MISCELLANY

BRAD BAXTER

Contents

1. Introduction 2
2. Roots of Unity 3
3. The Length of the Day 4
3.1. The Length of the Day at the Solstices 4
3.2. The Variation in the Length of the Day during the Year 4
4. Distance Seen and Height 6
5. The Railway problem 7
6. A Derivation of the FFT 8
7. Constrained Optimization 9
7.1. Lagrange Multipliers 10
8. The Cholesky Factorization 12
9. Conics 14
9.1. The Ellipse and Hyperbola 14
9.2. The Reflector Property 14
9.3. Conic sections really are conic sections 15
10. The Birthday Problem 17
11. The Bike Problem via the Inclusion–Exclusion Formula 19
11.1. The Inclusion–Exclusion Formula 19
References 21

1. Introduction

This is a short collection of miscellanies intended for teaching.

Version: 201912181519

2. Roots of Unity

Example 2.1. *Let* $\omega = e^{2\pi i/3}$. *Thus* $1, \omega, \omega^2$ *are the three cube roots of unity. Then it is easily checked that* $|1-\omega| = |1-\omega^2| = \sqrt{3}$, *so that*

$$|1-\omega||1-\omega^2| = 3.$$

Example 2.2. *Suppose we take the* 4 *points* ± 1 *and* $\pm i$. *Then* $|1-i| = \sqrt{2}$, *so that*

$$|1-i||1-(-1)||1-(-i)| = 4.$$

These examples lead to a conjecture:

$$\prod_{k=1}^{n-1} |1-\omega^k| = n, \tag{1}$$

where $\omega = e^{2\pi i/n}$ and $n \geq 2$, and here is the Matlab code to check this.

```
I=sqrt(-1);
n=5;omega=exp(2*pi*I/n); P=1; for k=1:n-1, P=P*abs(1-omega^k); end; P
```

In fact, we shall see that a stronger statement is true:

Let $n > 1$ be an integer and let $\omega = e^{2\pi i/n}$. Thus the complex numbers $\{\omega^k : k = 0, 1, \ldots, n-1\}$ are the nth roots of unity. Thus

$$z^n - 1 = (z-1)\prod_{k=1}^{n-1}\left(z-\omega^k\right).$$

Hence

$$\prod_{k=1}^{n-1}\left(1-\omega^k\right) = \lim_{z\to 1}\frac{z^n-1}{z-1} = n,$$

by de L'Hôpital's rule.

3. The Length of the Day

3.1. **The Length of the Day at the Solstices.** We shall compute the length of the day at the Summer solstice in the northern hemisphere. The origin of our coordinate system will be at the centre of the Earth, the x-axis will point directly towards the Sun, and the z-axis will be perpendicular to the Earth's orbital plane and will be directed into the northern hemisphere. We need the following orthonormal vectors to describe the motion of a point on the Earth's surface:

$$\mathbf{u}_1 = \begin{pmatrix} \cos\alpha \\ 0 \\ -\sin\alpha \end{pmatrix}, \mathbf{u}_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \text{ and } \mathbf{u}_3 = \begin{pmatrix} \sin\alpha \\ 0 \\ \cos\alpha \end{pmatrix}, \tag{2}$$

where $\alpha = 23.5$ degrees approximately for the Earth.

The motion of a point at latitude θ is then

$$\begin{pmatrix} x(t) \\ y(t) \\ z(t) \end{pmatrix} = (\mathbf{u}_1 \cos t + \mathbf{u}_2 \sin t)\cos\theta + \sin\theta \mathbf{u}_3. \tag{3}$$

In particular, we have

$$\begin{aligned} x(t) &= \cos\theta\cos\alpha\cos t + \sin\theta\sin\alpha \\ y(t) &= \cos\theta \sin t. \end{aligned} \tag{4}$$

At the Summer solstice, day corresponds to $x(t) > 0$. Thus, solving $x(t) = 0$, we obtain the length of day as a function of the latitude θ:

$$L(\theta) = 2\cos^{-1}\left(-\tan\theta\tan\alpha\right), \quad |\theta| \le 90 - \alpha, \tag{5}$$

and this gives the length of the day in *degrees*. Thus the length of the day in hours is given by$L_h(\theta) = (24/360)L(\theta) = (1/15)L(\theta)$, i.e.

$$L_h(\theta) = \frac{2}{15}\cos^{-1}\left(-\tan\theta\tan\alpha\right), \quad |\theta| \le 90 - \alpha, \tag{6}$$

For $\theta \in (90-\alpha, 90)$, $L_h(\theta) = 24$; similarly $L(\theta) = 0$ for $\theta \in (-90, -90+\alpha)$.

The following Matlab code generates the ratio of the longest day to the shortest day.

```
alpha= 23.5*pi/180;
theta=0:pi/100:(pi/2) - alpha;
y = acos(-tan(alpha)*tan(theta));
R = y ./ (pi - y);
plot(theta,R)
plot((180/pi)*theta,R)
grid
```

3.2. **The Variation in the Length of the Day during the Year.** We solve the equation

$$\begin{pmatrix} x(t) \\ y(t) \end{pmatrix}^T \begin{pmatrix} \cos u \\ \sin u \end{pmatrix} = 0, \tag{7}$$

where $0 \le u \le 360$ measures orbital time in degrees, i.e. one year corresponds to 360 degrees. Expanding (7), we obtain

$$\cos\theta\cos\alpha\cos u\cos t + \cos\theta\sin u\sin t = -\sin\theta\sin\alpha\cos u, \tag{8}$$

