Appendix C

Extensions

From Continued Fractions to Liouville Numbers

Continued fractions, CF are of two kinds, simple SCF and generalized GCF.

$$x = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \cfrac{1}{\ddots + \cfrac{1}{a_n}}}} \qquad y = a_0 + \cfrac{b_1}{a_1 + \cfrac{b_2}{a_2 + \cfrac{b_3}{\ddots + \cfrac{b_n}{a_n}}}}$$

Simple continued fraction with $a_0 \in \mathbb{Z}$ and $a_i \in \mathbb{Z}^+$ for $i \geq 1$.

Generalized continued fraction where $a_i, b_i \in \mathbb{R}$ or $\mathbb{C}$ depending on context.

More convenient notations for continued fractions are as follows:

$$x = [a_0; a_1, a_2, \dots, a_n] \qquad y = a_0 + \frac{b_1}{a_1 +}\frac{b_2}{a_2 +}\cdots\frac{b_n}{a_n} = a_0 + \underset{i=1}{\overset{n}{\mathrm{K}}}\frac{b_i}{a_i}$$

Theorem 1. $[a_0; a_1, \dots, a_n] \in \mathbb{Q}$ and each $x \in \mathbb{Q}$ has two representations, $x = [a_0; a_1, \dots, a_{n-1}, a_n]$ and $x = [a_0; a_1 \dots a_{n-1}, a_n - 1, 1]$ and no others.

Proof. The first statement follows from calculating fractions from the bottom and up. By applying the Euclidean algorithm to $x = p/q$ we get.

$$\frac{p}{q} = \underbrace{\left\lfloor\frac{p}{q}\right\rfloor}_{a_0} + \frac{p_1}{q_1} \qquad 0 \leq p_1 < q_1 \qquad \frac{p}{q} = a_0 + \frac{1}{q_1/p_1}$$

$$q_1 = a_1 p_1 + r_1 \qquad 0 \leq r_1 < p_1 \qquad = a_0 + \frac{1}{a_1 + r_1/p_1}$$

$$p_1 = a_2 r_1 + r_2 \qquad 0 \leq r_2 < r_1 \qquad = a_0 + \frac{1}{a_1 + \frac{1}{a_2 + r_2/r_1}}$$

$$\vdots \qquad \vdots \qquad \vdots$$

$$r_{n-2} = a_n r_{n-1} + r_n \qquad r_n = 0 \qquad = a_0 + \frac{1}{a_1 + \frac{1}{\ddots + \frac{1}{a_n}}} = [a_0; a_1, \dots, a_n]$$

Rewriting $\frac{1}{a_n} = \frac{1}{(a_n - 1) + 1/1}$ gives the second representation.
The requirement $a_i \in \mathbb{Z}^+$ means $x = a_0 + d$ with $d \in [0,1]$
$d \in (0,1)$ fixates a_0 to $\lfloor x \rfloor$.
If $d \in \{0,1\}$ then $x = [a_0]$ or $x = [a_0 - 1; 1]$ with $a_0 = \lfloor x \rfloor$
The same argument fixates a_i at all levels but the last which has two options.

If we always choose the shorter representation then every $x \in \mathbb{Q}$ has a unique representation $[a_0; a_1, a_2, \dots, a_n]$ where $a_n > 1$ if $n > 0$.

To make sense of infinite CF we need to introduce truncated c.f. They are called convergents and as the name suggests, they will converge to their CF

Definition 1. The nth **convergent** $c_n \in \mathbb{Q}$ of $[a_0; a_1, \dots]$ (finite or infinite) is defined by $c_n \equiv [a_0; a_1, \dots, a_n]$ where $n \in \{0,1,\dots\}$

Definition 2. Two sequences (p_n) *and* (q_n) for $[a_0; a_1, \dots]$ are defined by the **fundamental recurrence relations**:

$$\begin{cases} p_n = a_n p_{n-1} + p_{n-2} & p_{-1} = 1 \text{ and } p_{-2} = 0 \\ q_n = a_n q_{n-1} + q_{n-2} & q_{-1} = 0 \text{ and } q_{-2} = 1 \end{cases}$$

Theorem 2. $c_n = p_n / q_n$

Proof. (By induction over n)

$n = 0$: $\dfrac{p_0}{q_0} = \dfrac{a_0 p_{-1} + p_{-2}}{a_0 q_{-1} + q_{-2}} = \dfrac{a_0}{1} = c_0$

Assume the theorem is true for all $k \leq n$.

$$\begin{aligned} c_{n+1} = [a_0; \dots, a_{n+1}] &= [a_0; \dots, a_{n-1}, a_n + 1/a_{n+1}] \\ &= \frac{(a_n + 1/a_{n+1}) p_{n-1} + p_{n-2}}{(a_n + 1/a_{n+1}) q_{n-1} + q_{n-2}} \\ &= \frac{a_{n+1}(a_n p_{n-1} + p_{n-2}) + p_{n-1}}{a_{n+1}(a_n q_{n-1} + q_{n-2}) + q_{n-1}} \\ &= \frac{a_{n+1} p_n + p_{n-1}}{a_{n+1} q_n + q_{n-1}} \\ &= \frac{p_{n+1}}{q_{n+1}} \end{aligned}$$

■

Lemma 1. $p_n q_{n-1} - p_{n-1} q_n = (-1)^{n-1}$ $\left(\text{Same as: } c_n - c_{n-1} = \frac{(-1)^{n-1}}{q_n q_{n-1}}\right)$

Proof. (By induction over n)

$n = -1$: $p_{-1} q_{-2} - p_{-2} q_{-1} = 1 \cdot 1 - 0 \cdot 0 = (-1)^{-2}$

Assume the lemma is true for all $k \leq n$.

$$\begin{aligned} p_{n+1} q_n - p_n q_{n+1} &= (a_{n+1} p_n + p_{n-1}) q_n - p_n (a_{n+1} q_n + q_{n-1}) \\ &= -(p_n q_{n-1} - p_{n-1} q_n) \\ &= (-1)^n \end{aligned}$$

■

Lemma 2. $p_n q_{n-2} - p_{n-2} q_n = (-1)^n a_n$ $\left(\text{Same as: } c_n - c_{n-2} = \frac{(-1)^n a_n}{q_n q_{n-2}}\right)$

Proof. (By induction over n)

$n = 0$: $\quad p_0 q_{-2} - p_{-2} q_0 = (-1)^0 a_0$

Assume the lemma is true for all $k \le n$.

$$\begin{aligned} p_{n+1}q_{n-1} - p_{n-1}q_{n+1} &= (a_{n+1}p_n + p_{n-1})q_{n-1} - p_{n-1}(a_{n+1}q_n + q_{n-1}) \\ &= a_{n+1}(p_n q_{n-1} - p_{n-1} q_n) \\ &= (-1)^{n+1} a_{n+1} \qquad \text{(By Lemma 1)} \end{aligned}$$ ■

Lemma 1 says that $Ap_n + Bq_n = \pm 1$ with $A, B \in \mathbb{Z}$. A basic theorem from number theory implies that $(p_n, q_n) = 1$ (no common factor).

The recurrence relations imply (p_n) and (q_n) are strictly increasing series. By the lemmas the convergents satisfy $c_1 > c_3 > c_5 > \cdots > c_4 > c_2 > c_0$.

The easiest way to calculate c_n for $[a_0; a_1, a_2, ...]$ is by using matrices. Let $A_n = \begin{pmatrix} a_n & 1 \\ 1 & 0 \end{pmatrix}$ $n \ge 0$ and $B_n = A_n A_{n-1} \dots A_1 A_0$ then

$$B_n = \begin{pmatrix} p_n & q_n \\ p_{n-1} & q_{n-1} \end{pmatrix}$$

proof. (By induction over n)

$n = 0$: $B_0 = A_0 = \begin{pmatrix} a_0 & 1 \\ 1 & 0 \end{pmatrix} = \begin{pmatrix} p_0 & q_0 \\ p_{-1} & q_{-1} \end{pmatrix}$

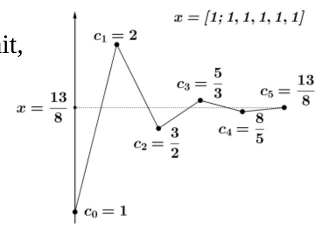
Assume statement true for all $k \le n$.

$$B_{n+1} = \begin{pmatrix} a_{n+1} & 1 \\ 1 & 0 \end{pmatrix} \begin{pmatrix} p_n & q_n \\ p_{n-1} & q_{n-1} \end{pmatrix} = \begin{pmatrix} a_{n+1}p_n + p_{n-1} & a_{n+1}q_n + q_{n-1} \\ p_n & q_n \end{pmatrix} = \begin{pmatrix} p_{n+1} & q_{n+1} \\ p_n & q_n \end{pmatrix}$$ ■

The difference between representing a rational number in a decimal system or as a continued fraction can be quite substantial. If we use the solution from the end of chapter one we will see a large reduction in the number of digits.

$$A = \frac{100\,000\,000\,000}{101\,000\,001\,001} = 0.\underbrace{\overline{990099000088226\ldots xyz}}_{25\,014\,018\,913 \text{ digits}} = [0; 1,99,1,9989,50,40,50]$$

Convergents will oscillate towards their limit, even convergents strictly decreasing and odd convergents strictly increasing.



