrete RV takes values on a discrete set (for example, the natural numbers is a discrete set, so is $\{1, 2, 3, 4, 5\}$). Moreover, if X is a discrete RV then $P(X = x) > 0$, where x is in the range of X. Examples of discrete RVs that you may have learned are: the Binomial, the Geometric, the Hypergeometric, the Poisson.

A continuous RV, on the other hand, takes values on an interval (or several intervals). Moreover, if X is a continuous RV then $P(X = x) = 0$, even if x is in the range of X. Examples of continuous RVs that you may have learned are: the Exponential, the Normal, the Uniform, the Gamma, the Cauchy.

Probability distribution, pdf, cdf

Discrete RV:

To characterize a discrete RV, we use the probability distribution function. It gives the formula for the probability that the RV takes some specific value. For example, if X has Bionimial(n,p) distribution, then $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ is the distribution function of X.

Continuous RV : To characterize a continuous RV, we use the probability density function (pdf). The pdf does not give a probability itself, but it is connected to a probability via the following formula:

$$P(X \leq x) = \int_{-\infty}^{x} f_X(u) du,$$

where f_X above is the pdf of the RV X.

Cdf:

Both continuous and discrete RVs can also be described via the cumulative distribution function, which gives the formula for the probability that the RV is less than or equal to some value:

$$F_X(x) = P(X \leq x).$$

Note that if X is a continuous RV, then F_X is differentiable, and its derivative is the density function f_X.

The moments

Discrete RV :

Let X be a discrete RV. Then its first moment, the Expectation, is defined as:

$$E(X) = \sum_n nP(X = n),$$

where the sum is understood to be taken over all values in the range of X.

It can be showed (note: not a definition) that for any function f, the expectation of the RV $f(X)$ is

$$E(f(X)) = \sum_n f(n)P(X = n).$$

In particular, we have the kth moment of X is $E(X^k) = \sum_n n^k P(X = n)$.

Continuous RV :

For a continuous RV X, we define the expectation as:

$$E(X) = \int_{-\infty}^{\infty} x f_X(x) dx.$$

More generally, for any function g, we have

$$E(g(X)) = \int_{-\infty}^{\infty} g(x) f_X(x) dx.$$

Variance, covariance, correlation :

Let X be a RV. We then define its variance as

$$Var(X) = E\Big[(X - E(X))^2\Big] = E(X^2) - E^2(X).$$

The variance measures how “spread out" the RV is from its mean.

Let X, Y be RVs. We define their covariance as

$$Cov(X, Y) = E\Big[(X - E(X))(Y - E(Y))\Big] = E(XY) - E(X)E(Y).$$

The covariance measures how “correlated" two RVs are with respect to each other. There is a catch, two different pair of RVs may have the same degree of correlation, but their covariance may be very different. For example, it is clear that

$$Cov(X, X) = Var(X).$$

Intuitively, the degree of "correlation" between X and X, versus $100X$ and $100X$ should be the same (they are perfectly correlated in each case). However, you can easily check that $Cov(100X, 100X) = 10000Cov(X, X)$. Thus we need to introduce another quantity that measures only the correlation and not affected by scaling of the RVs. That is the correlation:

Let X, Y be RVs. We definte their correlation as

$$Corr(X, Y) = \frac{Cov(X, Y)}{\sqrt{Var(X)Var(Y)}}.$$

Joint distribution, joint pdf

When we have 2 RVs X, Y, besides describing each individual distribution of X, Y, we also need to know how they interact together. The joint distribution (in the discrete case) or the joint pdf (in the continuous case) gives us this information. In fact, to calculate $E(XY)$ in the Covariance formula we would need to use the joint distribution of X, Y.

a. Discrete: Let X, Y be discrete RVs. Then the joint distribution of X, Y is $P(X = x, Y = y)$.

b. Continuous: Let X, Y be continuous RVs. Then their joint pdf, denoted $f_{X,Y}(x, y)$ is such that

$$P(X \leq x, Y \leq y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(u, v) du dv.$$

Some elementary properties:

a.

$$\sum_{x,y} P(X = x, Y = y) = 1.$$

b.

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(u, v) du dv = 1.$$

c. Discrete:

$$E(XY) = \sum_{x,y} xy P(X = x, Y = y).$$

d. Continuous:

$$E(XY) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} uv f_{X,Y}(u, v) du dv.$$

More generally

e. Discrete:

$$E(g(X, Y)) = \sum_{x,y} g(x, y) P(X = x, Y = y).$$

f. Continuous:

$$E(g(X, Y)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(u, v) f_{X,Y}(u, v) du dv.$$

Independence

Two random variables X, Y are independent if all events they generated are independent. More specifically, X, Y are independent if for all x, y:

$$P(X \leq x, Y \leq Y) = P(X \leq x) Y(\leq y).$$

An easier criterion to check is if the joint distribution "splits", i.e.

$$P(X = x, Y = y) = P(X = x) P(Y = y) (\text{ discrete}), \text{ or}$$

$$f_{XY}(x,y) = f_X(x)f_Y(y) \text{ (continous)} .$$

An important property is that if X, Y are independent then $E(XY) = E(X)E(Y)$. Note that, the reverse implication is not generally true. That is $E(XY) = E(X)E(Y)$ does NOT imply that X, Y are independent. See the following example.