or

$$\cos(t-\beta) = \frac{\sin\theta\sin\alpha\cos u}{\gamma}, \tag{9}$$

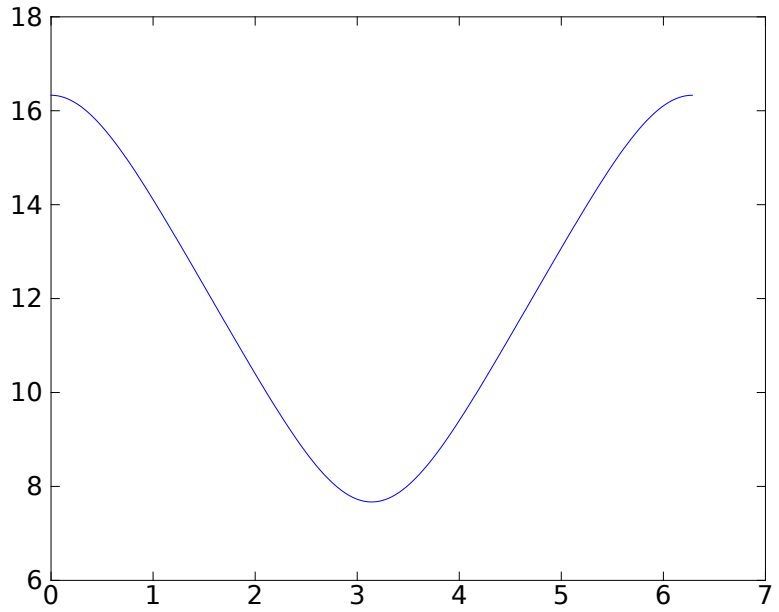

FIGURE 1. Annual variation in day-length at 51 degrees North

where

$$\gamma^2 = \cos^2\theta\cos^2\alpha\cos^2 u + \cos^2\theta\sin^2 u. \tag{10}$$

Hence the sunrise and sunset times are given by

$$t_\pm - \beta = \pm\cos^{-1}\left(\frac{-\sin\theta\sin\alpha\cos u}{\gamma}\right), \tag{11}$$

and the length of the day is then

$$t_+ - t_- = 2\cos^{-1}\left(\frac{-\sin\theta\sin\alpha\cos u}{\gamma}\right), \tag{12}$$

```
%
% Displays the yearly variation in the length of the day
% (in hours) at latitude theta, where |theta| < pi/2 - alpha.
%
alpha= 23.5*pi/180;
theta=51*pi/180;
u=0:pi/1000:2*pi;
A = -sin(theta)*sin(alpha)*cos(u);
B = cos(theta)*sqrt( (cos(alpha)^2)*(cos(u).^2) + (sin(u).^2) );
D = 2*acos(A ./ B)*12/pi;
plot(u,D)
%
% D is quite close to sinusoidal
%
%hold on
%plot(u, 12+(max(D)-12)*cos(u),'r')
%hold off
```

4. Distance Seen and Height

If we take the Earth to be a perfect sphere of radius R, then the distance seen D at height H is given by

$$D = R\theta \tag{13}$$

where

$$(R + H)\cos\theta = R. \tag{14}$$

It's useful to nondimensionalize these by introducing

$$d = \frac{D}{R} \quad \text{and} \quad h = \frac{H}{R}. \tag{15}$$

Thus (13) and (14) become

$$d = \theta \quad \text{and} \quad (1 + h)\cos\theta = 1. \tag{16}$$

Eliminating θ from (16) we obtain

$$\cos d = \left[\frac{1}{1+h}\right]. \tag{17}$$

If h is small, then d must also be small, so we have

$$1 - d^2/2 + \cdots = 1 - h + \cdots, \tag{18}$$

or

$$d^2 = 2h. \tag{19}$$

Returning to our original variables, we find

$$D^2 = 2HR. \tag{20}$$

Example 4.1. *Taking $R = 6.4 \times 10^6$ m and $H = 100$ m. Then $\sqrt{2HR} \approx 36$ km.*

Example 4.2. *Here is some MATLAB code to illustrate the approximation's worth.*

```
R=6.4e6; h = 0:100:100000;
dtrue = R*acos((1 + h/R).^(-1));
dapprox = (2*R*h).^(1/2);
```

5. The Railway problem

This is a very old chestnut indeed. We imagine a straight piece of rail of unit length which, under thermal expansion, becomes a circular arc of length $1+\delta$, where $1 \gg \delta > 0$. The rail will bow upwards, attaining a maximum height h at its centre, and the problem is to determine h, which is surprisingly large.

If we let R denote the radius of the circular arc after expansion, and θ denote the half-angle subtended at the centre of the circle, then we have the equations

$$2R\theta = 1+\delta, \tag{21}$$

$$R - h = R\cos\theta, \tag{22}$$

$$\frac{1}{2} = R\sin\theta. \tag{23}$$

Of course

$$h = R\left(1-\cos\theta\right). \tag{24}$$

Eliminating R from (21) and (23), we obtain

$$\frac{\sin\theta}{2\theta} = \frac{1/2}{1+\delta}$$

or

$$\frac{\sin\theta}{\theta} = \frac{1}{1+\delta}. \tag{25}$$

Now $0 < \delta \ll 1$ implies that θ is also small, so that

$$\frac{\sin\theta}{\theta} = 1 - \frac{1}{6}\theta^2 + \cdots = 1 - \delta + \cdots, \tag{26}$$

or

$$\theta^2 \approx 6\delta. \tag{27}$$

Substituting this approximation in (21) yields

$$R = \frac{1+\delta}{2\theta} \approx \frac{1+\delta}{2\sqrt{6\delta}}. \tag{28}$$

Substituing (28) in (24) then provides

$$h \approx R\theta^2/2 \approx \left(\frac{1+\delta}{2\sqrt{6\delta}}\right)\frac{6\delta}{2} = \frac{\sqrt{6\delta}\,(1+\delta)}{4}$$

i.e.

$$h \approx \sqrt{3\delta/8}. \tag{29}$$

This is at the root of the surprising size of h: $\sqrt{\delta}$ dominates δ for small δ.

Example 5.1. *Suppose $\delta = 10^{-4}$, which corresponds to expansion of 10 cm for a rail of length one kilometre. In this case $h = \sqrt{3\times 10^{-4}/8} = 6.1m$.*

6. A Derivation of the FFT

We choose $n = 2^M$ and illustrate the Fast Fourier Transform algorithm, which computes the DFT in $O(M2^M)$ operations.