Even convergents c_{2n} of $[a_0; a_1, \ldots]$ are strictly increasing and limited from above by c_1 so they have a limit L. Odd convergents c_{2n+1} are strictly decreasing and limited from below by c_0 so they have a limit U. These limits coincide since $|c_{2n+1} - c_{2n}| = 1/(q_{2n+1}q_{2n})$ and $q_n \to \infty$ as $n \to \infty$. $[a_0; a_1, \ldots]$, $a_0 \in \mathbb{Z}$ and $a_i \in \mathbb{Z}^+$ for $i \geq 1$ is therefore well-defined:

Definition 3. $[a_0; a_1, \ldots] \equiv \lim_{n\to\infty} [a_0; a_1, \ldots, a_n]$ $(a_0 \in \mathbb{Z}, a_{i>0} \in \mathbb{Z}^+)$ SCF

Theorem 3. Every infinite continued fraction, ICF $[a_0; a_1, \ldots] \in \mathbb{R} \setminus \mathbb{Q}$ and every $x \in \mathbb{R} \setminus \mathbb{Q}$ can be written as $[a_0; a_1, \ldots]$ in a unique way with

$$a_k = \lfloor x_k \rfloor \text{ where } x_0 = x \text{ and } x_{k+1} = \frac{1}{x_k - a_k} \text{ for } k \geq 1$$

Proof.
Part 1, let $x = [a_0; a_1, \ldots]$, suppose $x = p/q$ with $p, q \in \mathbb{Z}$ and $q \neq 0$.

$$\begin{aligned}
c_{2k} <& \quad p/q && < c_{2k+1} \\
0 <& \quad p/q - c_{2k} && < c_{2k+1} - c_{2k} \\
0 <& \quad \frac{p}{q} - \frac{p_{2k}}{q_{2k}} && < \frac{1}{q_{2k+1}q_{2k}} \\
0 <& \quad \underbrace{pq_{2k} - p_{2k}q}_{\in\mathbb{Z}} && < \frac{q}{q_{2k+1}} \quad (\to 0 \text{ as k} \to \infty)
\end{aligned}$$

This is a contradiction so x can't be a rational number, $x \in \mathbb{R} \setminus \mathbb{Q}$.

Part 2, Given $x \in \mathbb{R} \setminus \mathbb{Q}$ and a_k as in theorem for $k \geq 0$.

$x_0 \in \mathbb{R} \setminus \mathbb{Q}$ and $x_k \in \mathbb{R} \setminus \mathbb{Q} \Rightarrow x_{k+1} \in \mathbb{R} \setminus \mathbb{Q}$ (since $x_{k+1} \in \mathbb{Q} \Rightarrow x_k \in \mathbb{Q}$)
so $x_k \in \mathbb{R} \setminus \mathbb{Q}$ for all $k \geq 0$

$a_k \in \mathbb{Z}$, $a_k < x_k < a_k + 1 \Rightarrow x_{k+1} = \frac{1}{x_k - a_k} > 1 \Rightarrow$
$a_{k+1} = \lfloor x_{k+1} \rfloor \geq 1 \Rightarrow a_{k+1} \in \mathbb{Z}^+$ so $a_0 \in \mathbb{Z}$ and $a_k \in \mathbb{Z}^+$

Show $\lim_{k\to\infty} [a_0; a_1, \ldots, a_k] = x$ by studying $\left| x - \frac{p_k}{q_k} \right|$

$$x = x_0 = a_0 + \frac{1}{x_1} = [a_0; x_1] = [a_0; a_1, x_2] = \cdots = [a_0; a_1, \ldots, a_k, x_{k+1}]$$

By the same proof as for the fundamental recursion formulas

$$x = [a_0; a_1, \ldots, a_k, x_{k+1}] = \frac{x_{k+1}p_k + p_{k-1}}{x_{k+1}q_k + q_{k-1}}$$

$$\left|x - \frac{p_k}{q_k}\right| = \left|\frac{p_{k-1}q_k - p_k q_{k-1}}{(x_{k+1}q_k - p_k q_{k-1})q_k}\right| = \frac{1}{(x_{k+1}q_k + q_{k-1})q_k}$$

$$a_{k+1} < x_{k+1} \Rightarrow (x_{k+1}q_k + q_{k-1}) > (a_{k+1}q_k + q_{k-1}) = q_{k+1}$$

$$\Rightarrow\ 0 \le \left|x - \frac{p_k}{q_k}\right| < \frac{1}{q_{k+1}q_k} \longrightarrow 0 \text{ as } k \to \infty$$

By the squeeze theorem from analysis: $x = [a_0; a_1, \dots]$

Uniqueness:
The method used to prove uniqueness of $[a_0; a_1, \dots, a_n]$ in theorem 1 works for $[a_0; a_1, \dots]$ as well when $d \in [0,1]$ is replaced with $d \in (0,1)$. ■

Example: $\pi = [3; 7,15,1,292, \dots]$ $(c_n) = (3, \frac{22}{7}, \frac{333}{106}, \frac{355}{113}, \frac{103\,993}{33\,102}, \dots)$
Surprisingly the geometric mean of a_i is almost always the same. It is named Khinchin's constant, $K_0 = \prod_{r=1}^{\infty}\left(1 + \frac{1}{r(r+2)}\right)^{\log_2 r} = 2.6854520010 \dots.$
$\left(\left\{x | x \in \mathbb{R} \setminus \mathbb{Q} \text{ and } \lim_{n\to\infty} (a_1 a_2 \dots a_n)^{1/n} \neq K_0 \right\} \text{ is a set of measure zero} \right)$

What can be said of **periodic CF** like $x = [1; 1,1, \dots] = 1 + \cfrac{1}{1 + \cfrac{1}{1 + \ddots}}$?
It must be irrational and satisfy $x = 1 + 1/x$.
The solution is the golden ratio $\varphi = (1 + \sqrt{5})/2$.
If we extend the notation of periodic decimal representations to ICF, what about the number $x = \left[a_0; a_1, \dots, a_m, \overline{b_1, b_2, \dots, b_p}\right] = a_0 + \cfrac{1}{a_1 + \cfrac{1}{a_2 + \ddots}}$?

$$(\dots((x - a_0)^{-1} - a_1)^{-1} \dots a_{m-1})^{-1} - a_m = \underbrace{\cfrac{1}{b_1 + \cfrac{1}{b_2 + \ddots}}}_{y}$$

$$\left(\dots((y^{-1} - b_1)^{-1} - b_2)^{-1} \dots - b_{p-1}\right)^{-1} - b_p = y$$

$$\begin{aligned}
(x - a_0)^{-1} &= \frac{1}{x - a_0} \\
((x - a_0)^{-1} - a_1)^{-1} &= \frac{Ax + B}{Cx + D} \\
\left(\frac{Ax + B}{Cx + D} - a_2\right)^{-1} &= \frac{A'x + B'}{C'x + D'} \\
&\ \vdots
\end{aligned}$$

$$\begin{aligned}
(y^{-1} - b_1)^{-1} &= \frac{y}{1 - b_1 y} \\
((y^{-1} - b_1)^{-1} - b_2)^{-1} &= \frac{Ay + B}{Cy + D} \\
&\ \vdots
\end{aligned}$$

$$\frac{\alpha y + \beta}{\gamma y + \delta} = y \to y = q_1 + q_2\sqrt{n} \quad y \in \mathbb{Q}(\sqrt{n})$$

$\mathbb{Q}(\sqrt{n})$ quadratic field (n is a square-free integer)

$$\frac{\alpha x + \beta}{\gamma x + \delta} = q_1 + q_2\sqrt{n} \to x \in \mathbb{Q}(\sqrt{n})$$

$\alpha, \beta, \gamma, \delta \in \mathbb{Z}$ $q_1, q_2 \in \mathbb{Q}$

All periodic continued fractions are quadratic irrationals, i.e. solutions to quadratic equations with $\mathbb{Q}$-coefficients.

Theorem 4. x is a quadratic irrational $\Leftrightarrow$ x has a periodic continued fraction.

Proof.
$\Leftarrow$ See previous page.
$\Rightarrow$ The proof is a bit long, I will omit it, try searching the internet for "Galois' Memory on Continued Fractions" and look at page 24.

Examples: $\sqrt{19} = [4;\overline{2,1,3,1,2,8}]$

$$\sqrt{k^2+c} = [k;\overline{2k/c\,,2k}]$$

$$\sqrt{d} = \left[\lfloor\sqrt{d}\rfloor;\overline{a_1,a_2,\dots,a_2,a_1,2\lfloor\sqrt{d}\rfloor}\right]$$ if d is a square-free integer.

It's time to look at generalized continued fractions (GCF), for this we need:

Theorem 5. Euler's continued fraction formula ($a_i \in \mathbb{R}$ or $\mathbb{C}$)

$$a_0 + a_0a_1 + \cdots + a_0a_1\cdots a_n = \cfrac{a_0}{1-\cfrac{a_1}{1+a_1-\cfrac{a_2}{\ddots-\cfrac{a_{n-1}}{1+a_{n-1}-\cfrac{a_n}{1+a_n}}}}}$$

Proof.
You can prove it by induction but I will try to rewrite the left side by pulling out common factors and replacing $1/x$ with $1-y$ when x is bigger than one, as if I did not know the final formula.

$$\sum_{i=0}^{n}\prod_{j=0}^{i}a_j = \frac{a_0}{\left(\frac{1}{1+\sum_{i=1}^{n}\prod_{j=1}^{i}a_j}\right)} = \frac{a_0}{1-\left(1-\frac{1}{1+\sum_{i=1}^{n}\prod_{j=1}^{i}a_j}\right)}$$

$$= \frac{a_0}{1-\left(\frac{\sum_{i=1}^{n}\prod_{j=1}^{i}a_j}{1+\sum_{i=1}^{n}\prod_{j=1}^{i}a_j}\right)} = \cfrac{a_0}{1-\cfrac{a_1}{\left(\frac{1+\sum_{i=1}^{n}\prod_{j=1}^{i}a_j}{1+\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}\right)}}$$

$$= \cfrac{a_0}{1-\cfrac{a_1}{\left(\frac{(1+a_1)\left(1+\sum_{i=2}^{n}\prod_{j=2}^{i}a_j\right)-\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}{1+\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}\right)}}$$

$$= \cfrac{a_0}{1-\cfrac{a_1}{1+a_1-\cfrac{\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}{1+\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}}} = \cfrac{a_0}{1-\cfrac{a_1}{1+a_1-\cfrac{a_2}{\left(\frac{1+\sum_{i=2}^{n}\prod_{j=2}^{i}a_j}{1+\sum_{i=3}^{n}\prod_{j=3}^{i}a_j}\right)}}} = \cdots$$