Example 4.1.9. *Let X have the following distribution: $P(X = 1) = P(X = 0) = P(X = -1) = 1/3$, and let $Y = X^2$. Then it is clear that X, Y are NOT independent (you should try to show this using the definition of independence). However, we can also easily check that*

$$E(XY) = E(X)E(Y) = 0.$$

4.1.4 Conditional expectation

Conditional distribution, conditional density

We have discussed conditional probability $P(A|B)$, which is the probability that A happened given the knowledge that B has happened. In a similar way, for 2 RVs X, Y, we can talk about the probability that X takes some value x given that we know Y has taken some y. If X and Y are correlated in some way, the fact that we have seen Y taking some value should change the probability that X taking value x. Formally, we define, for 2 discrete RVs X, Y

$$P(X = x|Y = y) = \frac{P(X = x, Y = y)}{P(Y = y)}.$$

For continuous RVs, we cannot talk about the probablity that X takes some value, given that we have observed Y taking some value. The reason is the probability that Y taking some value is 0, since it is a continuous RV. This poses a slight problem, since in reality, we always observe Y taking some particular value, even if it is a continuous RV (think about the amount of time you wait for the bus to arrive, for example. You always have to wait a particular amount of time until the bus arrives, even if the probability that the continuous random variable representing the time you wait taking that particular value is 0). So for continuous RVs, we talk about the conditional density instead. Formally, we define, for 2 continuous RVs X, Y

$$f_{X|Y}(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)}.$$

Remark: In the two formulas above, we think of y as *fixed*, and x as taking any possible values in the range of X. Thus the conditional distribution, or conditional density, is a function of x, given a fixed value y. Moreover, for a fixed y, the conditional distribution (or probability density), is a probability distribution (or density). That is

$$\sum_x P(X = x|Y = y) = 1; \tag{4.1}$$

$$\int_{-\infty}^{\infty} f_{X|Y}(x|y)dx = 1. \tag{4.2}$$

Proof. Left as an exercise.

Conditional probability and conditional expectation

Discrete : Let X, Y be discrete RVs. The conditional probability $P(X = i|Y = j)$ was defined naturally using the definition of conditional expectation as above. Note that we also have

$$P(X \leq k|Y = j) = \sum_{i \leq k} P(X = i|Y = j).$$

In this way, for every y, conditioned on $Y = y$, $P(X < a|Y = y)$ is a proper cumulative distribution function, even though $P(Y = y) = 0$. This is related to the notion of regular conditional probability distribution, discussed below.

We define the conditional expectation of X, given $Y = y$ as

$$E(X|Y = y) = \sum_{x} xP(X = x|Y = y).$$

Continuous :

Let X, Y be continuous RVs. Note that we can NOT define $P(X < a|Y = y)$ using the definition of conditional expectation, because $P(Y = y) = 0$. However, we can define it as followed:

$$P(X < a|Y = y) = \int_{-\infty}^{a} f_{X|Y}(x|y)dx.$$

We define the conditional expectation of X, given $Y = y$ as

$$E(X|Y = y) = \int_{-\infty}^{\infty} x f_{X|Y}(x|y)dx.$$

Interpretation: Besides the fact that conditional expectation is the average (or mean) value of X given $Y = y$, it is also the *best guess* of X given $Y = y$, in some precise sense that we will discuss below.

Remark: Note that in these definitions, $E(X|Y = y)$ is *a real number*. This will be constrasted with $E(X|Y)$, which is a RV, the definition of which is given below.

Abstract definition of conditional expectation

The above definitions of $E(X|Y = y)$, while useful, is rather restrictive. It is because we do not have to observe the value of Y to be able to talk about the expectation of X conditioned on Y in a meaningful way. An example will explain. It is clear that the stock price of today depends on the stock price of yesterday (for simplicity let's suppose that stock price only changes discretely from day n to day $n + 1$). Suppose we are at day 0,

which is today, and we want to discuss our "expectation", or our best guess, of the stock price on day $n+1$, the guess being made on day n. It is clear that on day n, we have the knowledge of the stock price of that day, say S_n. So what we're asking for is $E(S_{n+1}|S_n)$. Since we are still at day 0, we do not know what value S_n is, it is a RV to us. However, to discuss our action on day n, in anticipation of day $n+1$, it is necessary that we make sense of the notion $E(S_{n+1}|S_n)$. Thus we need an abstract definition of conditional expectation, one that doesn't require us to plug in an observed value for the RV being conditioned on. We will als refer to this as *the measure theoretic definition* of conditional expectation.

Definition 4.1.10. *Let X, Y be RVs. The conditional expectation $E(X|Y)$ is a function of Y, such that for any function g, we have*

$$E\big[E(X|Y)g(Y)\big] = E\big[Xg(Y)\big].$$

Remark: Note that in contrast with the above, as we already said, $E(X|Y)$ is a RV, since it is a function of Y (in some trivial case it could be the constant function, but this does not happen usually). The interpretation of the equality in the definition is that as far as taking expectation with respect to function of Y, it does not matter if we use the conditional expectation $E(X|Y)$ or X itself. Thus the conditional expectation $E(X|Y)$ is a guess of X, in terms of the random variable Y, which satisfies some "indifference" property in terms of expectation.

Perhaps a more satisfactory property of $E(X|Y)$ is that not only it is a guess of X given Y, it is *the best* guess of X given Y in the following sense:

Lemma 4.1.11. *Let X, Y be RVs. Then for any function g we have*

$$E\Big([E(X|Y) - X]^2\Big) \leq E\Big([g(Y) - X]^2\Big).$$

Proof. Left as an exercise.

Some elementary properties :

The definition (4.1.10) unfortunately does not, most of the time, give us an easy way to compute what $E(X|Y)$ is. So the followings are some elementary properties of conditional expectation that will help us do that. You should try to prove these properties yourself.

a. $E(E(X|Y)) = E(X)$.

b. $E(aX + bY|Z) = a(EX|Z) + bE(Y|Z), a, b$ constant .

c. If X is independent of Y then

$$E(X|Y) = E(X).$$

d. For any function g,

$$E(g(Y)X|Y) = g(Y)E(X|Y).$$

e. *The independence lemma:* If X is independent of Y then for any function g

$$E\big[g(X,Y)|Y\big] = E\big[g(X,y)\big]|_{y=Y}.$$

(Properties f and g are not used in this course except in the discussion of regular conditional probability. You can skip these.)

f. If $X_n \geq 0, X_n \uparrow X$ then $E(X_n|\mathcal{G}) \uparrow E(X|\mathcal{G})$.

g. If $X \in L_2(\Omega, \mathcal{F}, P)$ then $E(X|\mathcal{G})$ is the orthogonal projection of X onto the subspace $L_2(\Omega, \mathcal{G}, P)$ in the Hilbert space $L_2(\Omega, \mathcal{F}, P)$ with inner product $\langle X, Y\rangle := E(XY)$.