Our primary data are the values $\{f(2\pi j/2^M) : j = 01, 2, \ldots, 2^M - 1\}$ of our function evaluated at the 2^M-th roots of one. For each $m \in \{0, 1, \ldots, M-1\}$, we define

$$F_{jk}^{(m)} = \sum_{p=0}^{2^m-1} f\left(e^{2\pi i\left(\frac{p}{2^m}+\frac{k}{2^M}\right)}\right) e^{-2\pi ijp/2^m}. \tag{30}$$

for $j = 0, 1, \ldots, 2^m - 1$ and $k = 0, 1, \ldots, 2^{M-m} - 1$. Thus $F^{(m)} \in \mathbb{C}^{2^m \times 2^{M-m}}$. In other words, each $F^{(m)}$ contains 2^M numbers, but their sizes are as follows:

$$\begin{aligned} F^{(0)} &\text{ is } 1 \times 2^M; \\ F^{(1)} &\text{ is } 2 \times 2^{M-1}; \\ F^{(2)} &\text{ is } 2^2 \times 2^{M-2}; \\ &\vdots \\ F^{(M-1)} &\text{ is } 2^{M-1} \times 2; \\ F^{(M)} &\text{ is } 2^M \times 1. \end{aligned}$$

In other words, $F^{(0)}$ is a row vector, $F^{(m)}$ has twice the number of rows as $F^{(m-1)}$, but half the number of columns, and $F^{(M)}$ is a column vector.

Example 6.1. *When $M = 3$ and $m = 2$, there are 2 4-transforms.*

Example 6.2. *When $M = 3$ and $m = 1$, there are 4 2-transforms.*

We now define a mapping constructing $F^{(m)}$ from $F^{(m-1)}$. Specifically, we divide the sum over p in (30) into even p and odd p, as follows

$$F_{jk}^{(m)} = E_m + O_m, \tag{31}$$

where

$$E_m = \sum_{q=0}^{2^{m-1}-1} f(\exp(2\pi i\left(\frac{q}{2^{m-1}} + \frac{k}{2^N}\right)\exp(-2\pi ijq/2^{m-1}) \tag{32}$$

and

$$O_m = \sum_{r=0}^{2^{m-1}-1} \left(f(\exp(2\pi i\left(\frac{r}{2^{m-1}} + \frac{k+2^{N-m}}{2^N}\right)\exp(-2\pi ijq/2^{m-1})\right)\exp(-\pi ij/2^{m-1}). \tag{33}$$

Now $F^{(m-1)}$ is a $2^{m-1} \times 2^{M-m+1}$ matrix, but it is useful to slightly abuse notation noting that $\mathbb{Z} \ni j \mapsto F_{jk}^{(m-1)}$ is a 2^{m-1}-periodic sequence. With this abuse of notation in mind, we obtain

$$F_{jk}^{(m)} = F_{jk}^{(m-1)} + e^{-\pi ij/2^{m-1}} F_{j,k+2^{M-m}}^{(m-1)}, \tag{34}$$

for $j = 0, 1, \ldots, 2^m - 1$, $k = 0, 1, \ldots, 2^{M-m} - 1$.

7. Constrained Optimization

Suppose we are considering investing money in two assets whose returns are independent random variables X_1 and X_2. Their distribution is unknown, but we do know the mean $\mu_k = \mathbb{E}X_k$ and the variance $\sigma_k^2 = \operatorname{var} X_k$, for $k = 1, 2$, and we shall assume that these variances are strictly positive.

Being risk-averse, we want to divide our investment between the two assets to minimize our risk. More formally, we have

$$Y = s_1 X_1 + s_2 X_2, \quad \text{where } s_1 + s_2 = 1. \tag{35}$$

Now, by the independence of X_1 and X_2, we have

$$\operatorname{var} Y = f(\mathbf{s}) = s_1^2\sigma_1^2 + s_2^2\sigma_2^2, \qquad \mathbf{s} = (s_1, s_2)^T \in \mathbb{R}^2. \tag{36}$$

Thus our problem is as follows:

$$\begin{array}{ll} \text{minimize} & f(\mathbf{s}) \\ \text{subject to} & g(\mathbf{s}) = 1, \end{array} \tag{37}$$

where

$$g(\mathbf{s}) = s_1 + s_2, \quad \mathbf{s} = (s_1, s_2)^T \in \mathbb{R}^2. \tag{38}$$

Now the function $f(\mathbf{s})$ satisfies

$$\nabla f(\mathbf{s}) = \begin{pmatrix} 2\sigma_1^2 s_1 \\ 2\sigma_2^2 s_2 \end{pmatrix}$$

and

$$D^2 f(\mathbf{s}) = \begin{pmatrix} 2\sigma_1^2 & 0 \\ 0 & 2\sigma_2^2 \end{pmatrix},$$

and all higher derivatives vanish. In other words, $f(\mathbf{s})$ is a quadratic and satisfies

$$f(\mathbf{s}+\mathbf{h}) = f(\mathbf{s}) + \mathbf{h}^T\nabla f(\mathbf{s}) + \frac{1}{2}\mathbf{h}^T D^2 f(\mathbf{s})\mathbf{h}. \tag{39}$$

Further, the constraint function $g(\mathbf{s})$ is linear and satisfies

$$g(\mathbf{s}+\mathbf{h}) = g(\mathbf{s}) + \mathbf{h}^T\nabla g(\mathbf{s}) = g(\mathbf{s}) + \mathbf{h}^T\mathbf{e}, \tag{40}$$

where

$$\nabla g(\mathbf{s}) \equiv \mathbf{e} = \begin{pmatrix} 1 \\ 1 \end{pmatrix}. \tag{41}$$

One way to understand such problems is via *line search*: we choose a point $\mathbf{s} \in \mathbb{R}^2$ and a search direction $\mathbf{d} \in \mathbb{R}^2$ and consider the univariate function

$$\phi(t) = f(\mathbf{s} + t\mathbf{d}), \qquad t \in \mathbb{R}. \tag{42}$$

Thus

$$\phi(t) = f(\mathbf{s}) + t\mathbf{d}^T\nabla f(\mathbf{s}) + \frac{1}{2}t^2\mathbf{d}^T D^2 f(\mathbf{s})\mathbf{d}, \tag{43}$$

but we also require the search direction to satisfy the linear constraint:

$$1 = g(\mathbf{s} + t\mathbf{d}) = g(\mathbf{s}) + t\mathbf{d}^T\nabla g(\mathbf{s}) = 1 + t\mathbf{d}^T\nabla g(\mathbf{s}), \tag{44}$$

or

$$\mathbf{d}^T\nabla g(\mathbf{s}) = 0 \tag{45}$$

When do we know we are at a minimum? In this case, we must have $\phi'(0) = 0$ for any $\mathbf{d}$ satisfying (45). Hence

$$\mathbf{d}^T\nabla f(\mathbf{s}) = \mathbf{d}^T\nabla g(\mathbf{s}) = 0, \tag{46}$$

which implies that

$$\nabla f(\mathbf{s}) = \lambda\nabla g(\mathbf{s}), \tag{47}$$

for some $\lambda \in \mathbb{R}$. In other words, we have

$$\begin{pmatrix} 2\sigma_1^2 s_1 \\ 2\sigma_2^2 s_2 \end{pmatrix} = \lambda \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \tag{48}$$

which imply that

$$s_k = \frac{1}{2}\lambda \sigma_k^{-2}, \quad k = 1, 2, \quad \text{and } s_1 + s_2 = 1. \tag{49}$$

