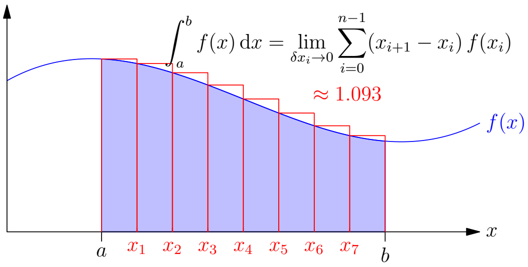
Advanced Machine Learning

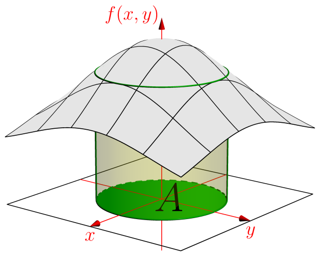
Integral Calculus



Riemann Integration, integration by parts, gaussian integrals

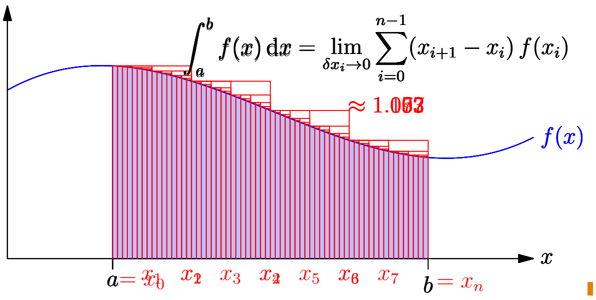
Outline

1. **Defining Integrals**
2. Doing Integrals
3. Gaussian Integrals



Riemann Integral

- Integrals represent area beneath a curve



Linearity of Integration

- Integration is a linear operator

$$\begin{aligned}
\int_a^b \big(r\,f(x) + s\,g(x)\big)\,\mathrm{d}x &= \lim_{\delta x_i \to 0} \sum_{i=0}^{n-1} (x_{i+1} - x_i)\,\big(r\,f(x_i) + s\,g(x_i)\big) \\
&= \lim_{\delta x_i \to 0} \left(\sum_{i=0}^{n-1} (x_{i+1} - x_i)\, r\,f(x_i) + \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, s\,g(x_i) \right) \\
&= \lim_{\delta x_i \to 0} \left(r \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, f(x_i) + s \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, g(x_i) \right) \\
&= r \lim_{\delta x_i \to 0} \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, f(x_i) + s \lim_{\delta x_i \to 0} \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, g(x_i) \\
&= r \int_a^b f(x)\,\mathrm{d}x + s \int_a^b f(x)\,\mathrm{d}x
\end{aligned}$$

Fundamental Law of Calculus

- Let

$$I(a,x) = \int_a^x f(z)\,\mathrm{d}z = \lim_{\delta z_i \to 0} \sum_{i=0}^{n-1} (z_{i+1} - z_i)\, f(z_i)$$

- Now for small δx

$$I(a, x + \delta x) = \int_a^{x+\delta x} f(z)\,\mathrm{d}z = \lim_{\delta z_i \to 0} \sum_{i=0}^{n-1} (z_{i+1} - z_i)\, f(z_i) + \delta x\, f(x)$$

- Thus

$$\frac{\mathrm{d}I(a,x)}{\mathrm{d}x} = \lim_{\delta x \to 0} \frac{I(x + \delta x) - I(x)}{\delta x} = \lim_{\delta x \to 0} \frac{\delta x\, f(x)}{\delta x} = f(x)$$

The Other Way Around

- Consider

$$\begin{aligned}
\int_a^b \frac{\mathrm{d}f(x)}{\mathrm{d}x}\,\mathrm{d}x &= \int_a^b \lim_{\delta x \to 0} \frac{f(x + \delta x) - f(x)}{\delta x}\,\mathrm{d}x \\
&= \lim_{x_{i+1} - x_i \to 0} \sum_{i=0}^{n-1} (x_{i+1} - x_i)\, \frac{f(x_{i+1}) - f(x_i)}{x_{i+1} - x_i} \\
&= \lim_{x_{i+1} - x_i \to 0} \sum_{i=0}^{n-1} \big(f(x_{i+1}) - f(x_i)\big) \\
&= \big(f(x_1) - f(x_0)\big) + \big(f(x_2) - f(x_1)\big) + \big(f(x_3) - f(x_2)\big) + \cdots \\
&\quad + \big(f(x_{n-1}) - f(x_{n-2})\big) + \big(f(x_n) - f(x_{n-1})\big) \\
&= f(x_n) - f(x_0) = f(b) - f(a)
\end{aligned}$$