The grey part repeats itself with new starting indices until it finally becomes.

$$\frac{1+\sum_{i=n-1}^{n}\prod_{j=n-1}^{i}a_j}{1+\sum_{i=n}^{n}\prod_{j=n}^{i}a_j}=\frac{1+a_{n-1}+a_{n-1}a_n}{1+a_n}=\frac{(1+a_{n-1})(1+a_n)-a_n}{1+a_n}=$$

$$1+a_{n-1}-\frac{a_n}{1+a_n}$$ ■

If $f(x)$ is analytic at zero it be expanded in an infinite Taylor series and its terms can be rewritten to suit Euler's formula if $f^{(n)}(0)\neq 0$ for all $n\in\mathbb{N}$.

$$f(x)=\sum_{n=0}^{\infty}\frac{f^{(n)}(0)}{n!}x^n\to\ f(x)=f(0)+f(0)\sum_{n=1}^{\infty}\prod_{m=1}^{n}\frac{f^{(m)}(0)x}{mf^{(m-1)}(0)}\to$$

$$a_0=f(0)\text{ and }a_n=\frac{f^{(n)}(0)x}{nf^{(n-1)}(0)}\text{ if }n>0\text{ in Eulers' formula }\to$$

$$f(x)=\cfrac{f(0)}{1-\cfrac{f'(0)x}{f(0)+f'(0)x-\cfrac{f(0)f''(0)x}{2f'(0)+f''(0)x-\cfrac{2f'(0)f'''(0)x}{3f''(0)+f'''(0)x-\ddots}}}}$$

$f(x)=e^z$ and $f^{(n)}(0)=1$ for all n:

$$e^z=\cfrac{1}{1-\cfrac{z}{1+z-\cfrac{z}{2+z-\cfrac{2z}{3+z-\cfrac{3z}{4+z-\ddots}}}}}\qquad e=\cfrac{1}{1-\cfrac{1}{2-\cfrac{1}{3-\cfrac{2}{4-\cfrac{3}{5-\ddots}}}}}$$

Similar methods on $\log\frac{1+z}{1-z}$ with $z=i$ and $\log\frac{1+i}{1-i}=\frac{i\pi}{2}$:

$$\log\frac{1+z}{1-z}=\cfrac{2z}{1-\cfrac{z^2}{3+z^2-\cfrac{(3z)^2}{5+3z^2-\cfrac{(5z)^2}{7+5z^2-\ddots}}}}\qquad \frac{\pi}{4}=\cfrac{1}{1+\cfrac{1^2}{2+\cfrac{3^2}{2+\cfrac{5^2}{2+\ddots}}}}$$

It seems we are close to proving e and π irrational but looks can deceive. There is no easy road in the mathematical landscape from GCF to SCF.

$$e=[2;1,2,1,1,4,1,1,6,1,1,8,1,1,10,1,\ldots]$$
$$\pi=[3;7,15,1,292,1,1,1,2,1,3,1,14,2,\ldots]\text{ no obvious pattern}$$

Square roots can be computed with GCFs: $\sqrt{x}=1+\frac{x-1}{1+\sqrt{x}}\to\sqrt{x}=1+\cfrac{x-1}{2+\cfrac{x-1}{2+\ddots}}$

The goal of this final section on CF is to look at how well convergents p_n/q_n approximate their SCF $[a_0; a_1, \dots]$ and to connect it to algebraic numbers.

Definition 4.
$x \in \mathbb{R}$ is **algebraic over** $\mathbb{Z}$ ($x \in \mathbb{A}$) iff:
x is a root of a polynomial with coefficients in $\mathbb{Z}$ ($p(x) = 0 \wedge p(x) \in \mathbb{Z}[x]$).
x is **algebraic of degree n** iff:
x is a root of an irreducible polynomial in $\mathbb{Z}[x]$ of degree n.
x is called a transcendental number if $x \in \mathbb{R} \setminus \mathbb{A}$.

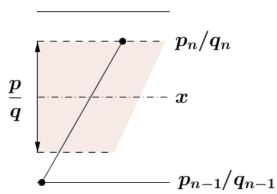
Examples:
$p/q \in \mathbb{Q}$ is algebraic (of degree 1) since it's a root of $qx - p \in \mathbb{Z}[x]$.
$2^{1/3}$ is algebraic of degree 3. $\left(2^{1/3}\right)^3 - 2 = 0,\ x^3 - 2$ is irreducible in $\mathbb{Z}[x]$.

Theorem 6. Let $x = [a_0; a_1, a_2, \dots] \in \mathbb{R} \setminus \mathbb{Q}$ and $c_n = \frac{p_n}{q_n} = [a_0; a_1, \dots, a_n]$.
If $\frac{p}{q} \in \mathbb{Q}$ is a reduced fraction $(p, q) = 1$ (trying to approximate x) then

$$\left|x - \frac{p}{q}\right| < \left|x - \frac{p_n}{q_n}\right| \Rightarrow q > q_n$$

Any rational approximation to x better than p_n/q_n has a bigger denominator.

Proof.
Lemma 1 and 2:



$$\left|x - \frac{p_n}{q_n}\right| < \left|x - \frac{p_{n+1}}{q_{n+1}}\right|$$

$$\left|\frac{p_{n-1}}{q_{n-1}} - \frac{p}{q}\right| < \left|\frac{p_n}{q_n} - \frac{p_{n-1}}{q_{n-1}}\right| = \left|\frac{p_n q_{n-1} - p_{n-1} q_n}{q_n q_{n-1}}\right| = \frac{1}{q_n q_{n-1}}$$

$$\parallel$$

$$\left|\frac{p_{n-1} q - q_{n-1} p}{q_{n-1} q}\right| \qquad \text{so} \qquad \frac{|p_{n-1} q - q_{n-1} p|}{q} < \frac{1}{q_n}$$

$|p_{n-1} q - q_{n-1} p| > 0$ since otherwise $\frac{p}{q} = \frac{p_{n-1}}{q_{n-1}}$ contradicting properties of $\frac{p}{q}$.

$$\frac{1}{s} \leq \frac{|p_{n-1} q - q_{n-1} p|}{q} < \frac{1}{q_n} \quad \Rightarrow \quad \frac{1}{s} < \frac{1}{q_n} \quad \Rightarrow \quad s > q_n \qquad \blacksquare$$

Theorem 7. Liouville's theorem
$\alpha \in \mathbb{A} \setminus \mathbb{Q}$ of degree $n \Rightarrow \exists C > 0$ such that for every rational number p/q:

$$\left|\alpha - \frac{p}{q}\right| > \frac{C}{q^n}$$

x is "hard" to approximate with rational numbers, the lower the degree the harder the approximation.

Stated in its contrapositive version it becomes: If $x \in \mathbb{R} \setminus \mathbb{Q}$ and if for any $C > 0$ and any $n \in \mathbb{N}$ there exists integers p and q $(q > 0)$ such that:

$\left|\alpha - \frac{p}{q}\right| \leq \frac{C}{q^n}$ then $x \in \mathbb{R} \setminus \mathbb{A}$. If x is "well" approximated with rationals then x is transcendental.

Proof.
Let $\alpha \in \mathbb{A} \setminus \mathbb{Q}$ of degree n with $f(x) = 0$ for $f(x) = \sum_{k=0}^{n} a_k x^k$ $a_k \in \mathbb{Z}$.
The factor theorem gives $f(x) = (x - \alpha)g(x)$. $g(\alpha) \neq 0$ since:

$$g(\alpha) = 0 \Rightarrow (x-\alpha)^2 | f(x) \Rightarrow (x-\alpha) | f'(x) \Rightarrow \begin{matrix} f'(\alpha) = 0 \;\; f'(x) \in \mathbb{Z}[x] \\ \deg(f') = n - 1 \\ \text{contradiction} \end{matrix}$$

$g(\alpha) \neq 0$ and $g \in C(\mathbb{R}) \Rightarrow \exists \delta > 0$ s.t. $g(x) \neq 0$ for $|x - \alpha| < \delta$
Pick integers p and q such that p/q is in this interval and $g(p/q) \neq 0$.

$$x - \alpha = \frac{f(x)}{g(x)} \Rightarrow$$

$$\frac{p}{q} - \alpha = \frac{f(p/q)}{g(p/q)} = \frac{q^n \sum_{k=0}^{n} a_k \left(\frac{p}{q}\right)^n}{q^n g\left(\frac{p}{q}\right)} = \frac{a_0 q^n + a_1 p q^{n-1} + \cdots + a_n p^n}{q^n g(p/q)}$$

$$f\left(\frac{p}{q}\right) = \underbrace{\left(\frac{p}{q} - \alpha\right)}_{\neq 0} \underbrace{g\left(\frac{p}{q}\right)}_{\neq 0} \Rightarrow |a_0 q^n + a_1 p q^{n-1} + \cdots + a_n p^n| \geq 1$$

Let $M = \sup\{|g(x)| \mid \alpha - \delta < x < \alpha + \delta\}$
$0 < M < \infty$ since a polynomial like g is bounded in any finite interval.

$$\left|\frac{p}{q} - \alpha\right| \geq \left|\frac{1}{q^n g\left(\frac{p}{q}\right)}\right| \geq \frac{1}{Mq^n} \text{ for any } \frac{p}{q} \text{ satisfying } \left|\alpha - \frac{p}{q}\right| < \delta$$

Other p/q satisfies: $\left|\alpha - \frac{p}{q}\right| > \delta > \frac{\delta}{q^n}$

With $C = \frac{1}{2}\min\left(\delta, \frac{1}{M}\right)$ we get $\left|\alpha - \frac{p}{q}\right| > \frac{C}{q^n}$ ■

To find a number that is transcendental, i.e. not a root of a polynomial with integer coefficients all we need is an irrational number α for which we can find integers p, q so that $|\alpha - p/q| \leq C/q^n$ for every $C > 0$ and $n \in \mathbb{N}$.