Thus

$$\lambda = \frac{2}{\sigma_1^{-2} + \sigma_2^{-2}}$$

and

$$s_k = \frac{\sigma_k^{-2}}{\sigma_1^{-2} + \sigma_2^{-2}}, \quad k = 1, 2. \tag{50}$$

The resulting minimal variance is then given by

$$\begin{aligned} \sigma^2 &\equiv f(\mathbf{s}) \\ &= \sigma_1^2 \frac{\sigma_1^{-4}}{\left(\sigma_1^{-2} + \sigma_2^{-2}\right)^2} + \sigma_2^2 \frac{\sigma_2^{-4}}{\left(\sigma_1^{-2} + \sigma_2^{-2}\right)^2} \\ &= \frac{1}{\sigma_1^{-2} + \sigma_2^{-2}}, \end{aligned}$$

or

$$\sigma^{-2} = \sigma_1^{-2} + \sigma_2^{-2}. \tag{51}$$

Example 7.1. *When $\sigma_1^2 = 1/10$ and $\sigma_2^2 = 1/5$, the minimal variance is given by $\sigma^{-2} = 10 + 5 = 15$, or $\sigma^2 = 1/15$.*

In general, we have

$$Y = s_1 X_1 + s_2 X_2 = \frac{\sigma_1^{-2} X_1 + \sigma_2^{-2} X_2}{\sigma_1^{-2} + \sigma_2^{-2}}$$

and

$$\mathbb{E}Y = s_1 \mu_1 + s_2 \mu_2 = \frac{\sigma_1^{-2} \mu_1 + \sigma_2^{-2} \mu_2}{\sigma_1^{-2} + \sigma_2^{-2}}$$

Thus, if $\sigma_1 \gg \sigma_2$, then $\mathbb{E}Y \approx \mu_2$, which is to be expected, whilst $\sigma_1 = \sigma_2$ implies $\mathbb{E}Y = (\mu_1 + \mu_2)/2$.

7.1. **Lagrange Multipliers.** The above technique is much more general. Suppose we have a risk-metric for investments in n assets which is given by

$$f(\mathbf{s}) = \mathbf{s}^T A \mathbf{s}, \qquad \mathbf{s} \in \mathbb{R}^n, \tag{52}$$

where $A \in \mathbb{R}^{n \times n}$ is a symmetric, positive definite matrix. We want to solve the constrained optimization problem

$$\begin{array}{ll} \text{minimize} & f(\mathbf{s}) \\ \text{subject to} & g(\mathbf{s}) = 1, \end{array} \tag{53}$$

where

$$g(\mathbf{s}) = \mathbf{w}^T \mathbf{s}, \qquad \mathbf{s} \in \mathbb{R}^n, \tag{54}$$

where $\mathbf{w} \in \mathbb{R}^n$ is some fixed vector. Then a similar argument implies that

$$\nabla f(\mathbf{s}) = \lambda \nabla g(\mathbf{s}), \tag{55}$$

where

$$\nabla f(\mathbf{s}) = 2A\mathbf{s} \quad \text{and} \quad \nabla g(\mathbf{s}) = \mathbf{w}. \tag{56}$$

Hence

$$\mathbf{s} = \frac{1}{2}\lambda A^{-1}\mathbf{w} \quad \text{and} \quad 1 = \mathbf{w}^T\mathbf{s}, \tag{57}$$

which implies

$$\lambda = \frac{2}{\mathbf{w}^T A^{-1}\mathbf{w}} \tag{58}$$

and

$$\mathbf{s} = \frac{A^{-1}\mathbf{w}}{\mathbf{w}^T A^{-1}\mathbf{w}}. \tag{59}$$

Exercise 7.1. *Prove that* (59) *implies that the corresponding minimal risk-metric is given by*

$$f(\mathbf{s}) = \mathbf{w}^T A^{-1}\mathbf{w}. \tag{60}$$

8. The Cholesky Factorization

Let $\mathbb{P}_n$ denote the set of all non-negative definite symmetric matrices in $\mathbb{R}^{n\times n}$. Given any $A_n \in \mathbb{P}_n$, there is a unique lower triangular matrix $L_n \in \mathbb{R}^{n\times n}$, with positive diagonal elements, for which $A_n = L_n L^T$, and this is called the Cholesky factorization. This section provides a constructive proof of this result, the factorization being obvious when $n = 1$.

Let us now consider the problem of computing the Cholesky factorization $A_{n+1} = L_{n+1}L_{n+1}^T$, where

$$A_{n+1} = \begin{pmatrix} A_n & \mathbf{a} \\ \mathbf{a}^T & b \end{pmatrix} \in \mathbb{R}^{(n+1)\times(n+1)}, \tag{61}$$

where $\mathbf{a} \in \mathbb{R}^n$, $b \geq 0$ and we assume that we have already computed the Cholesky factorization $A_n = L_n L_n^T$. We define

$$L_{n+1} = \begin{pmatrix} L_n & \mathbf{0} \\ \mathbf{p}^T & q \end{pmatrix}, \tag{62}$$

where $\mathbf{p} \in \mathbb{R}^n$ and $q \geq 0$ are to be determined. Then

$$A_{n+1} = \begin{pmatrix} A_n & \mathbf{a} \\ \mathbf{a}^T & b \end{pmatrix} = \begin{pmatrix} L_n & \mathbf{0} \\ \mathbf{p}^T & q \end{pmatrix} \begin{pmatrix} L_n^T & \mathbf{p} \\ \mathbf{0}^T & q \end{pmatrix}, \tag{63}$$

and $\mathbf{p}$ and q must therefore satisfy the equations

$$L_n \mathbf{p} = \mathbf{a} \tag{64}$$

and

$$\|\mathbf{p}\|^2 + q^2 = b. \tag{65}$$

It is (65) that presents the difficulty: we must prove that

$$b \geq \|\mathbf{p}\|^2 = \|L_n^{-1}\mathbf{a}\|^2 \tag{66}$$

to ensure that $q^2 \geq 0$. To this end, we shall first deal with the simpler case when $A_n = I_n$.