Remark: The expression $E\big[g(X,y)\big]|_{y=Y}$ means that we just evaluate $E\big[g(X,y)\big]$ as a *regular expectation* (it is only a random variable in terms of X, y is understood to be a constant (or just a dummy variable) here. Note that $E\big[g(X,y)\big]$ is *a function of* y. Thus we are free to plug in the random variable Y after we compute what $E\big[g(X,y)\big]$ is.

Example 4.1.12. *Let X be a Bernoulli(1/2) random variable and Y has Normal(0,1) distribution, X independent of Y. Compute $E(Y^X|Y)$.*

Ans: We have

$$E(y^X) = y^0 \cdot \frac{1}{2} + y^1 \cdot \frac{1}{2} = \frac{1}{2}(1+y).$$

Thus y the independence lemma, $E(Y^X|Y) = \frac{1}{2}(1+Y)$. Note how the distribution of Y is irrelevant in this computation.

Expectation conditional on more than one random variables

In applications, a random variable X may be correlated to not just 1 random variable Y, but possibly to n random variables $Y_1, Y_2, \cdots, Y_n$ (it is reasonable to build a model of stock so that the stock price today does not just depend on its performance yesterday, but on its performance in the past month). To discuss the behavior of X given our observations of $Y_1, \cdots Y_n$, we need to extend our notion of conditional expectation to more than 1 random variable. The extension actually is straightforward.

Definition 4.1.13. *Let $X, Y_1, Y_2, \cdots, Y_n$ be RVs. The conditional expectation $E(X\big|Y_1, \cdots, Y_n)$ is a function of $Y_1, Y_2, \cdots, Y_n$, such that for any function g, we have*

$$E\big[E(X\big|Y_1, Y_2, \cdots, Y_n)g(Y_1, Y_2, \cdots, Y_n)\big] = E\big[Xg(Y_1, Y_2, \cdots, Y_n)\big].$$

Remark: Actually in subsection (4.1.4), there is no restriction on what the RV Y can be. Thus one could select it to be a *multi-dimensional RV*, effectively making it a random vector with n components

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ \cdots \\ Y_n \end{bmatrix}.$$

Even more generaly, X itself can also be a multi-dimensional RV,

$$X = \begin{bmatrix} X_1 \\ X_2 \\ \cdots \\ X_m \end{bmatrix}.$$

Thus we see that we have covered the case of expectation conditional on more than one random variables:

$$E(X_1, X_2, \cdots X_m | Y_1, Y_2, \cdots, Y_n)$$

in subsection (4.1.4), including the elementary properties. One just needs to interpret the symbol accordingly, for example in $E(aX + bY|Z)$, a, b has to be understood as vector, aX and bY as vector dot products if X, Y are multi-dimensional RV.

Probability as an expectation

You may observe that in the abstract definition of conditional expectation, we did not mention about conditional probability. Surely we would want to have a definition for $P(X \leq x|Y)$. It turns out that our definition of conditional expectation *already covers conditional probability as a special case.* To be precise, we first need to introduct the following so-called indicator function of an event E, denoted as $\mathbf{1}_E$

$$\begin{aligned} \mathbf{1}_E(\omega) &= 1 \text{ if } \omega \in E \\ &= 0 \text{ if } \omega \notin E. \end{aligned}$$

Basically the indicator function is a logical indicator, it's 1 if E happens and 0 if E does not happen. For example $\mathbf{1}_{\{0<1\}} = 1$ and $\mathbf{1}_{\{1+1<3\}} = 0$. But now note that suppose we have a random variable X, and say it has a density function $f_X(x)$ then

$$\begin{aligned} E(\mathbf{1}_{\{X \leq x\}}) &= \int_{-\infty}^{\infty} \mathbf{1}_{\{y \leq x\}}(y) f_X(y) dy \\ &= \int_{-\infty}^{x} f_X(y) dy = P(X \leq x). \end{aligned}$$

where the second equality is because $\mathbf{1}_{\{y \leq x\}} = 0$ for all values of $y > x$ so we just stop the integration limit at x. Similarly, you can check that

$$\begin{aligned} E(\mathbf{1}_{\{X \geq x\}}) &= P(X \geq x) \\ E(\mathbf{1}_{\{X = x\}}) &= P(X = x). \end{aligned}$$

Thus probability can be expressed as an expectation. More importantly for us, this is still true at the conditional expectation level. More precisely we have the following

Lemma 4.1.14. *Let X, Y be random variables. Let $f(Y) = E(\mathbf{1}_{\{X \leq x\}}|Y)$. Then $f(y) = P(X \leq x|Y = y)$ where $P(X \leq x|Y = y)$ is understood in the sense of subsection (4.1.4). Similarly for $P(X \geq x|Y = y), P(X = x|Y = y)$.*

Remark: For a fixed x, the expression $\mathbf{1}_{\{X \leq x\}}$ here is understood as function of X. Thus the expression $E(\mathbf{1}_{\{X \leq x\}}|Y)$ is understood in the sense of $E(g(X)|Y)$ where $g(X)$ is just a random variable.

4.1.5 Connection between the measure theoretic and classical definition of conditional expectations

Discrete RVs

Let X, Y be two RVs. We have seen that we can define $E(X|Y)$ abstractly via definition (4.1.10). Suppose X, Y are both discrete. Then we also have alternative definitions of $E(X|Y = y)$ via classical probability theory. How are these two connected?