Let $\alpha = [a_0; a_1, a_2, \dots]$ with $a_0 = 0, a_1 = 1$ and $a_n = (q_{n-1})^{n-2} + 1$.

$$\left|\alpha - \frac{p_n}{q_n}\right| < \frac{1}{q_n q_{n+1}} < \frac{1}{a_{n+1} {q_n}^2} < \frac{1}{{q_n}^{n+1}} = \left(\frac{1}{q_n}\right) / {q_n}^n$$

By making n large enough this is smaller than C/q^n for any n and C.

When you have convinced yourself of each of the inequalities you have found a transcendental number $\alpha = [0; 1,2,4,170,2213^3 + 1, \dots]$. The first proof of transcendental numbers was given by Liouville in 1844.

Definition 5. Liouville numbers $\mathbb{L}$ are irrational numbers x such that for any positive integer n there exists integers p, q $(q > 1)$ that fulfills:

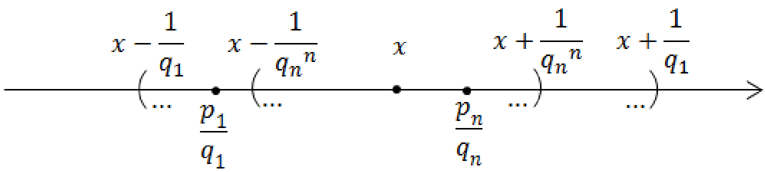
$$\left|x - \frac{p}{q}\right| < \frac{1}{q^n}$$

In other words, they have close rational number approximations.

An example of such a number is the binary Liouville‘s constant:

$$\sum_{k=1}^{\infty} \frac{1}{2^{k!}} = (0.110001000000000000000001 \dots)_2 = 2^{-1} + 2^{-2} + 2^{-6} + 2^{-24} + \cdots$$

All Liouville numbers are transcendental but most transcendental numbers are not Liouville numbers. They are an uncountable dense subset of the real numbers but their Lebesgue measure is zero $\lambda(\mathbb{L}) = 0$, most transcendental numbers can't be approximated all the way by p_n/q_n in this image:



How big n can be when approximating x with a sequence of rational numbers obeying $0 < |x - p_n/q_n| < q_n^{-n}$ is given by Liouville-Roth's constant $\mu(x)$.

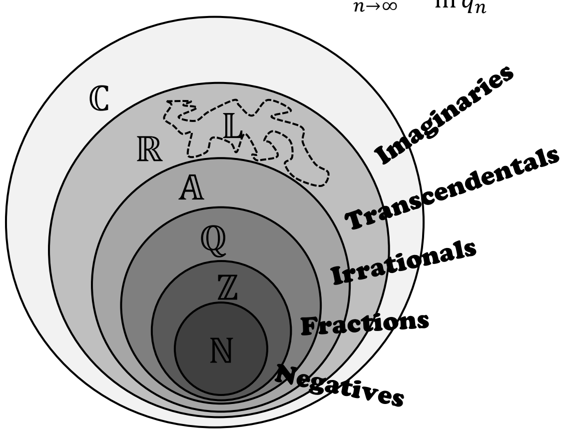
$$\mu(x) = \inf\left\{\alpha \in \mathbb{R}: \left|\left\{(p,q) \in \mathbb{Z} \times \mathbb{N}_2 : 0 < \left|x - \frac{p}{q}\right| < \frac{1}{q^\alpha}\right\}\right| = \infty\right\} \quad \mathbb{N}_2 = \{2,3, \dots\}$$

Almost all real numbers have $\mu(x) = 2$ (the complement being a null set).
If $x = [a_0; a_1, \dots]$ with convergents p_n/q_n then:

$$\mu(x) = 1 + \limsup_{n\to\infty} \frac{\ln q_{n+1}}{\ln q_n}$$

$\mu(e) = 2$ and $\mu(\pi) < 7.6$
$x \in \mathbb{Q} \Rightarrow \mu(x) = 1$
$x \in \mathbb{A} \setminus \mathbb{Q} \Rightarrow \mu(x) = 2$ (*)
$x \in \mathbb{R} \setminus \mathbb{A} \Rightarrow \mu(x) \geq 2$
$x \in \mathbb{L} \Leftrightarrow \mu(x) = \infty$

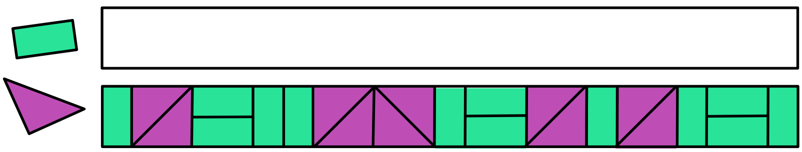
(*) proved by Klaus Roth, for which he was awarded the Fields medal in 1958.



A few examples on generating functions (GFs)

Using a gf to count extended domino tilings

This example is inspired by the introduction to generating functions given in *Concrete Mathematics* by D. Knuth et al. In how many ways can you cover a $2 \times n$ rectangle without overlap? You have two pieces at your disposal, a 2×1 domino-tile and a triangular piece cut diagonally from a 2×2 square.



Let T_n be the number of different tilings of a $2 \times n$ rectangle. The first step is always important. You must choose zero bricks for T_0 and there is only way to order the empty set so $T_0 = 1$. This tiling is illustrated with a stroke |. Let T be the sum of all possible tilings. This will be our entrance to a GF of $\langle T_n \rangle$.

$$T = | + ▯ + ▯▯ + ⊟ + ⧄ + ⧅ + ▯▯▯ + ▯⊟ + ▯⧄ + ▯⧅ + ⊟▯ + \dots$$

Addition in this context is just a summing up of different patterns. Putting two patterns after each other represents multiplication. This operation does not commute, ▯·⊟ ≠ ⊟·▯ and sometimes it leads to improper tilings ▯⧅ = ▯⧅. Rearranging terms and applying "domino" arithmetic leads to:

$$T = | + ▯T + ⊟T + ⧄T + ⧅T \quad \rightarrow \quad T = \frac{|}{| - (▯ + ⊟ + ⧄ + ⧅)}$$

$$\frac{1}{1-z} = 1 + z + z^2 + \cdots \quad \rightarrow \quad T = | + (▯ + ⊟ + ⧄ + ⧅)^2 + (▯ + ⊟ + ⧄ + ⧅)^3 + \cdots$$

This sum contains more information than needed. Skipping some information and treating ⊟, ⧄ and ⧅ as ⊠ gives:

$$T = \sum_{k \geq 0} (▯ + ▭^2 + 2◺^2)^k = \sum_{k \geq 0} (▯ + 3⊠^2)^k = \sum_{j,m} \binom{j+m}{j} 3^m ▯^j ⊠^{2m}$$

$\binom{j+m}{j} 3^m$ is the number of ways to tile a $2 \times (j + 2m)$ rectangle composed of j ▯-pieces and m occurrences of ⊟, ⧄ or ⧅ in the tiling.

$$T = \frac{1}{1 - z - 3z^2} \quad \rightarrow \quad T_n = [z^n] \frac{1}{1 - z - 3z^2}$$

If we don't use triangular pieces the gf will be $T_{▯} = \frac{1}{1-z-z^2}$ which if we add a factor z in the numerator is the same gf as Fibonacci's.

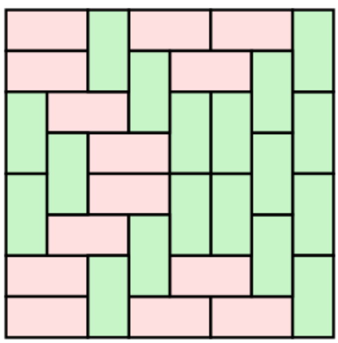
$$F_{Fib} = \frac{z}{1-z-z^2} \to \widehat{\boldsymbol{F}}_{Fib} = \langle 0,1,1,2,3,5,8,11,\dots \rangle$$

$$T_{▯} = \frac{1}{1-z-z^2} \to \widehat{\boldsymbol{T}}_{▯} = \langle 1,1,2,3,5,\dots \rangle = \langle F_{n+1} \rangle$$

A closed form for T_n can be obtained by partial fraction decomposition of T or by solving a recurrence relation for T_n as a difference equation. The easiest way to count the number of tilings is to use the recurrence relation.

$$\begin{cases} T_n = T_{n-1} + 3T_{n-2} \\ T_0 = 1 \\ T_1 = 1 \end{cases} \to \langle T_n \rangle = \langle 1,1,4,7,19,40,\dots \rangle$$

Let me finally mention a gf that is much harder to find. How many ways can you tile a rectangle of size $m \times n$ with dominoes? The answer was found by Kasteleyn and Fisher–Temperley in 1961. The problem is connected to graph theory, statistical mechanics and phase transitions.



$$\text{GF} = 2^{mn/2} \prod_{\substack{1 \le j \le m \\ 1 \le k \le m}} \left(\left(\cos^2 \frac{j\pi}{m+1} \right) ▯^2 + \left(\cos^2 \frac{k\pi}{n+1} \right) ▭^2 \right)^{1/4}$$

The coefficient of $▯^j ▭^k$ counts tilings with j vertical and k horizontal tiles. The number of tilings of a $2m \times 2n$ rectangle with $2mn$ dominos is:

$$4^{mn} \prod_{j=1}^{m} \prod_{k=1}^{n} \left(\cos^2 \frac{j\pi}{2m+1} + \cos^2 \frac{k\pi}{2n+1} \right)$$

A gf for money

How many ways are there to give change worth 50 cent with Pennies $\boxed{1}$, Nickels $\boxed{5}$, Dimes $\boxed{10}$, Quarters $\boxed{25}$ and Half-dollars $\boxed{50}$. This exercise was popularized by Georg Pólya (1887–1985), a polymath from Hungary with great achievements in combinatorics and many other fields including education. His most known book *How to solve it* from 1945 is a guidebook on how to solve mathematical problems. One way to solve the 50¢ problem is to use a gf that corresponds to money, each term is a selection of coins.