Lemma 8.1. *Let $A_n = I_n$. Then $b \geq \|\mathbf{a}\|^2$.*

Proof. For any $\mathbf{v} \in \mathbb{R}^n$ and $w \in \mathbb{R}$ we have

$$\begin{aligned} 0 &\leq \begin{pmatrix} \mathbf{v} \\ w \end{pmatrix}^T \begin{pmatrix} I_n & \mathbf{a} \\ \mathbf{a}^T & b \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ w \end{pmatrix} \\ &= \mathbf{v}^T\mathbf{v} + 2w\mathbf{v}^T\mathbf{a} + bw^2 \\ &= \|\mathbf{v} + w\mathbf{a}\|^2 + \left(b - \|\mathbf{a}\|^2\right) w^2. \end{aligned}$$

Setting $w = 1$ and $v = -a$, we obtain $0 \leq b - \|\mathbf{a}\|^2$, as desired. □

To extend this result to the original case, we use the following trick to relate the general A_{n+1} to the case where $A_n = I_n$.

$$\begin{aligned} &\begin{pmatrix} L_n^{-1} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \begin{pmatrix} A_n & \mathbf{a} \\ \mathbf{a}^T & b \end{pmatrix} \begin{pmatrix} L_n^{-T} & \mathbf{0} \\ \mathbf{0}^T & 1 \end{pmatrix} \\ &\quad = \begin{pmatrix} L_n^{-1}A_nL_n^{-T} & L_n^{-1}\mathbf{a} \\ \left(L_n^{-1}\mathbf{a}\right)^T & b \end{pmatrix} \\ &\quad = \begin{pmatrix} I_n & L_n^{-1}\mathbf{a} \\ \left(L_n^{-1}\mathbf{a}\right)^T & b \end{pmatrix}. \end{aligned}$$

Hence Lemma 8.1 implies that

$$b \geq \|L_n^{-1}\mathbf{a}\|^2,$$

which is (66), as required.

Students are often be more familiar with the square-root defined by $A^{1/2} = QD^{1/2}A^TQ$, where $A = QDQ^T$ is the spectral factorization of A, rather than the Cholesky factorization $A = LL^T$. Thus $(A^{1/2})^2 = LL^T$, and it can be shown that $L = A^{1/2}W$, where W is an orthogonal matrix. [Essentially the argument is as follows. If we compute the SVD $L = USV^T$, where U and V are orthogonal matrices and S is the diagonal matrix of singular values of L, then $LL^T = (USV^T)(VSU^T) = US^2U^T = A = QDQ^T$. Hence $U = Q$ and $S = D^{1/2}$. Thus $L = QD^{1/2}V^T = A^{1/2}W$, where $W = QV^T$.]

With this in mind, we see that (66) becomes

$$b \geq \|L_n^{-1}\mathbf{a}\|^2 = \|A_n^{-1/2}\mathbf{a}\|^2 = \mathbf{a}^TA_n^{-1}\mathbf{a}. \tag{67}$$

Once we know condition (67), it's possible to remove all of the scaffolding used above, although I believe most readers will find the more circuitous route described above useful: it's often good to leave some scaffolding in place!

Lemma 8.2. *Let $A_n \in \mathbb{R}^{n\times n}$ be any symmetric non-negative definite matrix and define $A_{n+1} \in \mathbb{R}^{(n+1)\times(n+1)}$ by (61). Then A_{n+1} is non-negative definite if and only if*

$$b \geq \mathbf{a}^TA_n^{-1}\mathbf{a}. \tag{68}$$

Proof. For any $\mathbf{v} \in \mathbb{R}^n$ and $w \in \mathbb{R}$ we have

$$\begin{aligned} 0 &\leq \begin{pmatrix} \mathbf{v} \\ w \end{pmatrix}^T \begin{pmatrix} A_n & \mathbf{a} \\ \mathbf{a}^T & b \end{pmatrix} \begin{pmatrix} \mathbf{v} \\ w \end{pmatrix} \\ &= \mathbf{v}^TA_n\mathbf{v} + 2w\mathbf{v}^T\mathbf{a} + bw^2 \\ &= \|A_n^{1/2}\mathbf{v} + wA_n^{-1/2}\mathbf{a}\|^2 + \left(b - \mathbf{a}^TA_n^{-1}\mathbf{a}\right)w^2. \end{aligned}$$

[How did I complete the square here? The key point is that $\mathbf{v}^TA_n\mathbf{v} = (A_n^{1/2}\mathbf{v})^T(A_n^{1/2}\mathbf{v})$, which implies that we must then write the second term as $\mathbf{v}^T\mathbf{a} = (A_n^{1/2}\mathbf{v})^T(A_n^{-1/2}\mathbf{a})$.] If A_{n+1} is non-negative definite, then setting $w = 1$ and $v = -a$ we obtain $0 \leq b - \mathbf{a}^TA_n^{-1}\mathbf{a}$. Conversely, if $b - \mathbf{a}^TA_n^{-1}\mathbf{a} \geq 0$, then A_{n+1} is non-negative definite. □

9. Conics

9.1. **The Ellipse and Hyperbola.** Let's begin with the ellipse and the hyperbola, which we shall define as contours of the functions

$$f_\pm(\mathbf{x}) = \|\mathbf{x}+\mathbf{s}\| \pm \|\mathbf{x}-\mathbf{s}\|, \qquad \mathbf{x} \in \mathbb{R}^n. \tag{69}$$