Note that $E(X|Y)$ by definition if a function of Y. Thus we can write $E(X|Y) = g(Y)$ for some function g. On the other hand, $E(X|Y = y)$ is also clearly a function of y. So you can expect that $\forall y$ on the event $\{Y = y\}$

$$E(X|Y) = E(X|Y = y).$$

That is

$$E(X|Y)\mathbf{1}_{Y=y} = E(X|Y = y)\mathbf{1}_{Y=y}.$$

Proof. We need to check that for any function $g(Y)$

$$E\Big[E(X|Y = y)\mathbf{1}_{Y=y}g(Y)\Big] = E\Big[X\mathbf{1}_{Y=y}g(Y)\Big].$$

The LHS is equal to

$$\begin{aligned} E(X|Y = y)g(y)P(Y = y) &= \sum_i \frac{iP(X = i, Y = y)}{P(Y = y)} g(y)P(Y = y) \\ &= \sum_i iP(X = i, Y = y)g(y) \\ &= E\Big[Xg(y)\mathbf{1}_{Y=y}\Big] = RHS. \end{aligned}$$

Continuous RVs

What about the case when both X, Y are continuous? Here we cannot use the above criterion, as the event $\{Y = y\}$ has probability 0. Rather we will turn it around, and observe

that since $E(X|Y=y)$ is a function of y, we can also write $E(X|Y=y)=g(y)$. So now we can plug the RV Y into the function g and we claim

$$E(X|Y)=g(Y), P \text{ a.s.},$$

where the a.s. notation means the equality holds outside an event of probability 0 with respect to P.
Proof.

$$E(X|Y=y)=\int_{-\infty}^{\infty} x\frac{f_{XY}(x,y)}{f_Y(y)}dx.$$

Therefore

$$g(Y)=\int_{-\infty}^{\infty} x\frac{f_{XY}(x,Y)}{f_Y(Y)}dx$$

We need to check that for any function $h(Y)$

$$E\Big(h(Y)\int_{-\infty}^{\infty} x\frac{f_{XY}(x,Y)}{f_Y(Y)}dx\Big)=E\Big(Xh(Y)\Big).$$

But the LHS is equal to

$$\begin{aligned}&\int_{-\infty}^{\infty} h(y)\Big[\int_{-\infty}^{\infty} x\frac{f_{XY}(x,y)}{f_Y(y)}dx\Big]f_Y(y)dy\\ =\ &\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} h(y)xf_{XY}(x,y)dxdy=E(Xh(Y))=RHS.\end{aligned}$$

4.1.6 Law of large number

The theorem

Theorem 4.1.15. *Let $X_1, X_2, \cdots$ be a sequence of independent identically distributed (abbreviated as i.i.d.) RVs such that $E|X_1|<\infty$. Then with probability 1,*

$$\frac{\sum_{i=1}^n X_k}{n}\to E(X_1).$$

Notation: It is usually denoted that $S_n=\sum_{i=1}^n X_k$, thus one usually sees $\frac{S_n}{n}\to E(X_1)$ in the statement of the law of large number (LLN).

Interpretation: Suppose that you play a game where your winning is random, which is represented by a RV X. As you play this game many times, you will find that your average earning (over time) is approximately the expected value of X.

Application

We'll give one application of the LLN, in pricing of a random game: Suppose there is a game of tossing a fair coin, where if the coin turns up H then you get paid 3 dollars. If it turns up T, then you get paid 1 dollars. Question: What is the fair price to charge for this game?

Ans: The fair price to charge for this game is $3\frac{1}{2} + 1\frac{1}{2} = 2$ (dollars). But can you explain why this is the fair price? The reason is the LLN. If this game is played *only once* (an important point, which we'll come back later when we discuss the fair price of financial instrument) then it is not clear that the price is fair. However, the assumption here is that the game will be played *many times,* by potentially many different players. Thus each player's winning is an independent, identically distributed random variable, which takes values 3 and 1 with probability 1/2 each. The total amount of money the house has to pay to these players, after n games have been played, is $\sum_{i=1}^{n} X_i$. By the LLN, this is approximately $nE(X_1)$, which is $2n$, which is the total amount charged by the house. So the house comes out even and this is a fair price for the game.

Remark: This is the main principle behind casino's operation (and profitability). Of course the players are not charged to play the games in the casino. But the game is set up so that the expectation is negative (even if you bet on a roulette table, say on an even number, your chance of winning is still less than 1/2, since there is a 0 and double 0's). Thus by the LLN, with a lot of customers, the casino will have a positive profit. Note that the LLN does allow for an occasional incident where someone plays 1 single game and win big. But if you play a lot of games at the casino, the LLN says that you will lose money eventually.

4.1.7 Central limit theorem

The theorem

The LLN gives us an estimate of $\frac{S_n}{n}$ (it is approximately $E(X_1)$). However, for various reasons, we may want a more precise estimate than that. Note that the LLN says nothing about how close to $E(X_1)$ $\frac{S_n}{n}$ is, or (perhaps surprisingly) what distribution we may approximate $\frac{S_n}{n}$ with. It is surprising because we do not have any restriction on the distribution of each individual X_i, but it turns out that the approximate distribution of $\frac{S_n}{n}$ is the normal distribution. The precise statement is as followed:

Theorem 4.1.16. *Let $X_1, X_2, \cdots$ be i.i.d. RV such that $E|X_1|^2 < \infty$. We will also denote $E(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Then for any real number x,*

$$P(\frac{S_n - n\mu}{\sigma\sqrt{n}} \leq x) \to P(Z \leq x),$$

where Z has standard normal $(N(0,1))$ distribution.

Application

The central limit theorem is used to estimate probability of the *sum* or the *average* of an i.i.d. sequence of RVs. Determining which case to use requires a close reading into the problem.