$P = \emptyset + \boxed{1} + \boxed{1}^2 + \cdots = \sum_{n_1} \boxed{1}^{n_1}$ $\quad P \sim$ Number of pennies, $(n_1 \in \mathbb{N}_0)$.
$N = \emptyset + \boxed{5} + \boxed{5}^2 + \cdots = \sum_{n_5} \boxed{5}^{n_5}$ $\quad N^- = N \cdot P \sim$ 5¢ or 1¢ coins.
$D = \sum_{n_{10}} \boxed{10}^{n_{10}}$ $\quad D^- = D \cdot N^- \sim$ 10¢, 5¢ or 1¢ coins.
$Q = \sum_{n_{25}} \boxed{25}^{n_{25}}$ $\quad Q^- = Q \cdot D^- \sim$ 25¢, 10¢, 5¢ or 1¢ coins.
$H = \sum_{n_{50}} \boxed{50}^{n_{50}}$ $\quad H^- = H \cdot Q^- \sim$ 50¢, 25¢, 10¢, 5¢ or 1¢ coins.

To find the number of terms in H^- representing a value of 50¢ let $\boxed{m} = z^m$.

$$H^- = \sum_{n_i \in \mathbb{N}_0} z^{n_1 + 5n_5 + 10n_{10} + 25n_{25} + 50n_{50}} = \frac{1}{(1-z)(1-z^5)(1-z^{10})(1-z^{25})(1-z^{50})}$$

$[z^{50}]H^-$ can be retrieved by using recursive relations found from the gf by identifying the coefficient of z^n in each relation.

$P = 1/(1-z)$ $\quad (1-z)P = 1 \ (1 = z^0)$ $\quad P_n = P_{n-1} + [n = 0]$
$N^- = P/(1-z^5)$ $\quad (1-z^5)N^- = P$ $\quad N_n^- = N_{n-5}^- + P_n$
$D^- = N^-/(1-z^{10})$ $\quad (1-z^{10})D^- = N^-$ $\quad D_n^- = D_{n-10}^- + N_n^-$
$Q^- = D^-/(1-z^{25})$ $\quad (1-z^{25})Q^- = D^-$ $\quad Q_n^- = Q_{n-25}^- + D_n^-$
$H^- = Q^-/(1-z^{50})$ $\quad (1-z^{50})H^- = Q^-$ $\quad H_n^- = H_{n-50}^- + Q_n^-$

n	**0**	**5**	**10**	**15**	**20**	**25**	**30**	**35**	**40**	**45**	**50**
$P_n = P_{n-1} + \delta_{n,0}$	1	1	1	1	1	1	1	1	1	1	1
$N_n^- = N_{n-5}^- + P_n$	1	2	3	4	5	6	7	8	9	10	11
$D_n^- = D_{n-10}^- + N_n^-$	1	2	4	6	9	12	16	20	25	30	36
$Q_n^- = Q_{n-25}^- + D_n^-$	1	2	4	6	9	13	18	24	31	39	49
$H_n^- = H_{n-50}^- + Q_n^-$	1	2	4	6	9	13	18	24	31	39	50

$\wp(n) = [z^n] \prod_{n \in \mathbb{N}_1} (1 - z^n)^{-1}$ is the number of partitions of n, the different ways to write n as a sum of positive integers without regard to order.

$5 = 4 + 1 = 3 + 2 = 3 + 1 + 1 = 2 + 2 + 1 = 2 + 1 + 1 + 1 = 1 + 1 + 1 + 1 + 1 \rightarrow \wp(5) = 7$

Faulhaber's formula and exponential generating functions

In the last example our goal will be to find a gf that can solve a task that has engaged mathematicians since the days of Pythagoras, to calculate $S_p(n)$:

$$S_p(n) = 0^p + 1^p + 2^p + \cdots + (n-1)^p \quad (\text{ Sum of } n \text{ terms })$$

$S_0(n) = 1 + \cdots + 1 = n$, $S_1(n) = 0 + 1 + \cdots + (n-1) = n(n-1)/2$. The last sum was calculated by Pythagoreans and a young German boy named Friedrich Gauss. Archimedes derived a polynomial for $S_2(n)$ and Aryabhata from India found one for $S_3(n)$ around 500 AD. In 1000 AD Abu Bakr al-Karaji, Bagdad proved a polynomial for $S_3(n)$ and al-Haytham derived $S_4(n)$ in the same period. Johann Faulhaber from Germany derived polynomials for $p = 1, \ldots, 17$ in 1600 AD. A general formula for $S_p(n)$ is named after him but he never proved it. The formula contain $\langle B_k \rangle = \langle 1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, 0, -\frac{1}{30}, 0, \frac{5}{66}, \ldots \rangle$, a sequence of rational numbers. The first proof was given 1834 by Carl Jacobi.

$$S_p(n) = \frac{1}{p+1} \sum_{k=0}^{p} \binom{p+1}{k} B_k n^{p+1-k}$$

$S_p(n)$	$p \backslash n$	$\Sigma_\emptyset$ 0	$0^p +$ 1	$1^p +$ 2	$2^p +$ 3	$3^p +$ 4	$4^p +$ 5	$+ \ldots + n^p$ n
$\sum_{k=0}^{n-1} k^0$	**0**	0	1	2	3	4	5	n
$\sum_{k=0}^{n-1} k^1$	**1**	0	0	1	3	6	10	$(n^2 - n)/2$
$\sum_{k=0}^{n-1} k^2$	**2**	0	0	1	5	14	30	$(2n^3 - 3n^2 + n)/6$
$\sum_{k=0}^{n-1} k^3$	**3**	0	0	1	9	36	100	$(n^4 - 2n^3 + n^2)/4$
$\sum_{k=0}^{n-1} k^4$	**4**	0	0	1	17	98	354	$(6n^5 - 15n^4 + 10n^3 - n)/30$
$\sum_{k=0}^{n-1} k^5$	**5**	0	0	1	33	276	1300	$(2n^6 - 6n^5 + 5n^4 - n^2)/12$

Sometimes a sequence $\langle g_n \rangle$ has a GF whose properties are quite complicated while the closely related $\langle g_n/n! \rangle$ has a gf that is much easier to handle. We can always multiply with $n!$ in the end.

Definition 1.

$$G^e(z) \equiv \sum_{n=0}^{\infty} g_n \frac{z^n}{n!}$$

This GF is called exponential, $G^e(z) = e^z = \sum_{n \in \mathbb{N}_0} z^n/n!$ is GF of $\hat{\boldsymbol{g}} = \langle 1 \rangle$. I will resist the temptation to call it gfe, the traditional choice would be EGF, I'll choose the more uplifting GF^e. Faulhaber's formula can be proved with induction but an approach with GF^e is more elegant and powerful.

GF^{e}s are useful because there is a natural convolution of sequences $\widehat{\boldsymbol{g}} \star^b \widehat{\boldsymbol{h}}$ with nice properties that will make $\star^b$ correspond to multiplication of GF^{e}s $G^e(z) \cdot H^e(z)$ in the same way as $\widehat{\boldsymbol{g}} \star \widehat{\boldsymbol{h}}$ corresponds to multiplication of GFs $G(z) \cdot H(z)$. Let us therefore first review GFs and ordinary convolution.

Definition 2.

$K^\omega \equiv \{(a_0, a_1, a_2, \ldots) | a_i \in K\}$ Set of sequences of some field K like $\mathbb{R}$ or $\mathbb{C}$.

$K[X] \equiv \{a_0 + a_1 X + a_2 X^2 + \cdots | a_i \in K\}$ Polynomial ring of some field K.

$gf: K^\omega \to K[X]$, $\langle a_n \rangle \mapsto \sum_{n=0}^{\infty} a_n X^n$ Function that gives GF of a sequence.

$gf \times gf: K^\omega \times K^\omega \to K[X] \times K[X]$, $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{b}}) \mapsto (A(X), B(X))$

$\otimes: K[X]^2 \to K[X]$ Multiplication $(a_0 + a_1 X + \cdots) \otimes (b_0 + b_1 X + \cdots)$
$A(X) \otimes B(X) \mapsto \sum_{n=0}^{\infty} (\sum_{k=0}^{n} a_k b_{n-k}) X^n$

$\star: K^\omega \times K^\omega \to K^\omega, (\langle a_n \rangle, \langle b_n \rangle) \mapsto \langle \sum_{k=0}^{n} a_k b_{n-k} \rangle$ Convolution of sequences.

A fancy way of saying that convolution of sequences corresponds to multiplication of GFs is that the diagram below commutes which means that by functional composition $\otimes \circ (gf \times gf) = gf \circ \star$. This is a direct consequence of the gray parts in the definitions.

$$\begin{array}{ccc} K^\omega \times K^\omega & \xrightarrow{gf \times gf} & K[X] \times K[X] \\ \star \downarrow & & \downarrow \otimes \\ K^\omega & \xrightarrow[gf]{} & K[X] \end{array}$$

Fig. C.1 Commuting diagram

Theorem 1.

Convolution of sequences $\widehat{\boldsymbol{a}} \star \widehat{\boldsymbol{b}}$ is a commutative and associative operator with an identity $\widehat{\boldsymbol{id}} = \langle 1,0,0, \ldots \rangle$ and a unique inverse for sequences starting with non-zero elements.