- We can think of integration as an **anti-derivative** it undoes differentiation

Indefinite Integrals

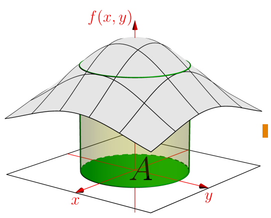
- So far we have considered **definite integrals** where we integrate between two points (a and b)
- However, when think about integration as an anti-derivative, it is useful to think of a function $F(x) = \int f(x)\,\mathrm{d}x$
- So that $F'(x) = f(x)$
- However the function $F(x)$, $F(x) + 1$, $F(x) + \pi$, etc. all have the same derivative so $F(x)$ is only defined up to an additive constant
- Note that the definite integral is given by

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a)$$

Multiple Integrals

- For functions involving many independent variables (e.g. $f(x,y)$, $f(x,y,z)$, $f(\boldsymbol{x})$) we can integrate over multiple dimensions
- For example

$$\iint_A f(x,y)\,\mathrm{d}x\,\mathrm{d}y$$

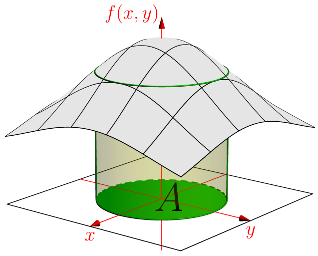


- It gets tedious writing multiple integral signs and I tend to write just one

$$\int \cdots \int f(x_1, x_2, \ldots, x_n)\,\mathrm{d}x_1\,\mathrm{d}x_2 \cdots \mathrm{d}x_n = \int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$

Outline

1. Defining Integrals
2. **Doing Integrals**
3. Gaussian Integrals



Performing Integration

- A key method for performing integrals is through knowledge of the anti-derivative
- If we know $F'(x) = f(x)$ then $F(x) + c = \int f(x)\mathrm{d}x$
- E.g. we know that $\mathrm{d}x^n/\mathrm{d}x = n x^{n-1}$ therefore

$$\int x^{n-1}\mathrm{d}x = \frac{1}{n}\int \frac{\mathrm{d}x^n}{\mathrm{d}x}\mathrm{d}x = \frac{x^n}{n} + c$$

and

$$\int_a^b x^{n-1}\mathrm{d}x = \frac{b^n}{n} - \frac{a^n}{n}$$

Is Integration Straightforward?

- We saw due to the product and chain rules that we can differentiate almost anything. Given integration is the anti-derivative can we integrate anything?
- Products and compositions

$$\int f(x)g(x)\mathrm{d}x = ? \qquad \int f(g(x))\mathrm{d}x = ?$$

- Unfortunately, unlike differentiation we don't have a small parameter we can expand in
- In general integration is hard

Integration by Parts

- Recall the product rule $\frac{\mathrm{d}f(x)g(x)}{\mathrm{d}x} = \frac{\mathrm{d}f(x)}{\mathrm{d}x}g(x) + f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x}$
- Integrating we get

$$\int_a^b \frac{\mathrm{d}f(x)g(x)}{\mathrm{d}x}\mathrm{d}x = \int_a^b \frac{\mathrm{d}f(x)}{\mathrm{d}x}g(x)\mathrm{d}x + \int_a^b f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x}\mathrm{d}x$$
$$= [f(x)g(x)]_a^b = f(b)g(b) - f(a)g(a)$$