The key trick is to observe that

$$4\mathbf{x}^T\mathbf{s} = \|\mathbf{x}+\mathbf{s}\|^2 - \|\mathbf{x}-\mathbf{s}\|^2 = f_+(\mathbf{x})f_-(\mathbf{x}). \tag{70}$$

If $f_+(\mathbf{x}) = \alpha$, then $f_-(\mathbf{x}) = 4\mathbf{x}^T\mathbf{s}/\alpha$. Adding these equations gives

$$f_+(\mathbf{x}) + f_-(\mathbf{x}) = 2\|\mathbf{x}+\mathbf{s}\| = \alpha + \frac{4\mathbf{x}^T\mathbf{s}}{\alpha} = \frac{\alpha^2 + 4\mathbf{x}^T\mathbf{s}}{\alpha}, \tag{71}$$

and squaring both sides yields the quadratic form

$$4\|\mathbf{x}+\mathbf{s}\|^2 = \left(\frac{\alpha^2 + 4\mathbf{x}^T\mathbf{s}}{\alpha}\right)^2. \tag{72}$$

The matrix occurring in this quadratic form is

$$M = 4\left(I_n - \frac{4}{\alpha^2}\mathbf{s}\mathbf{s}^T\right). \tag{73}$$

Similarly, if $f_-(\mathbf{x}) = \alpha$, then $f_+(\mathbf{x}) = 4\mathbf{x}^T\mathbf{s}/\alpha$ and adding them yields (72) and (73). The distinction between the contours of $f_\pm$ lies in the eigenvalues of M, which are 1 (with multiplicity $n-1$) and

$$\lambda = 4\left(1 - \frac{4\|\mathbf{s}\|^2}{\alpha^2}\right).$$

If $f_+(\mathbf{x}) = \alpha$, then the triangle inequality implies that

$$\|\mathbf{s}\| \le \frac{1}{2}\left(\|\mathbf{s}+\mathbf{x}\| + \|\mathbf{s}-\mathbf{x}\|\right) = \frac{\alpha}{2},$$

i.e.

$$4\|\mathbf{s}\|^2 \le \alpha^2,$$

which implies $\lambda \ge 0$, with inequality if and only if $\mathbf{x}$ and $\pm\mathbf{s}$ are collinear. Thus M is non-negative definite on contours of f_+ and M is positive definite when the contour is not the line segment joining $\pm\mathbf{s}$.

In contrast, the triangle inequality also implies that

$$\alpha = \|\mathbf{x}+\mathbf{s}\| - \|\mathbf{x}-\mathbf{s}\| \le \|\mathbf{x}+\mathbf{s}-(\mathbf{x}-\mathbf{s})\| = 2\|\mathbf{s}\|,$$

or $4\|\mathbf{s}\|^2/\alpha^2 \ge 1$ on contours of f_-, i.e. $\lambda \le 0$.

9.2. **The Reflector Property.** We have

$$\nabla f_\pm(\mathbf{x}) = \frac{\mathbf{x}+\mathbf{s}}{\|\mathbf{x}+\mathbf{s}\|} \pm \frac{\mathbf{x}-\mathbf{s}}{\|\mathbf{x}-\mathbf{s}\|}. \tag{74}$$

Then

$$\left(\frac{\mathbf{x}+\mathbf{s}}{\|\mathbf{x}+\mathbf{s}\|}\right)^T \nabla f_\pm(\mathbf{x}) = 1 \pm \frac{(\mathbf{x}+\mathbf{s})^T(\mathbf{x}-\mathbf{s})}{\|\mathbf{x}+\mathbf{s}\|\|\mathbf{x}-\mathbf{s}\|}. \tag{75}$$

and

$$\left(\frac{\mathbf{x}-\mathbf{s}}{\|\mathbf{x}-\mathbf{s}\|}\right)^T \nabla f_\pm(\mathbf{x}) = \frac{(\mathbf{x}+\mathbf{s})^T(\mathbf{x}-\mathbf{s})}{\|\mathbf{x}+\mathbf{s}\|\|\mathbf{x}-\mathbf{s}\|} \pm 1. \tag{76}$$

Thus

$$\left(\frac{\mathbf{x}+\mathbf{s}}{\|\mathbf{x}+\mathbf{s}\|}\right)^T \nabla f_\pm(\mathbf{x}) = \pm\left(\frac{\mathbf{x}-\mathbf{s}}{\|\mathbf{x}-\mathbf{s}\|}\right)^T \nabla f_\pm(\mathbf{x}), \tag{77}$$

which is the reflector property for the ellipse and hyperbola.

9.3. **Conic sections really are conic sections.** Let us take a cone of semi-angle θ in $\mathbb{R}^n$ whose axis is the line generated by a unit vector $\mathbf{u} \in \mathbb{R}^n$, i.e.

$$C = \{\mathbf{x} \in \mathbb{R}^n : \frac{\mathbf{u}^T\mathbf{x}}{\|\mathbf{x}\|} = \pm\cos\theta\}. \tag{78}$$

In other words, the equation for the cone is

$$\left(\mathbf{u}^T\mathbf{x}\right)^2 = \|\mathbf{x}\|^2\cos^2\theta, \tag{79}$$

or

$$\mathbf{x}^T\left(I_n\cos^2\theta - \mathbf{u}\mathbf{u}^T\right)\mathbf{x} = 0. \tag{80}$$

Now let $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ be any orthonormal basis for $\mathbb{R}^n$ and consider the hyperplane P with normal vector $\mathbf{v}_n$ at signed distance z_n from the origin. In other words,

$$P = \{\mathbf{x} = \sum_{k=1}^{n} z_k\mathbf{v}_k : z_1, z_2, \ldots, z_{n-1} \in \mathbb{R}\}. \tag{81}$$

If we let $V \in \mathbb{R}^{n\times n}$ be the orthogonal matrix with columns $\mathbf{v}_1, \ldots, \mathbf{v}_n$ and substitute $\mathbf{x} = V\mathbf{z}$ in (79), then we obtain

$$\left(\mathbf{z}^T V^T\mathbf{u}\right)^2 = \|\mathbf{z}\|^2\cos^2\theta. \tag{82}$$