I leave the proof of the theorem as an exercise to the reader. The next step is to define a convolution that makes the diagram commute with gf replaced by $gf^e: \langle f_k \rangle \mapsto f_0 + f_1 X + f_2 \frac{X^2}{2!} + f_3 \frac{X^3}{3!} + \cdots = F^e(X)$.

$$F^e(X) G^e(X) = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \frac{f_k}{k!} \cdot \frac{g_{n-k}}{(n-k)!} \right) X^n = \sum_{n=0}^{\infty} \left(\sum_{k=0}^{n} \binom{n}{k} f_k g_{n-k} \right) \frac{X^n}{n!}$$

Definition 3.

$$\langle f_n\rangle \star^b \langle g_n\rangle \equiv \langle \sum_{k=0}^{n} \binom{n}{k} f_k g_{n-k}\rangle$$

This definition of **binomial convolution** restores the commutativity of the diagram when gf is replaced by gf^e, $\hat{\boldsymbol{f}} \star^b \hat{\boldsymbol{g}}$ translates to $F^e(X)G^e(X)$.

Theorem 2.

All properties listed in theorem 1 apply to binomial convolutions as well.

Example

Let $\langle r_n\rangle = \langle 1, \frac{1}{2}, \frac{1}{3}, \dots\rangle = \langle \frac{1}{n+1}\rangle$ be the reciprocals whose cumulative sums are the harmonic numbers $H_n = 1 + 1/2 + 1/3 + \cdots + 1/n$. Our interest will be the inverse of $\hat{\boldsymbol{r}}$ and its GF^e. Denote that sequence $\hat{\boldsymbol{B}}$.

$$R^e(X) = 1 + \frac{1}{2}X + \frac{1}{3}\frac{X^2}{2!} + \cdots = \frac{1}{X}\left(X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots\right) = \frac{1}{X}(e^X - 1)$$

$$R^e(X) \cdot B^e(X) = 1 \rightarrow B^e(X) = \frac{X}{e^X - 1}$$

$$\hat{\boldsymbol{r}} \star^b \hat{\boldsymbol{B}} = \widehat{\boldsymbol{id}} \rightarrow \sum_{k=0}^{n} \binom{n}{k} B_{n-k} r_k = [n = 0] \rightarrow$$

$$\langle B_0 r_0, B_1 r_0 + B_0 r_1, B_2 r_0 + 2B_1 r_1 + B_0 r_2, B_3 r_0 + 3B_2 r_1 + 3B_1 r_2 + B_0 r_3, \dots\rangle = \langle \delta_{n0}\rangle$$

$$\langle B_n\rangle = \langle 1, -\frac{1}{2}, \frac{1}{6}, 0, -\frac{1}{30}, 0, \frac{1}{42}, 0, -\frac{1}{30}, 0, \frac{5}{66}, 0, \frac{691}{2730}, 0, \frac{7}{6}, 0, -\frac{3617}{510}, \dots\rangle$$

These numbers crop up all over the mathematical landscape. They are called Bernoulli numbers in honor of the Swiss Jakob Bernoulli but it was Faulhaber the ‘Arithmetician of Ulm’ who was the first to study them in the 1630s. His work was carried on by Bernoulli. In his famous *Ars Conjectandi*, published posthumously in 1713 he claims to have calculated within half a quarter:

$$1^{10} + 2^{10} + \cdots + 1000^{10} = 91409924241424243424241924242500$$

Faulhaber’s formula reduces this to a sum of just 11 terms. The numbers of Bernoulli were independently discovered and described by Seki Kowa from Japan in a book published posthumously in 1712. The first algorithm ever made for a machine was written in 1842 by Ada Lovelace. Her father was the English poet Lord Byron, the machine called “The Analytical Engine” was made by Charles Babbage and the algorithm produced Bernoulli numbers.

Bernoulli numbers

Bernoulli numbers B_n are of great importance in number theory and elsewhere. They have $B_{2k+1} = 0$ when $k \geq 1$ and alternate between positive and negative values, $|B_{2n}| \sim 4\sqrt{\pi n}(n/\pi e)^{2n}$ for big n. There are two versions. The first version that is used in this book has $B_1 = -½$ while the second version has B_1'=½ which gives $B_n' = (-1)^n B_n$.

Recursively

$$\sum_{k=0}^{n-1} \binom{n}{k} B_k = \delta_{n1}$$

1=	**B_0**
0=	$1B_0+2$**B_1**
0=	$1B_0+3B_1+3$**B_2**
0=	$1B_0+4B_1+6B_2+4$**B_3**
0=	$1B_0+5B_1+10B_2+10B_3+5$**B_4**

GF[e]

$$\frac{z}{e^z - 1} = \sum_{k=0}^{\infty} B_k \frac{z^k}{k!}$$

Analytically

$$B_n = \frac{n!}{2\pi i} \oint \frac{z}{e^z - 1} \frac{dz}{z^{n+1}}$$

Explicitly

$$B_n = \sum_{k=0}^{n} \frac{1}{k+1} \sum_{r=0}^{k} (-1)^r \binom{k}{r} r^n$$

Taylor series of $\tan(z)$ and $\tanh(z)$ contain B_n, Riemann's zeta function $\zeta(z) = \sum_{k=1}^{\infty} k^{-z}$ is related to them, $B_n = -n\zeta(1-n)$. The Bernoulli polynomials defined by $B_n(x) \equiv \frac{D}{e^D - 1} x^n$ have $B_n(0) = B_n$. One reason for their frequent presence is their GF[e]. $f(z) = 1/(e^z - 1)$ has a simple pole at $z = 0$. When it is removed by a factor z we get $G(z) = z/(e^z - 1)$ holomorphic in $\mathbb{C}$. What is it that makes these functions so special?

Euler-Maclaurin's important formula connects $I = \int_a^b f(x)dx$ with the sum $S = f(a+1) + \cdots + f(b)$. Their difference is approximated with a term containing B_n and derivatives at the endpoints and then there is an error term that can be made small, and it disappears when $f(x)$ is a polynomial.

$$S - I = \sum_{k=1}^{p} \frac{B_k}{k!} \left(f^{k-1}(b) - f^{k-1}(a) \right) + R \text{ with } |R| \leq \frac{2\zeta(2p)}{(2\pi)^{2p}} \int_a^b \left| f^{(2p)}(x) \right| dx$$

To explain the occurrence of B_k, look at $f(n+1) = \sum_{k\geq 0} \frac{f^k(n)}{k!}$. The shift operator becomes $S = e^D$ and $\Delta \equiv S - 1 = e^D - 1$. As integration is the inverse of derivation D^{-1}, summing is the inverse of difference operation $\Delta^{-1} = 1/(e^D - 1)$, which when expanded in powers of D starts with a D^{-1} term for integration followed by D^{k-1}-terms with coefficients $B_k/k!$.

The tools for calculating $S_p(n)$ are now in place. Let n be fixed and vary p, look at the table of $S_p(n)$ vertically instead of horizontally.

$$\begin{aligned} Sn^e(z) &= \sum_{p\ge 0} S_p(n)\frac{z^p}{p!} = \sum_{p\ge 0}\left(\sum_{k=0}^{n-1} k^p\right)\frac{z^p}{p!} = \sum_{k=0}^{n-1}\sum_{p\ge 0}\frac{(kz)^p}{p!} = \sum_{k=0}^{n-1} e^{kz} \\ &= \frac{e^{nz}-1}{e^z-1} = B^e(z)\frac{e^{nz}-1}{z} \end{aligned}$$

A sequence with GF[e] equal to $H^e(z)$ has $h_p = p!\,[z^p]H^e(z)$ ($=\langle h_0, h_1, \dots\rangle_p$)

If $H^e(z) = F^e(z)G^e(z)$ then $\langle h_p\rangle = \langle f_i\rangle \star^b \langle g_j\rangle = \langle \sum_{k=0}^{p}\binom{p}{k} f_k g_{p-k}\rangle$

$$G^e(z) \equiv \frac{e^{nz}-1}{z} = \sum_{k=1}^{\infty}\frac{(nz)^k}{k!\,z} = \sum_{k=0}^{\infty}\underbrace{\frac{n^{k+1}}{k+1}}_{g_k}\cdot\frac{z^k}{k!}$$

$Sn^e(z) = B^e(z)\cdot G^e(z) \rightarrow$

$$\begin{aligned} S_p(n) &= \left(\langle B_i\rangle \star^b \langle\frac{n^{j+1}}{j+1}\rangle\right)_p \\ &= \sum_{k=0}^{p}\binom{p}{k} B_k \frac{n^{p-k+1}}{p-k+1} \\ &= \frac{1}{p+1}\sum_{k=0}^{p}\binom{p}{k}\frac{p+1}{p+1-k} B_k n^{p+1-k} \\ &= \frac{1}{p+1}\sum_{k=0}^{p}\binom{p+1}{k} B_k n^{p+1-k} \end{aligned}$$

$$0^p + 1^p + 2^p + \cdots + (n-1)^p = \frac{1}{p+1}\sum_{k=0}^{p}\binom{p+1}{k} B_k n^{p+1-k}$$

With Bernoulli numbers B_k defined by:

$$\widehat{\boldsymbol{B}} \star^b \hat{\boldsymbol{r}} = \widehat{\boldsymbol{\iota d}}$$

or

$$\sum_{k=0}^{m-1}\binom{m}{k} B_k = [m=1]$$

Thermodynamics and Neural Networks

(From Blue book on NN)

Statistical Mechanics and Phase Transitions

(Ising Model included, from course in Statistical Mechanics)

Discriminants and Symmetric Polynomials

A quadratic polynomial $f(x) = x^2 + px + q = (x - r_1)(x - r_2)$ with roots $r_{1,2} = (-p \pm \sqrt{p^2 - 4q})/2$ has a multiple root $r_1 = r_2$ if $\Delta = p^2 - 4q$ is zero. This test can be derived from the relations between roots and coefficients:

$$\begin{cases} r_1 r_2 = q \\ r_1 + r_2 = -p \end{cases} \rightarrow (r_1 - r_2)^2 = (r_1 + r_2)^2 - 4r_1 r_2 = p^2 - 4q$$

The goal will be to generalize this to polynomials of all degrees.