- Unfortunately we get two integrals, but we can turn this around

$$\int_a^b f(x)\frac{\mathrm{d}g(x)}{\mathrm{d}x}\mathrm{d}x = [f(x)g(x)]_a^b - \int_a^b \frac{\mathrm{d}f(x)}{\mathrm{d}x}g(x)\mathrm{d}x$$

whether this is helpful depends on $f(x)$ and $g(x)$

Example of Integration by Parts

- Consider

$$\Pi(z) = \int_0^\infty x^z \mathrm{e}^{-x}\mathrm{d}x = \int_0^\infty x^z \frac{\mathrm{d}(-\mathrm{e}^{-x})}{\mathrm{d}x}\mathrm{d}x$$
$$= \left[x^z(-\mathrm{e}^{-x})\right]_0^\infty - \int_0^\infty \frac{\mathrm{d}x^z}{\mathrm{d}x}(-\mathrm{e}^{-x})\mathrm{d}x$$
$$= \int_0^\infty (z x^{z-1})\mathrm{e}^{-x}\mathrm{d}x = z\int_0^\infty x^{z-1}\mathrm{e}^{-x}\mathrm{d}x = z\Pi(z-1)$$

- Thus $\Pi(z) = z\Pi(z-1)$, but

$$\Pi(0) = \int_0^\infty \mathrm{e}^{-z}\mathrm{d}z = \left[-\mathrm{e}^{-x}\right]_0^\infty = -\mathrm{e}^{-\infty} - (-\mathrm{e}^0) = 1$$

- Now
$\Pi(n) = n\Pi(n-1) = n(n-1)\Pi(n-2) = n(n-1)(n-2)\ldots 1 = n!$

Substitution

- We can make a transformation from x to $u = u(x)$

$$\int_a^b f(x)\mathrm{d}x = \lim_{\delta x_i \to 0}\sum_{i=0}^{n-1} f(x_i)(x_{i+1} - x_i)$$
$$= \lim_{\delta u_i \to 0}\sum_{i=0}^{n-1} f(x(u_i))\frac{x(u_{i+1}) - x(u_i)}{u_{i+1} - u_i}(u_{i+1} - u_i)$$
$$= \int_{u(a)}^{u(b)} f(x(u))\frac{\mathrm{d}x(u)}{\mathrm{d}u}\mathrm{d}u$$

- ★ where u_i is such that $x(u_i) = x_i$ or $u_i = u(x_i)$ where $u(x)$ is the inverse of $x(u)$
- ★ using $\lim_{\delta u_i \to 0}\frac{x(u_{i+1}) - x(u_i)}{u_{i+1} - u_i} = \frac{\mathrm{d}x(u_i)}{\mathrm{d}u}$

Example of Integration by Substitution

- We consider $I(n) = \int\limits_0^\infty x^n \mathrm{e}^{-x^2/2}\mathrm{d}x$
- Let $u(x) = x^2/2$ or $x(u) = \sqrt{2u}$ so that

$$\frac{\mathrm{d}x(u)}{\mathrm{d}u} = \frac{1}{\sqrt{2u}} \qquad u(0) = 0 \qquad u(\infty) = \infty$$

- Thus

$$I(n) = \int_0^\infty \left(\sqrt{2u}\right)^n \mathrm{e}^{-u}\frac{1}{\sqrt{2u}}\mathrm{d}u$$
$$= 2^{\frac{n-1}{2}}\int_0^\infty u^{\frac{n-1}{2}}\mathrm{e}^{-u}\mathrm{d}u = 2^{\frac{n-1}{2}}\,\Pi\left(\frac{n-1}{2}\right)$$

- $I(1) = 1$, $I(3) = 2 \times 1! = 2$, $I(5) = 2^2 \times 2! = 8$, but $I(0) = \Pi(-1/2)/\sqrt{2}$, $I(2) = \sqrt{2}\Pi(1/2) = \Pi(-1/2)/\sqrt{2}$

Changing Variables in Multidimensional Space

- When changing variables in many dimensions $\boldsymbol{x} \to \boldsymbol{u}$ the change of variables involves the Jacobian