Setting

$$\mathbf{U} = V^T\mathbf{u}, \tag{83}$$

we see that (82) becomes

$$\left(\sum_{k=1}^{n} z_k U_k\right)^2 = \left(\sum_{k=1}^{n} z_k^2\right)\cos^2\theta, \tag{84}$$

or

$$\sum_{k,\ell=1}^{n-1} z_k z_\ell U_k U_\ell + 2z_n U_n\sum_{k=1}^{n-1} z_k U_k + z_n^2 U_n^2 = \left(\sum_{k=1}^{n-1} z_k^2\right)\cos^2\theta + z_n^2\cos^2\theta. \tag{85}$$

Hence, writing

$$\widehat{\mathbf{z}} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_{n-1} \end{pmatrix} \quad \text{and} \quad \widehat{\mathbf{U}} = \begin{pmatrix} U_1 \\ U_2 \\ \vdots \\ U_{n-1} \end{pmatrix}$$

(85) becomes the quadratic form

$$\widehat{\mathbf{z}}^T M\widehat{\mathbf{z}} - 2z_n U_n\widehat{\mathbf{z}}^T\widehat{\mathbf{U}} + z_n^2\left(\cos^2\theta - U_n^2\right) = 0, \tag{86}$$

where the matrix $M \in \mathbb{R}^{(n-1)\times(n-1)}$ is given by

$$M = I_{n-1}\cos^2\theta - \widehat{\mathbf{U}}\widehat{\mathbf{U}}^T. \tag{87}$$

Example 9.1. *Let us choose* $\mathbf{v}_n = \mathbf{u}$*, so that* $\widehat{\mathbf{U}} = 0$ *and* $U_n = 1$*. Then* (86) *becomes*

$$\left(\sum_{k=1}^{n-1} z_k^2\right)\cos^2\theta - z_n^2\sin^2\theta = 0,$$

or

$$\sum_{k=1}^{n-1} z_k^2 = z_n^2\tan^2\theta.$$

The eigenvalues of M are $\cos^2\theta$ (with multiplicity $n-2$) and

$$\mu := \cos^2\theta - \|\widehat{\mathbf{U}}\|^2.$$

Now

$$1 = \|\mathbf{u}\|^2 = \sum_{k=1}^{n} \left(\mathbf{u}^T\mathbf{v}_k\right)^2 = \|\widehat{\mathbf{U}}\|^2 + U_n^2,$$

which implies

$$\mu = \cos^2\theta - \left(1 - U_n^2\right) = U_n^2 - \sin^2\theta. \tag{88}$$

Example 9.2. *Let $n = 3$ and suppose $\mu = 0$. Then*

$$\|\widehat{\mathbf{U}}\|^2 = \cos^2\theta$$

and

$$U_n = \pm\sin\theta.$$

If $\mathbf{q}_1 = \widehat{\mathbf{U}}/\|\widehat{\mathbf{U}}\|$ and $\mathbf{q}_2 \in \mathbb{R}^2$ is orthogonal to $\widehat{\mathbf{U}}$, then the matrix

$$Q = \begin{pmatrix} \mathbf{q}_1 & \mathbf{q}_2 \end{pmatrix} \in \mathbb{R}^{2\times 2}$$

is orthogonal and

$$D := Q^T M Q = \begin{pmatrix} \mu & 0 \\ 0 & \cos^2\theta \end{pmatrix}.$$

If we let $\widehat{\mathbf{z}} = Q\mathbf{x}$, then

$$\cos^2\theta\ x_2^2 \pm 2z_3\sin\theta\cos\theta x_1 + z_3^2\left(\cos^2\theta - \sin^2\theta\right) = 0. \tag{89}$$

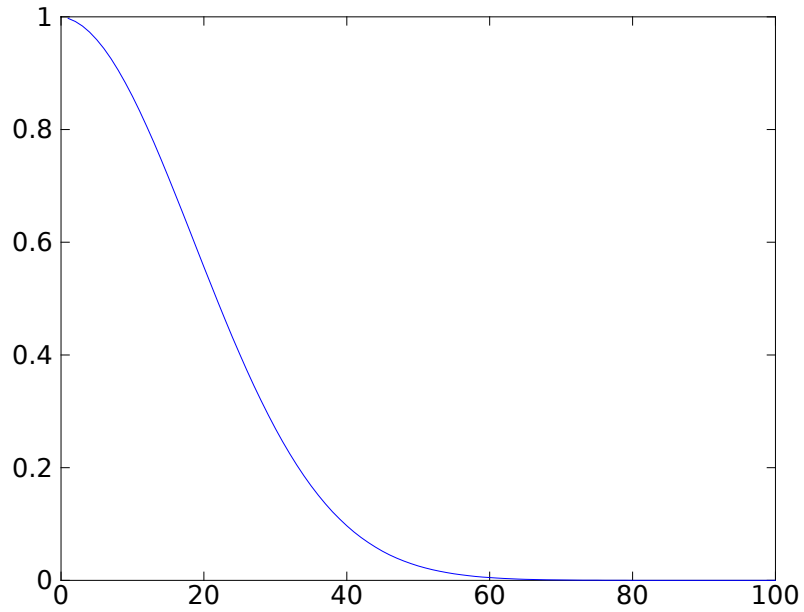

FIGURE 2. The probability that all n birthdays are different

10. The Birthday Problem

This is a traditional probabilistic problem: given n people, whose birthdays are assumed to be uniformly distributed over the $N = 365$ days of the year (ignoring leap years), find the probability that at least two of them share a birthday. Now

$$\mathbb{P}(\text{at least two share a birthday}) = 1 - \mathbb{P}(\text{all birthdays distinct}) =: 1 - p_n. \tag{90}$$

Now

$$p_n = \frac{N(N-1)(N-2)\cdots(N-n+1)}{N^n} \tag{91}$$

and, dividing numerator and denominator by N^n, we obtain

$$p_n = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n-1}{N}\right). \tag{92}$$

In one sense the problem is now solved. The surprise is that p_n tends to zero rather quickly. Indeed, $p_{23} = 0.5073$, by direct calculation. However, plotting p_n reveals a suspiciously Gaussian curve, as we see in Figure 2. Why does p_n decay so quickly and can we understand the seemingly Gaussian behaviour?

```
N=365; n=100; p=1; prob=zeros(1,n);
%
% prob(k) = (1 - 1/N)(1 - 2/N)...(1-k/N)
%         = prob(all k+1 bdays different)
%
for k=1:n
  p=p*(1-k/N);
  prob(k)=p;
end
```