Definition 1. The **discriminant** Δ is a function of a polynomial:

$$\Delta: \mathbb{F}[X] \rightarrow \mathbb{F}, P = \sum_{k=0}^{n} a_k X^k \mapsto a_n^{2n-2} \cdot \prod_{1 \le i < j \le n} (r_i - r_j)^2$$

where $(r_k)_{k=1}^n$ are the roots of P in the algebraic closure of $\mathbb{F}$.

Theorem 1. (Vieta's formulas, François Viète 1540–1603)
The coefficients of $f(z) = \sum_{k=0}^{n} a_n z^n$ are related to its roots by formulas:

$$a_{n-k} = a_n \cdot (-1)^k \cdot \sum_{1 \le i_1 < i_2 < \cdots < i_k \le n} r_{i_1} r_{i_2} \dots r_{i_k} \quad \text{for } k = 1,2,\dots,n$$

Proof. It follows directly from expansion.

$$a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0 = a_n (z - r_1)(z - r_2) \dots (z - r_n) \qquad \blacksquare$$

Clearly, $\Delta(P) = 0 \Leftrightarrow P$ has a multiple root. Discriminants for quadratic and cubic polynomials are:

$n = 2$: $ax^2 + bx + c$ $\quad \Delta = b^2 - 4ac$ $\quad$ ← Check! ↓

$n = 3$: $ax^3 + bx^2 + cx + d$ $\quad \Delta = b^2c^2 - 4ac^3 - 4b^3d - 27a^2d^2 + 18abcd$

What is the general form of Δ, a multivariate polynomial with coefficients in $\mathbb{Z}, \mathbb{F}$ or maybe $\overline{\mathbb{F}}$? A **polynomial in several variables** in $\mathbb{F}[\boldsymbol{X}]$ is of the form:

$$P(X_1, X_2, \dots, X_n) = \sum_{i_k \in \mathbb{N}_0} \underbrace{a_{i_1 i_2 \dots i_n}}_{c} X_1^{i_1} X_2^{i_2} \dots X_n^{i_n} \quad c \in \mathbb{F}$$

The **degree of a term** is $\sum_{k=1}^{n} i_k$, its maximum over all terms is the **degree of the polynomial**. If $P(\alpha \boldsymbol{X}) = \alpha^k P(\boldsymbol{X})$ then P is **homogeneous of degree k**. The Vandermondes polynomial V_n is homogeneous of degree $n(n-1)/2$ and alternating in sign when two variables are switched.

$$V_n(r_1, r_2, \dots, r_n) = \prod_{1 \le i < j \le n} (r_j - r_i) = \begin{vmatrix} 1 & r_1 & r_1^2 & \cdots & r_1^{n-1} \\ 1 & r_2 & r_2^2 & \dots & r_2^{n-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & r_n & r_n^2 & \dots & r_n^{n-1} \end{vmatrix}$$

A polynomial that is unaffected by any switch of two variables is called a **symmetric** polynomial. This means that for any permutation σ of the indices, $P\left(X_{\sigma(1)}, X_{\sigma(2)}, \dots, X_{\sigma(n)}\right) = P(X_1, X_2, \dots, X_n)$. The discriminant $\Delta(\boldsymbol{r})$ is a symmetric function of the roots. The roots all have similar roles in the factorization into linear factors.

The discriminant is homogeneous to degree $2n-2$ in the coefficients since multiplying all coefficents with λ does not change any roots, only the prefactor a_n^{2n-2}. V_n^2 is homogeneous of degree $n(n-1)$. This leads to two requirements on the exponents of coefficients ${a_0}^{i_0}{a_1}^{i_1} \dots {a_n}^{i_n}$ in $\Delta(\boldsymbol{a})$.

$$\begin{cases} i_n + \cdots + i_1 + i_0 = 2n-2 \\ n i_n + \cdots + 1 i_1 + 0 i_0 = n(n-1) \end{cases} \quad \text{For } n = 3 \quad \begin{cases} i_a + i_b + i_c + i_d = 4 \\ 3i_a + 2i_b + i_c = 6 \end{cases}$$

P has a multiple root $\Leftrightarrow$ P and P' have a common root. This suggests an expanded discriminant to test for common roots of two polynomials P and Q.

Definition 2. The **resultant** R is a function of two polynomials:

$$R: \mathbb{F}[X] \times \mathbb{F}[Y] \to \mathbb{F}$$
$$(P, Q) = \left(\sum_{k=0}^{n} a_k X^k, \sum_{k=0}^{m} b_k Y^k \right) \mapsto {a_n}^m {b_m}^n \prod_{i=1}^{n} \prod_{j=1}^{m} \left(\alpha_i - \beta_j\right)$$

where $(\alpha_k)_{k=1}^{n}$ are the roots of P and $(\beta_k)_{k=1}^{m}$ are the roots of Q in $\overline{\mathbb{F}}_{\text{Alg}}$.

The resultant can be computed as the determinant of an $m+n$ square matrix called the Sylvester matrix. Its definition starts from two polynomials.

Definition 3. The definition of the **Sylvester matrix** is best illustrated by a concrete example. Let P and Q be as above with $n = 5$ and $m = 3$.

$$S_{P,Q} = \begin{pmatrix} a_5 & a_4 & a_3 & a_2 & a_1 & a_0 & 0 & 0 \\ 0 & a_5 & a_4 & a_3 & a_2 & a_1 & a_0 & 0 \\ 0 & 0 & a_5 & a_4 & a_3 & a_2 & a_1 & a_0 \\ b_3 & b_2 & b_1 & b_0 & 0 & 0 & 0 & 0 \\ 0 & b_3 & b_2 & b_1 & b_0 & 0 & 0 & 0 \\ 0 & 0 & b_3 & b_2 & b_1 & b_0 & 0 & 0 \\ 0 & 0 & 0 & b_3 & b_2 & b_1 & b_0 & 0 \\ 0 & 0 & 0 & 0 & b_3 & b_2 & b_1 & b_0 \end{pmatrix}$$

The resultant can be calculated as $R(P, Q) = \left|S_{P,Q}\right|$ and the discriminant is connected to the resultant via $\Delta(P) = (-1)^{n(n-1)/2} R(P, P')/a_n$.

An alternative approach for finding common roots is the Euclidean algo–rithm and gcd. P and Q have a root in common $\Leftrightarrow \deg\big(\gcd(\mathrm{P},\mathrm{Q})\big) \geq 1$. This approach is basically the same since the Sylvester matrix is involved in both methods, $\deg(\gcd(P,Q)) = n + m - \operatorname{rank} S_{P,Q}$.

The part that we still miss is how to get from a symmetric polynomial of the roots to a polynomial of the coefficients. The route goes via Vieta's formulas. They contain a set of polynomials called elementary symmetric polynomials that can be used as building blocks for symmetric polynomials. There is a trio of building blocks, homogeneous of degree k and with n variables.

Elementary symmetric polynomials:

$$e_{k,n} = \sum_{1\leq i_1<\cdots<i_k\leq n} X_{i_1} \dots X_{i_k} \qquad e_{2,3} = X_1X_2 + X_1X_3 + X_2X_3$$

Complete symmetric polynomials:

$$h_{k,n} = \sum_{1\leq i_1\leq\cdots\leq i_k\leq n} X_{i_1} \dots X_{i_k} \qquad h_{2,3} = {X_1}^2 + X_1X_2 + X_1X_3 + {X_2}^2 + X_2X_3 + {X_3}^2$$

Power sum symmetric polynomials:

$$p_{k,n} = \sum_{i=1}^{n} {X_i}^k \qquad p_{2,3} = {X_1}^2 + {X_2}^2 + {X_3}^2$$

$e_{k,n}$ is a shorter version of $\boldsymbol{e}_k(X_1, X_2, \dots, X_n)$ and similarly for $h_{k,n}$ and $p_{k,n}$.

Theorem 2. (The fundamental theorem of symmetric polynomials)
Every symmetric polynomial $P(X_1, X_2 \dots, X_n) \in A[X_1, X_2 \dots, X_n]$
has a unique representation $P\big(X_1, X_2 \dots, X_n\big) = Q(e_{1,n}, e_{2,n}, \dots, e_{n,n})$
for some polynomial $Q \in A[Y_1, Y_2, \dots, Y_n]$.
A can be any commutative ring, so $\mathbb{Z}, \mathbb{Q}, \mathbb{A}, \mathbb{R}$ or $\mathbb{C}$ are all possible.

The theorem works both with $e_{k,n}$ and $h_{k,n}$ as a basis but not with $p_{k,n}$.

Proof.
Let P be a symmetric polynomial in $A[X_1, X_2 \dots, X_n]$. Permutations do not alter the number of terms or the degree of individual terms. Collect terms of degree d into P_d. $P = P_0 + P_1 + \cdots + P_{\deg(P)}$ where each part can be treated separately. Assume n and d is fixed. Order the X-factors of each term after index and sort the terms lexicographically. An example with $n = 4$, $d = 6$:

$$\prod_{1\leq i<j\leq 4} \big(X_i - X_j\big) = (X_1 - X_2)(X_1 - X_3)(X_1 - X_4)(X_2 - X_3)(X_2 - X_4)(X_3 - X_4)$$

In decreasing lexical order it becomes: $c_1 \underbrace{X_1{}^3 X_2{}^2 X_3{}^1}_{111223} + \cdots + c_m \underbrace{X_2{}^1 X_3{}^2 X_4{}^3}_{233444}$.