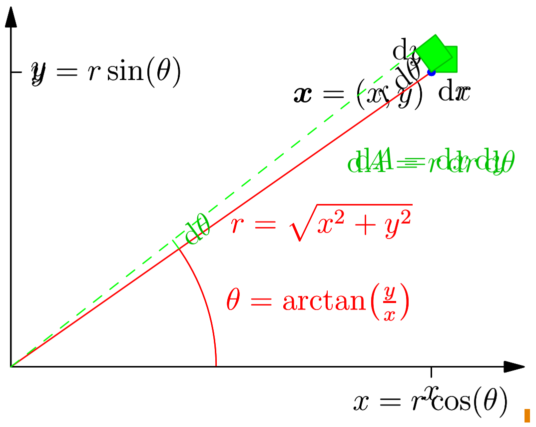
$$\int f(\boldsymbol{x})\mathrm{d}\boldsymbol{x} = \int f(\boldsymbol{x}(\boldsymbol{u}))|\det(\mathbf{J})|\mathrm{d}\boldsymbol{u}, \qquad \boldsymbol{J} = \begin{pmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_1}{\partial u_2} & \cdots & \frac{\partial x_1}{\partial u_n} \\ \frac{\partial x_1}{\partial u_2} & \frac{\partial x_2}{\partial u_2} & \cdots & \frac{\partial x_2}{\partial u_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial x_n}{\partial u_1} & \frac{\partial x_n}{\partial u_2} & \cdots & \frac{\partial x_n}{\partial u_n} \end{pmatrix}$$

- E.g. transforming from Cartesian coordinates (x,y) to polar coordinates (r,θ) then $x = r\cos(\theta)$ and $y = r\sin(\theta)$

$$|\det(\mathbf{J})| = \left|\det\begin{pmatrix} \frac{\partial r\cos(\theta)}{\partial r} & \frac{\partial r\cos(\theta)}{\partial \theta} \\ \frac{\partial r\sin(\theta)}{\partial r} & \frac{\partial r\sin(\theta)}{\partial \theta} \end{pmatrix}\right| = \left|\det\begin{pmatrix} \cos(\theta) & -r\sin(\theta) \\ \sin(\theta) & r\cos(\theta) \end{pmatrix}\right|$$
$$= r\left(\cos^2(\theta) + \sin^2(\theta)\right) = r$$

- That is, $\mathrm{d}x\mathrm{d}y = r\,\mathrm{d}r\mathrm{d}\theta$

Change of Variables in Pictures



Differentiating Through the Integral

- A trick that sometimes works is differentiating through an integral, e.g. consider finding moments

$$M_n = \mathbb{E}[X^n] = \int_{-\infty}^{\infty} x^n f_X(x)\,\mathrm{d}x$$

- We can define a momentum generating function

$$Z(\ell) = \int_{-\infty}^{\infty} \mathrm{e}^{\ell x} f_X(x)\,\mathrm{d}x$$

- Then $M_n = Z^{(n)}(0)$

$$\left.\frac{\mathrm{d}^n Z(\ell)}{\mathrm{d}\ell^n}\right|_{\ell=0} = \int_{-\infty}^{\infty} \left.\frac{\mathrm{d}^n \mathrm{e}^{\ell x}}{\mathrm{d}\ell^n}\right|_{\ell=0} f_X(x)\,\mathrm{d}x = \int_{-\infty}^{\infty} x^n f_X(x)\,\mathrm{d}x = M_n$$

Cumulant Generating Function

- Note that $\mathrm{e}^{\ell x} = 1 + \ell x + \frac{1}{2}\ell^2 x^2 + \frac{1}{3!}\ell^3 x^3 + \cdots$
- So

$$Z(\ell) = \int_{-\infty}^{\infty} \mathrm{e}^{\ell x} f_X(x)\,\mathrm{d}x = 1 + \ell M_1 + \frac{1}{2}\ell^2 M_2 + \frac{1}{3!}\ell^3 M_3 + \cdots$$

- Now using $\log(1+\epsilon) = \epsilon - \frac{1}{2}\epsilon^2 + \frac{1}{3}\epsilon^3 + \cdots$

$$G(\ell) = \log(Z(\ell)) = \ell M_1 + \frac{1}{2}\ell^2 (M_2 - M_1^2) + \frac{1}{3!}\ell^3 (M_3 - 3M_2M_1 + 2M_1^3) + \cdots$$

- So that $\kappa_n = G^{(n)}(0)$, with $\kappa_1 = M_1$ (the mean), $\kappa_2 = M_2 - M_1^2$ (the variance), $\kappa_3 = M_3 - 3M_2M_1 + 2M_1^3$ (the third cumulant related to the skewness)