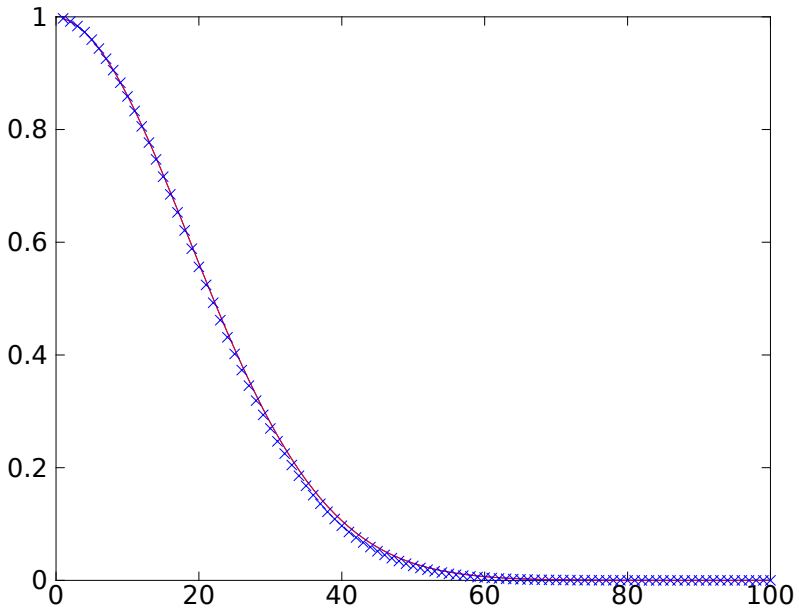

FIGURE 3. x: p_n; line: $\exp(-n(n-1)/(2N))$

First take logarithms:

$$\log p_n = \sum_{k=1}^{n-1} \log\left(1 - \frac{k}{N}\right). \tag{93}$$

Now

$$\log(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots,$$

and the series is convergent for $|x| < 1$. Thus

$$\log(1-x) = -x - \frac{x^2}{2} - \frac{x^3}{3} - \frac{x^4}{4} - \cdots \leq -x, \tag{94}$$

for $0 \leq x < 1$. If $|x| \ll 1$, then we also have the approximation $\log(1+x) \approx -x$. Thus

$$\log p_n \leq -\sum_{k=1}^{n-1} \frac{k}{N} = -\frac{n(n-1)}{2N}, \tag{95}$$

which implies

$$p_n \leq e^{-n(n-1)/(2N)}. \tag{96}$$

This explains the rapid decay and the Gaussian resemblance, as we see in Figure 3.

11. The Bike Problem via the Inclusion–Exclusion Formula

Suppose n cyclists randomly permute their bikes. What is the probability that at least one cyclist has the correct bike?

More formally, the sample space X consits of all positive permutations of the integers $1, 2, \ldots, n$, i.e.

$$X = \{(i_1, i_2, \ldots, i_n) : i_1, \ldots, i_n \text{ a permutation of } 1, \ldots, n\}. \tag{97}$$

We shall assign each of these permutations the same probability $1/n!$.

Further, define

$$A_k = \{x \in X : i_k = k\}, \qquad \text{for } k = 1, 2, \ldots, n. \tag{98}$$

Thus A_k is the set of outcomes for which cyclist k gets bike k. We want to calculate the probability

$$\mathbb{P}\left(A_1 \cup A_2 \cup \cdots \cup A_n\right).$$

Example 11.1. *If $n = 3$, then the sample space is*

$$X = \{(123), (132), (213), (231), (312), (321)\}.$$

Then $A_1 = \{(123), (132)\}$, $A_2 = \{(123), (321)\}$ and $A_3 = \{(123), (213)\}$, whilst $\mathbb{P}(A_1 \cup A_2 \cup A_3) = 4/6 = 2/3$.

The solution requires the inclusion–exclusion formula:

$$\mathbb{P}(A_1 \cup A_2 \cup \cdots \cup A_k) = \sum_{\ell=1}^{n} (-1)^{\ell-1} \sum_{1 \le k_1 < k_2 < \cdots < k_\ell \le n} \mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_\ell}). \tag{99}$$

Exercise 11.1.

$$\mathbb{P}(A_{k_1} \cap \cdots \cap A_{k_m} = \frac{(n-m)!}{n!}.$$

Thus

$$\sum_{1 \le k_1 < k_2 < \cdots < k_\ell \le n} \mathbb{P}(A_{k_1} \cap A_{k_2} \cap \cdots \cap A_{k_\ell}) = \binom{n}{m} \frac{(n-m)!}{n!} = \frac{1}{m!}.$$

Hence

$$\mathbb{P}(A_1 \cup \cdots \cup A_n) = 1 - \frac{1}{2!} + \frac{1}{3!} + \cdots + \frac{(-1)^{n-1}}{n!} \to 1 - e^{-1}. \tag{100}$$

11.1. **The Inclusion–Exclusion Formula.** For each subset A of the sample space X, the *indicator function* $I_A : X \to \{0, 1\}$ is defined by $I_A(x) = 1$ if and only if $x \in A$.

Example 11.2. *For Example 11.1 we have*

$$I_{A_1}(123) = I_{A_1}(132) = 1$$

but

$$I_{A_1}(213) = I_{A_1}(231) = I_{A_1}(312) = I_{A_1}(321) = 0.$$

The indicator function has some crucial properties. Firstly

$$1 - I_A(x) = I_{A^c}(x)$$

where A_c is the complement of A, i.e. $X \setminus A$. Further,

$$I_{A \cap B}(x) = I_A(x) I_B(x).$$

Further, we use de Morgan's Law:

$$(A_1 \cup \cdots \cup A_n)^c = A_1^c \cap \cdots \cap A_n^c.$$

Thus

$$\begin{aligned} I_{A_1 \cup \cdots \cup A_n}(x) &= 1 - I_{(A_1 \cup \cdots \cup A_n)^c}(x) \\ &= 1 - \prod_{k=1}^{n} \left(1 - I_{A_k}(x)\right). \end{aligned}$$

References

[1] Beardon (2005), *Algebra and Geometry*, CUP.