Example 4.1.17. *The bus arrives at the Hill center according to a uniform* $[0, 12]$ *distribution. Suppose you wait at the Hill center bus stop for 30 days. What is the approximate probability that your* **average** *wait time is more than 5 minutes?*

Ans: Let $X_1, X_2, \cdots$ *be i.i.d.* $U[0, 12]$*. Then* $E(X_1) = 6$ *and* $Var(X_1) = 12$*. Thus*

$$P(\frac{S_{30}}{30} \geq 5) = P(\frac{S_{30} - 6 \times 30}{\sqrt{12 \times 30}} \geq \frac{5\sqrt{30}}{\sqrt{12}} - \frac{6\sqrt{30}}{\sqrt{12}}) \approx P(Z \geq -1.58).$$

Example 4.1.18. *The earning per day of a casino is distributed as an Exponential(1) RV. (1 here stands for 1 million, we omit the unit). What is the approximate probability that the casino's earning in 1 month is more than 35 millions?*

Ans: Note that here we're asked for the **total** *earning. Thus let* $X_1, X_2, \cdots$ *be i.i.d.* $Exp(1)$*. Then* $E(X_1) = 1$ *and* $Var(X_1) = 1$*. Thus*

$$P(S_{30} \geq 35) = P(\frac{S_{30} - 30}{\sqrt{30}} \geq \frac{5}{\sqrt{30}}) \approx P(Z \geq \frac{5}{\sqrt{30}}).$$

4.1.8 Moment generating function and characteristic function

Given a random variable X, the moment generating function of X is

$$M_X(t) := \mathbb{E}(e^{tX}) \text{ if it exists}$$

and the characteristic function of X is

$$\Phi_X(t) := \mathbb{E}(e^{itX}).$$

The values of t so that $\mathbb{E}(e^{tX}) < \infty$ is the domain of the moment generating function $M_X(t)$. This domain can be empty, for example with the Cauchy distribution. On the other hand, the characteristic function always exists because $|e^{itX} \leq 1|$.

If the domain of $M_X(t)$ includes an open interval around 0 then moments of all order of X exist and in particualr $E(X^n) = M_X^{(n)}(0)$. Moreover, $M_X(t) = M_Y(t)$ for two random variables X, Y if and only if X, Y have the same distribution.

If a characteristic function $\Phi_X(t)$ has a k-th derivative at zero, then the random variable X has all moments up to k if k is even, but only up to k âĂŞ 1 if k is odd. If a random variable X has moments up to k-th order, then the characteristic function $\Phi_X(t)$ is k times continuously differentiable on the entire real line. In either case

$$\Phi_X^{(k)}(0) = i^k \mathbb{E}[X^k].$$

We also have $\Phi_X(t) = \Phi_Y(t)$ for two random variables X, Y if and only if X, Y have the same distribution.

4.1.9 Multivariate normal distribution

The multivariate normal distribution is ubiquitous in financial mathematics and other disciplines concerned with modeling. Underlying this ubiquity is the central limit theorem and the remarkable fact that the (finite) sum of independent normal distributions also has normal distribution.

Definition: $X_1, X_2, \cdots, X_n$ has multivariate normal distribution with mean $\boldsymbol{\mu}$ and (invertible) covariance matrix Σ if their joint density has the form :

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{1}{\sqrt{(2\pi)^n det(\Sigma)}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu})}.$$

Here $\mathbf{x} := [x_1, x_2, \cdots, x_n]^T$, $\boldsymbol{\mu} = [\mu_1, \mu_2, \cdots, \mu_n]^T$.

The following is a key result for multivariate normal distribution : Let $\mathbf{Z} := [Z_1, Z_2, \cdots, Z_n]^T$ be independent standard Normals. Let A be a $m \times n$ matrix with independent columns and $\boldsymbol{\mu}$ an $\mathbb{R}^m$ vector. Then $A\mathbf{Z} + \boldsymbol{\mu}$ has a multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma = AA^T$. This result can be verified by a standard application of multivariate change of variables technique.

Remark: If A, B are such that $AA^T = BB^T$ then $\mathbf{X}_1 = A\mathbf{Z} + \boldsymbol{\mu}$ and $\mathbf{X}_2 = B\mathbf{Z} + \boldsymbol{\mu}$ has the same multivariate normal distribution. This does NOT mean $\mathbf{X}_1 = \mathbf{X}_2$. A simple example is $X_1 = -2Z + 1$ and $X_2 = 2_Z + 1$. X_1, X_2 are both Normally distributed with mean 1 and variance 4. But they obviously are two different random variables in terms of their realization.

An important corrolary from this result is a linear transformation of a multivariate normal distribution is also a multivariate normal distribution. That is if $\mathbf{X}$ has multivariate normal distribution and $\mathbf{Y} = A\mathbf{X} + \mathbf{b}$ then $\mathbf{Y}$ also has multivariate normal distribution with appropriate mean and covariance matrix.

In Monte Carlo simulation, to generate a multivariate normal distribution $\mathbf{X}$ with a given mean $\boldsymbol{\mu}$ and covariance matrix Σ, one starts with vector $\mathbf{Z}$ of independent standard normals. Let A be the Cholesky decomposition of Σ. Then $\mathbf{X} = A\mathbf{Z} + \mu$ has the desired multivariate normal distribution.

4.2 Problems

1. Our friend John tells us that he has two daughters. we also know that he has three children in total. What is the probability that his youngest child is a girl (assuming that a boy and a girl are equally likely)?

2. The arrival time of shuttles at a bus stop from 5:00pm to 5:30pm on a weekday is uniformly distributed. In other words, let X be the waiting time (in minutes) after 5:00 pm until the arrival of a particular shuttle, then X has the p.d.f

$$f_X(x) = \frac{1}{30}, 0 \leq x \leq 30. \tag{4.3}$$

Suppose there are 30 shuttles arriving at the bus stop from 5:00 pm to 5:30 pm, and their distribution are i.i.d. What, then, is the approximate probability that their average arrival time on a particular weekday is before 5:12 pm?