More Integration

- Although we have a few tricks, integration is hard
- Surprisingly integration sometimes is easier when carried out in the complex plane
- This is a beautiful part of mathematics (due largely to Cauchy)—but beyond the scope of this course
- Interestingly, also there is an algorithm that allows us to integrate a lot of function. It is sufficiently complicated that you need to write a computer algorithm of considerable complexity to implement it. Most symbolic manipulation packages (e.g. Mathematica) have implemented some part of this algorithm

Special Functions

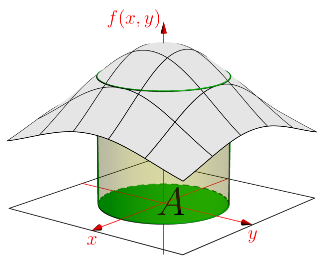
- There are integrals with no known closed form solution
- We saw that $\Pi(z) = \int\limits_0^\infty x^z \mathrm{e}^{-x}\mathrm{d}x$ satisfies $\Pi(z) = z\Pi(z-1)$
- For integer n then $\Pi(n) = n!$, but for general z, the integal $\Pi(z)$ can't be written in terms of elementary functions
- We consider $\Pi(z)$ as a special function in its own right
- Although, history has left us with the gamma function instead

$$\Gamma(z) = \int_0^\infty x^{z-1}\mathrm{e}^{-x}\mathrm{d}x = \Pi(z-1)$$

- Other special function defined by integrals exist (e.g. the Bessel , Aire, hypergeometric, elliptic, error functions, ...)

Outline

1. Defining Integrals
2. Doing Integrals
3. **Gaussian Integrals**



Gaussian Integrals

- Gaussian integrals are integrals involving e^{-x^2}, e.g.

$$\int_{-\infty}^{\infty} \mathrm{e}^{-x^2}\mathrm{d}x \qquad\qquad \int_{-\infty}^{\infty} x^4 \mathrm{e}^{-ax^2 - bx}\mathrm{d}x$$

- They are important in computing integrals with respect to the normal distribution

$$\mathcal{N}(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$$

- The great news is that these integrals are all doable
- The bad news is that they are quite tricky to do

The Gaussian Integral

- The integral over a Gaussian is surprisingly difficult

$$I_1 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\mathrm{d}x$$

- There is a nice trick which is to consider

$$I_1^2 = \int_{-\infty}^{\infty} \mathrm{e}^{-x^2/2}\mathrm{d}x \int_{-\infty}^{\infty} \mathrm{e}^{-y^2/2}\mathrm{d}y = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \mathrm{e}^{-(x^2+y^2)/2}\mathrm{d}x\mathrm{d}y$$

- Making the change of variables $r = \sqrt{x^2+y^2}$ and $\theta = \arctan(y/x)$ (so that $x = r\cos(\theta)$, $y = r\sin(\theta)$ and $x^2 + y^2 = r^2$)

$$I_1^2 = \int_0^{2\pi} \mathrm{d}\theta \int_0^\infty r e^{-r^2/2}\mathrm{d}r = 2\pi \int_0^\infty r e^{-r^2/2}\mathrm{d}r$$

The Gaussian Integral Continued

- From before

$$I_1^2 = 2\pi \int_0^\infty r e^{-r^2/2} \mathrm{d}r$$

- Finally let $u = r^2/2$ so that $\mathrm{d}u/\mathrm{d}r = r$ or $\mathrm{d}u = r\mathrm{d}r$ we get

$$I_1^2 = 2\pi \int_0^\infty e^{-u} \mathrm{d}u = 2\pi$$

- So that $I_1 = \sqrt{2\pi}$
- Incidentally, $I_1 = \sqrt{2}\Pi(-1/2)$ so $\Pi(-1/2) = \Gamma(1/2) = \sqrt{\pi}$

Normal Distribution

- We consider

$$I_2 = \int_{-\infty}^{\infty} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)} \mathrm{d}x$$

- Making the change of variables $z = (x-\mu)/\sigma$ so that $\mathrm{d}z = \mathrm{d}x/\sigma$ or $\mathrm{d}x = \sigma \mathrm{d}z$. Then

$$I_2 = \sigma \int_{-\infty}^{\infty} \mathrm{e}^{-z^2/2} \mathrm{d}z = \sigma I_1 = \sqrt{2\pi}\sigma$$

- Note that the *probability density function* (PDF) for a normally distributed random variable is given by