3*. Studying probability and statistics has a positive effect on students' job placement. A student who has succesfully completed these classes has a 70% probabilty of landing a job with Goldman Sachs. A student who did not succesfully complete the program, however, only has a 40% probability of landing such a job. our friend, Tom, just got a position with Goldman Sach. Suppose the probability of a student succesfully completing the Financial Math program at Rutgers is 80%. What is the probability that Tom succesfully finished his Financial Math program at Rutgers?

4. The breakdown time of the Apple Ipad is an exponential(5) random variable. In other words, let X be the time until break down (in years) of a particular Ipad, then X has the p.d.f

$$f_X(x) = \frac{1}{5}e^{-\frac{1}{5}x}, 0 \leq x < \infty. \tag{4.4}$$

Suppose an Apple store has 30 Ipads, and the distribution of their break down times are i.i.d. What, then, is the approximate probability that the average break down time of these Ipads is before 2 years?

5. There are 120 students in the Introduction to Probability class. Suppose that each student has a probability of .3 of getting an A in this class, and the students' performance is independent of one another. What, then, is the approximate probability that the class will have at least 20 students getting an A ?

6. In this problem we will verify that the conditional expectation $E(X|Y)$ is the best guess of X given Y in the following sense

$$E\big[(X - E(X|Y))^2\big] \leq E\big[(X - g(Y))^2\big], \text{ for all } g(Y). \tag{4.5}$$

a. Show that

$$E\big[E(X|Y)X\big] = E\big[E(X|Y)^2\big].$$

Hint:

$$E\big[E(X|Y)X\big] = E\Big[E\Big(E(X|Y)X\Big|Y\Big)\Big].$$

Proceed using properties of conditional expectation.

b. Use the result of part a to prove (4.5).

7. Let X, Y be independent random variables, Y having Normal(0,1) distribution and X has distribution $P(X = 1.5) = P(X = 0.5) = 1/2$. Let $Z = XY$. Compute

a. $E(Z|Y)$.

b. $E(Z^2|Y)$.

8.

a. Let X have distribution Uniform[0,Y] distribution, where Y has Exp(1) distribution. Compute the joint distribution of X, Y and $E(X)$.

b*. Let X have distribution Exponential(Y) distribution where Y has Uniform[1,2] distribution. Compute the joint distribution of X, Y and $E(X)$.

9. Prove equations (4.1) and (4.2).

10. For $\alpha > 0$ we define the function

$$\varphi(x) \triangleq \frac{1}{\alpha} x^{-\frac{1+\alpha}{\alpha}}, \quad x > 1.$$

a. Explain why φ is a density function.

b. Let X be a random variable with density φ. Compute the expectation $\mathbb{E}[\log(X)]$ (as always(!) $x \to \log(x)$ denotes the natural logarithm, i.e., $\log(e^x) = x$).

11*. Let X be a random variable with density function $f(x) \triangleq \frac{1}{2} e^{-|x|}$, $x \in \mathbb{R}$. The corresponding distribution is called the Laplace distribution.

a. Compute $\mathbb{E}[|X|^n]$ and $\mathbb{E}[X^n]$ for all $n \in \mathbb{N}$.

b. Find X's characteristic function, $t \to \mathbb{E}[e^{itX}]$.

12.Let $\alpha > 0$ and $\mu \in \mathbb{R}$ be given and define the function $f : \mathbb{R} \to \mathbb{R}$ by

$$f(x) \triangleq ce^{-\alpha|x-\mu|}, \quad x \in \mathbb{R}.$$

a. Find the value of c such that f is a density function.

b. Let Y be a random variable with density f. Compute Y's moment generating function $\mathbb{E}[e^{tY}]$ and use this function to compute the mean $\mathbb{E}[Y]$ and the variance $\mathbb{V}[Y]$ provided they exist. *Warning: this function is not finite for all t and you need to figure out 1) when it is finite and 2) what value it has when it is finite.*

c. Let X be a strictly positive random variable such that the density for $\log(X)$ is the function f (with the constant c as determined in the first question). Compute the mean $\mathbb{E}[X]$ and the variance $\mathbb{V}[X]$ provided they exist.

13. We let a be a strictly positive constant, $a > 0$, and we define the function

$$F(x) = \begin{cases} 1 - x^{-a} & \text{for } x \geq 1 \\ 0 & \text{else.} \end{cases}$$

Let X be a random variable with F as its distribution.

a. Compute X's density function.

b. For which values of a does X have a finite mean?

14. Let X be a standard normally distributed random variable, $X \sim \mathcal{N}(0, 1)$.

a. Find the density function of the random variable Y given by $Y \triangleq X^2$. This distribution is called the χ^2 distribution and plays a key role in term structure theory.

b. Find the density function of the random variable Y given by $Y \triangleq |X|$.

15. Define the function

$$f(x) \triangleq \begin{cases} 0 & \text{for } x < 0 \\ 0.5 & \text{for } x \in [0,1] \\ \frac{1}{x^3} & \text{for } x > 1. \end{cases}$$

a. Explain why f is a density function.

Let X be a random variable with f as its density function.

b. Compute X's mean, $\mathbb{E}[X]$. Does X also have a variance?

c. Compute the probability $\mathbb{P}(X \in [0,2])$.

16. Let σ be a positive constant and let $\varphi(y)$ be the standard normal density $\varphi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2}$.

a. Use the fact that $\int_{-\infty}^{\infty} \varphi(y)\, dy = 1$ to compute

$$\int_{-\infty}^{\infty} e^{\sigma y} \varphi(y)\, dy \quad \text{and} \quad e^{\sigma \int_{-\infty}^{\infty} y\varphi(y) dy}.$$

Which of these is larger? Note that if Y is a standard normal random variable, we are comparing $\mathbb{E}e^{\sigma Y}$ and $e^{\sigma \mathbb{E} Y}$.

b. Relate the previous question to Jensen's inequality.

17. Let X be a standard normally distributed random variable, $X \sim \mathcal{N}(0,1)$.

a. Compute the characteristic function $t \to \mathbb{E}[e^{itX}]$ for $t \in \mathbb{R}$. (***Hint: derive an ODE that completely characterizes this function***).

b. Use the previous question to compute $\mathbb{E}[X]$ and $\mathbb{V}[X]$ (***Hint: take derivatives***).

18. Let X and Y be two independent standard normals, $X \sim \mathcal{N}(0,1)$ and $Y \sim \mathcal{N}(0,1)$. Compute the probability $\mathbb{P}(X \leq Y)$.

19. Let (X,Y) be uniformly distributed in the triangle $T \triangleq \{(x,y) | x \geq 0, y \geq 0, x + y \leq 2\}$

a. Find the density function for (X,Y).

b. Find the density function of X.

c. Find the conditional density function of Y given $X = x$.

d. Compute the term $\mathbb{E}[Y|X = x]$.

19. Let X and Y be two independent random variables both uniformly distributed between zero and one. Define the following random variables

$$U \triangleq XY, \quad V \triangleq \frac{X}{Y}.$$

a. Compute the marginal density functions for both U and V.

b. Find the two dimensional density function for the pair (U,V). Are U and V independent? Why? Why not?

20. Let X and Y be two independent random variables with the same density function f given by

$$f(z) = \begin{cases} 2e^{-2z} & \text{for } z > 0 \\ 0 & \text{else.} \end{cases}$$

Define the random variables

$$U \triangleq X + Y, \quad V \triangleq X - Y.$$

a. Compute the following terms: $\mathbb{E}[U], \mathbb{E}[V], \mathbb{V}[U], \mathbb{V}[V]$ and $Cov(U, V)$.

b. Find the two dimensional density function for pair of random variables (U, V).

c. Find the marginal density functions for both U and V. Are U and V independent?

21. Let X and Y be independent random variables with the same exponential density. Define the random variables U and V by $U \triangleq X + Y$ and $V \triangleq X/Y$.

a. Compute the marginal density functions for both U and V.

b. Find the two dimensional density function for the pair (U, V). Are U and V independent?

22. Let X and Y be two independent random variables both normally distributed with mean zero and variance one. Define the following random variables

$$U \triangleq \sqrt{X^2 + Y^2}, \quad V \triangleq \frac{X}{Y} \text{ if } Y \neq 0 \text{ otherwise } V \triangleq 0.$$

a. Compute the marginal density functions for both U and V.

b. Find the two dimensional density function for the pair (U, V). Are U and V independent?

23. Let X and Y be two random variables both normally distributed with mean zero, variance one and correlation coefficient ρ.

a. Find the density and the distribution function of the random variable $Z \triangleq \frac{X}{Y}$.

b. Compute the probability $\mathbb{P}(X < 0, Y > 0)$.

24. Let X be a random variable with density function

$$f(x) = \begin{cases} 2(e^{-x} - e^{-2x}) & \text{for } x > 0 \\ 0 & \text{else.} \end{cases}$$

a. Explain why f is a density function and find the corresponding distribution function for X.

b. Find the Laplace transform for X, $t \to \mathbb{E}[e^{-tX}]$, as well as the moment generating function for X, $t \to \mathbb{E}[e^{tX}]$ (as usual you have to be careful about whether or not the function is finite valued).

c. Assume that U and V are two independent random variable both having density function f. Find the Laplace transform of $U + V$.

25*. Let X be exponential with parameter $\mu > 0$ and let Y and Z be two random variables that are both exponential with parameter $\lambda > 0$. Assume that X, Y and Z are all independent.

a. Compute the probability that X is bigger than 10.

b. Compute the probability that X is at least twice as big as both Y and Z.

26*. Let the two dimensional random variable (X, Y) have density function

$$f(x,y) = \begin{cases} \frac{1}{2}e^{-x} & \text{for } -x < y < x, \quad x > 0 \\ 0 & \text{else.} \end{cases}$$

a. Find the marginal densities for X and for Y.

b. Are X and Y independent? Why? Why not?

c. Define the two random variables $U \triangleq X + Y$ and $V \triangleq X - Y$ and find the two dimensional density function for the pair (U, V).

d. Are U and V independent? Why? Why not?

27. Let (X, Y) be a two dimensional normally distribution random variable with the properties

$$\mathbb{E}[X] = \mathbb{E}[Y] = 0, \quad \mathbb{V}[X] = \mathbb{V}[Y] = 1, \quad Cov(X,Y) = -0.5.$$

a. What is the joint density function for (X, Y)?

b. Define the random variables

$$U \triangleq X, \quad V = \frac{1}{\sqrt{3}}(X + 2Y)$$

and find the joint density for U and V. Why are U and V independent?

27. Let X and Y be two independent random variables. X has density function

$$f(x) \triangleq \begin{cases} e^{-x} & \text{for } x \geq 0 \\ 0 & \text{else,} \end{cases}$$

whereas Y has density

$$g(x) \triangleq \begin{cases} xe^{-x} & \text{for } x \geq 0 \\ 0 & \text{else,} \end{cases}$$

a. Compute the following terms: $\mathbb{V}[X], \mathbb{V}[Y], \mathbb{V}[X + Y]$ and $Cov[X, X + Y]$.

b. Define the random variables

$$U \triangleq X, \quad V \triangleq X + Y$$

and compute the joint density function for the pair (U, V).

c. Compute the conditional density function for X given $X + Y = v$ for all $v > 0$.