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-(x-\mu)^2/(2\sigma^2)}$$

Multi-dimensional Gaussians

- Consider

$$I_3 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\|\boldsymbol{x}\|_2^2} \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

where $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)^\mathsf{T}$

- Note that $\|\boldsymbol{x}\|_2^2 = x_1^2 + x_2^2 + \cdots + x_n^2$ and using $\mathrm{e}^{\sum_i a_i} = \prod_i \mathrm{e}^{a_i}$

$$\begin{aligned} I_3 &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\sum_{i=1}^{n} x_i^2} \mathrm{d}x_1 \cdots \mathrm{d}x_n \\ &= \prod_{i=1}^{n} \int_{-\infty}^{\infty} \mathrm{e}^{-x_i^2/2} \mathrm{d}x_i = \prod_{i=1}^{n} \sqrt{2\pi} = (2\pi)^{n/2} \end{aligned}$$

Full Multi-variate Normal

- Consider

$$I_4 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T} \boldsymbol{\Xi}^{-1} (\boldsymbol{x}-\boldsymbol{\mu})} \mathrm{d}x_1 \cdots \mathrm{d}x_n$$

- Let $\boldsymbol{\Xi}^{-1} = \mathbf{V}\boldsymbol{\Lambda}^{-1}\mathbf{V}^\mathsf{T}$ and make the change of variables $\boldsymbol{y} = \mathbf{V}^\mathsf{T}(\boldsymbol{x} - \boldsymbol{\mu})$
- The Jacobian $\mathbf{J}$ has elements (note that $\boldsymbol{x} = \mathbf{V}\boldsymbol{y} + \boldsymbol{\mu}$)

$$J_{ij} = \frac{\partial x_i}{\partial y_j} = \frac{\partial}{\partial y_j}\left(\sum_{k=1}^{n} V_{ik} y_k + \mu_i\right) = V_{ij}$$

- So that $\mathbf{J} = \mathbf{V}$ and consequently $|\det(\mathbf{J})| = |\det(\mathbf{V})| = 1$ then

$$I_4 = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \mathrm{e}^{-\frac{1}{2}\boldsymbol{y}^\mathsf{T}\boldsymbol{\Lambda}^{-1}\boldsymbol{y}} \mathrm{d}y_1 \cdots \mathrm{d}y_n. = \prod_{i=1}^{n} \int_{-\infty}^{\infty} \mathrm{e}^{-y_i^2/(2\lambda_i)} \mathrm{d}y_i = \prod_i \sqrt{2\pi\lambda_i}$$

Determinants

- Using the facts, that $\boldsymbol{\Xi} = \mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}$ then

$$\det(\boldsymbol{\Xi}) = \det(\mathbf{V}\boldsymbol{\Lambda}\mathbf{V}^\mathsf{T}) = \det(\mathbf{V})\det(\boldsymbol{\Lambda})\det(\mathbf{V}^\mathsf{T}) = \det(\boldsymbol{\Lambda}) = \prod_{i=1}^{n} \lambda_i$$

using $\det(\mathbf{A}\mathbf{B}) = \det(\mathbf{A})\det(\mathbf{B})$ and $\det(\mathbf{V}) = 1$

- Recall $I_4 = \prod_i \sqrt{2\pi\lambda_i} = (2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Xi})}$
- We note for an $n \times n$ matrix $\mathbf{M}$ then $\det(c\mathbf{M}) = c^n \det(\mathbf{M})$ so that

$$I_4 = (2\pi)^{n/2}\sqrt{\det(\boldsymbol{\Xi})} = \sqrt{\det(2\pi\boldsymbol{\Xi})}$$

- Finally, we get that for the PDF of a normal to integrate to 1

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Xi}) = \frac{1}{\sqrt{\det(2\pi\boldsymbol{\Xi})}} \mathrm{e}^{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\mathsf{T}\boldsymbol{\Xi}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})}$$

Summary

- Integration is extra-ordinarily useful as a tool of analysis
- It occurs when you work with probabilities densities for continuous random variables
- Integration is beautiful, but hard—often impossible
- Normal distributions lucky almost always give raise to integrals that can be computed in closed form, although often it requires quite a bit of work
- Making friends with integration will give you a super-power that not too many people share