28. Let X and Y have joint density function

$$f(x,y) \triangleq \begin{cases} c & \text{for } x \in [0,1], \quad -x \leq y \leq x \\ 0 & \text{else,} \end{cases}$$

where c is a constant.

a. Find c.

b. Compute X and Y's marginal density functions. Are X and Y independent?

c. Compute the terms: $\mathbb{E}[X], \mathbb{E}[Y], \mathbb{V}[X]$ and $Cov(X, Y)$.

d. Compute the probability $\mathbb{P}(X + Y \geq 1.5)$.

29. Let X, Y and Z be three independent random variables, all uniformly distributed between zero and one. We denote by F the uniform distribution function and we then define the two random variables

$$U \triangleq \max(X, Y, Z), \quad V \triangleq \min(X, Y, Z).$$

a. Express U's and V's distribution functions F_U and F_V in terms of F.

b. Compute the terms: $\mathbb{E}[U], \mathbb{E}[V], \mathbb{V}[U]$ and $\mathbb{V}[V]$.

c. Find the joint density function for (U, V) and use it to compute $Cov(U, V)$

30*. Let X and Y be two independent standard normally distributed random variables. Compute the expectation $\mathbb{E}[\max(X, Y)]$.

31. Let X and Y have density

$$f(x, y) \triangleq e^{-xy^2}, \quad x > 0, \quad y > 1.$$

a. Find Y's marginal density.

b. Compute X's conditional density given $Y = y$ and use this density to compute the conditional expectation $\mathbb{E}[X|Y = y]$ for $y > 1$.

c. Compute the mean of X, $\mathbb{E}[X]$.

31. Let X and Y have density

$$f(x, y) \triangleq e^{-2x}, \quad x > 0, \quad |y| < 2x.$$

a. Find X's marginal density.

b. Compute the mean of $X + Y$, $\mathbb{E}[X + Y]$.

c. Define the random variables

$$U \triangleq X + \frac{1}{2}Y, \quad V \triangleq X - \frac{1}{2}Y$$

and compute the joint density function for the pair (U, V).

d. Are U and V independent?

32. Let (X, Y) be a pair of random variables with joint density function

$$f_{X,Y}(x, y) = \begin{cases} \frac{2|x|+y}{\sqrt{2\pi}} \exp\left\{-\frac{(2|x|+y)^2}{2}\right\} & \text{if} \quad y \geq -|x|, \\ 0 & \text{if} \quad y < -|x|. \end{cases}$$

a. Show that X and Y are each standard normal random variables. In other words, determine the marginal densities of X and Y and show that these are standard normal.

b. Show that X and Y are uncorrelated, i.e., show that

$$\mathbb{E}[XY] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} xy f_{X,Y}(x,y)\, dx\, dy$$

is equal to zero. (Hint: Use the fact that $f_{X,Y}(x,y) = f_{X,Y}(-x,y)$.)

c. Suppose you are told that X takes some value x. Conditioned on this information, is Y still standard normal?

33*. Let X be standard normal, $X \sim \mathcal{N}(0,1)$, and let Z be independent of X and satisfy

$$\mathbb{P}(Z=1) = \mathbb{P}(Z=-1) = 0.5.$$

a. Show that $Y \triangleq ZX$ is also a standard normal.

b. Show that X and Y are uncorrelated but not independent.

34. Give an example of two random variables X and Y satisfying the property that X and Y are uncorrelated but not independent.

35. Let

$$\varphi_n(x) = \frac{1}{\sqrt{2\pi n}}\, e^{-x^2/(2n)}$$

be the normal density with mean zero and variance n.

a.

$$\lim_{n\to\infty} \int_{-\infty}^{\infty} \varphi_n(x)\, dx.$$

(Hint: To compute $\int_{-\infty}^{\infty} \varphi_n(x)\, dx$, you may use without proof the fact that the standard normal density integrates to 1: $\int_{-\infty}^{\infty} \varphi_1(y)\, dy = 1$.)

b.

$$\int_{-\infty}^{\infty} \lim_{n\to\infty} \varphi_n(x)\, dx.$$

Did you get the same answers in a and b?

36. Let X be a random variable with cdf $F_X(x)$. Suppose that $F_X(x)$ is strictly increasing so that $F_X^{-1}(x)$ is well defined. Show that $F_X(X)$ has a Uniform[0,1] distribution and $F_X^{-1}(U)$ where U is a Uniform $[0,1]$ distribution has the same distribution as X. This is the basis for Monte Carlo simulation and also for the concept of copula.

37. Bayesian formula has direct application in Bayesian estimates in statistics. In this framework, we suppose that our data follows a distribution that depends on some parameter θ. The goal is to have an estimate of θ based on the observation x of X. Before observing x, we have some idea about the distribution of θ. This is called the apriori distribution. After observing x we update this distribution and refer to it as the posterior distribution $f(\theta|x)$. The posterior estimate of θ based on the observation of x, denoted as $\widehat{\theta}(x)$ using the minimum mean square error (MSE) as standard, is defined as

$$\widehat{\theta}(x) = \mathbb{E}(\theta|x) = \int \theta f(\theta|x) d\theta.$$

Suppose that $X|\theta$ has a Normal(θ, σ^2) distribution and θ has an apriori distribution that is Normal(μ, τ^2). Our sample consists of only 1 data point x. Show that the posterior MSE estimate is

$$\widehat{\theta}(x) = \frac{\sigma^2}{\sigma^2 + \tau^2}\mu + \frac{\tau^2}{\sigma^2 + \tau^2}x.$$

References

[1] Boyce, W. E., DiPrima, R. C., & Meade, D. B. (1992). Elementary differential equations and boundary value problems. New York: Wiley.

[2] Lay, David C. “Linear Algebra and its applications, 1997."

[3] Stefanica, Dan. “A Linear Algebra Primer for Financial Engineering." FE Press 946 (2014): 947.

[4] Stewart, James. “Calculus."Lin Cengage Learning, 2011.