For $n = 4$ there will be the following e.s.p's to build a polynomial from.

$e_1 = X_1 + \cdots + X_4$

$e_2 = X_1X_2 + \cdots + X_3X_4$

$e_3 = X_1X_2X_3 + \cdots + X_2X_3X_4$

$e_4 = X_1X_2X_3X_4$

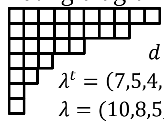
$e_{i_1} e_{i_2} \dots e_{i_k}$

$i_1 \geq i_2 \geq \cdots \geq i_k$

$i_1 + \cdots + i_k = d$

$\lambda^t = (i_1, i_2, \dots, i_k)$

Young diagram

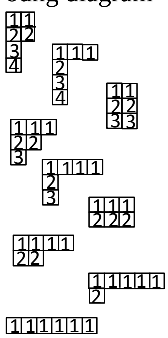


$d = 30$

$\lambda^t = (7,5,4,3,3,2,2,2,1,1)$

$\lambda = (10,8,5,3,2,1,1)$

Products of e.s.p's are parametrized by the indices in weakly decreasing order and the parameter is denoted λ^t (the transpose of λ). The parametrization is a partition of d that is illustrated by block figures called **Young diagrams**. The first column of $e_{\lambda^t} = e_{i_1} \dots e_{i_k}$ has i_1 blocks and the last has i_k blocks.

$\boldsymbol{\lambda^t}$	$\boldsymbol{e}_{\lambda^t}$	Young diagram	$\boldsymbol{\lambda}$ (rows)	Leading term	Lexical order ↓+
(4,2)	e_4e_2		(2,2,1,1)	$X_1^2X_2^2X_3X_4$	$X_1^2X_2^2X_3X_4$
(4,1,1)	$e_4e_1^2$		(3,1,1,1)	$X_1^3X_2X_3X_4$	$X_1^2X_2^2X_3^2$
(3,3)	e_3^2		(2,2,2)	$X_1^2X_2^2X_3^2$	$X_1^3X_2X_3X_4$
(3,2,1)	$e_3e_2e_1$		(3,2,1)	$X_1^3X_2^2X_3$	$X_1^3X_2^2X_3$
(3,1,1,1)	$e_3e_1^3$		(4,1,1)	$X_1^4X_2X_3$	$X_1^3X_2^3$
(2,2,2)	e_2^3		(3,3)	$X_1^3X_2^3$	$X_1^4X_2X_3$
(2,2,1,1)	$e_2^2e_1^2$		(4,2)	$X_1^4X_2^2$	$X_1^4X_2^2$
(2,1,1,1,1)	$e_2^2e_1^4$		(5,1)	$X_1^5X_2$	$X_1^5X_2$
(1,1,1,1,1,1)	e_1^6		(6)	X_1^6	X_1^6

The numbers inside the Young-diagram-columns of the table show which indeterminates occur in the leading terms of $e_{i_1}, \dots, e_{i_k}$. It's clear that $\boldsymbol{e}_{\lambda^t}$ has a leading term $\boldsymbol{X}^\lambda$, indexed by λ in **multi-index notation**. The proof proceeds by induction over lexicographic order of the leading term in P. Since P is symmetric the leading term has weakly decreasing exponents which means that it equals X^λ for some partition of d. If its coefficient is c then $P - c\boldsymbol{e}_{\lambda^t}$ is zero or a symmetric polynomial with strictly smaller leading term on which the induction assumption can be used. P is retrieved as a polynomial in e.s.p's by adding $c\boldsymbol{e}_{\lambda^t}$ to $P - c\boldsymbol{e}_{\lambda^t}$. The uniqueness of this representation follows from the fact that all $\boldsymbol{e}_{\lambda^t}$ has different leading terms. ■

Note that the Vandermonde polynomial used to illustrate the lexicographical ordering is not symmetric. Its square, that occurs in the discriminant is symmetric and can be represented as a polynomial of e.s.p's. Via Vieta's formulas we can replace these with coefficients of the original polynomial. All coefficients c from the proof will be integers so $\Delta \in \mathbb{Z}[a_0, a_1, \dots, a_n]$.

Solutions to Ordinary Differential Equations, ODE

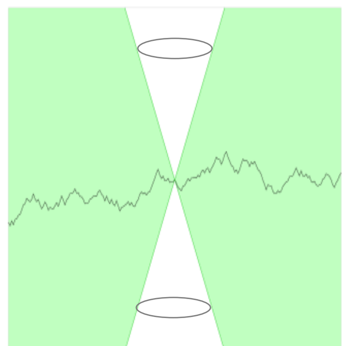
Existence and uniqueness of solutions to $y' = f(x, y)$ with $y(x_0) = y_0$ will depend on $f(x, y)$ which describes the slope space in the (x, y)-plane. y will be assumed to be a vector $\boldsymbol{y} \in \mathbb{R}^n$ and f will be a function $\boldsymbol{f}: \mathbb{R} \times \mathbb{R}^n \to \mathbb{R}^n$ with t as free variable.

$$\dot{\boldsymbol{y}} = \begin{pmatrix} \dot{y}_1 \\ \dot{y}_2 \\ \vdots \\ \dot{y}_n \end{pmatrix} = \begin{pmatrix} f_1(t, y_1, y_2, \dots, y_n) \\ f_2(t, y_1, y_2, \dots, y_n) \\ \vdots \\ f_n(t, y_1, y_2, \dots, y_n) \end{pmatrix} = \boldsymbol{f}(t, \boldsymbol{y}) \quad \boldsymbol{y}(t_0) = \boldsymbol{y}_0$$

Definition 1.

A function $f: X \to Y$ between two metric spaces (X, d_X) and (Y, d_Y) is **Lipschitz continuous** if there is a $K \geq 0$ such that, for all $x_1, x_2 \in X$:

$$d_Y\big(f(x_1), f(x_2)\big) \leq K d_X(x_1, x_2)$$



It sets an upper limit for the slope between two points on the graph. The graph of f is always outside a given double cone with vertex moving along the graph. Functions with bounded first derivatives are Lipschitz continuous.

$\boldsymbol{f}(t, \boldsymbol{y})$ satisfies a **Lipschitz condition** in $\Omega \subseteq \mathbb{R} \times \mathbb{R}^n$ if there is a constant $K \geq 0$ such that:

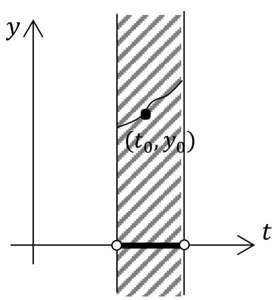
$$(t, \boldsymbol{y}_1), (t, \boldsymbol{y}_2) \in \Omega \Rightarrow |\boldsymbol{f}(t, \boldsymbol{y}_1) - \boldsymbol{f}(t, \boldsymbol{y}_2)| \leq K|\boldsymbol{y}_1 - \boldsymbol{y}_2|$$

A Lipschitz condition is fulfilled if Ω is convex and limited, and $f \in C^1$ in a neighborhood of $\overline{\Omega}$.

Theorem 1. (Picard-Lindelöf theorem)
If $\boldsymbol{f}(t, \boldsymbol{y})$ is continuous and fulfills a Lipschitz condition in the strip $\{(t, \boldsymbol{y}): |t - t_0| \leq a, \boldsymbol{y} \in \mathbb{R}^n\}$ then

$$\dot{\boldsymbol{y}} = \boldsymbol{f}(t, \boldsymbol{y})$$
$$\boldsymbol{y}(t_0) = \boldsymbol{y}_0$$

has a unique solution in the interval $I = |t - t_0| < a$.



Proof.
Rewrite the initial value problem as an integral equation.

$$\begin{matrix}\dot{\boldsymbol{y}} = \boldsymbol{f}(t, \boldsymbol{y}) \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0\end{matrix} \Leftrightarrow \boldsymbol{y}(t) = \boldsymbol{y}(t_0) + \int_{t_0}^{t} \boldsymbol{f}(s, \boldsymbol{y}(s))ds \equiv T[\boldsymbol{y}(t)]$$

The theorem states that the function space operator $T\colon \mathrm{C}(I, \mathbb{R}^n) \to \mathrm{C}(I, \mathbb{R}^n)$ has a unique fixed point that solves $T\boldsymbol{y} = \boldsymbol{y}$. Define a sequence of functions

$$\boldsymbol{y}_0(t) \equiv \boldsymbol{y}_0$$
$$\boldsymbol{y}_{n+1}(t) = T[\boldsymbol{y}_n(t)]$$

$\boldsymbol{y}_n$ will converge uniformly on I to some function $\boldsymbol{y}$ that solves $T\boldsymbol{y} = \boldsymbol{y}$.

Pointwise vs. uniform convergence

Definition 2.
A sequence of functions (f_n) with the same domain and range, $f_n\colon V \to W$ **converges pointwise** to f iff:

$$\forall x \in V\colon \lim_{n\to\infty}\big(f_n(x) - f(x)\big) = 0$$

A sequence of functions (f_n) with the same domain and range, $f_n\colon V \to W$ **converges uniformly** to f iff:

$$\lim_{n\to\infty}(\sup\ \{x \in V\colon \|f_n(x) - f(x)\|\}) = 0$$

Uniform convergence to $f \Rightarrow$ pointwise convergence to f, but not the reverse. $f_n(x) = x^n$ with $\mathrm{D}_{f_n} = [0,1)$ converges pointwise to $f(x) \equiv 0$ but $\sup\{x \in [0,1)\colon |f_n - f|\} = 1$, f_n has no uniformly convergent limit.

Many properties such as continuity and Riemann integrability are transferred to the limit function, but only for uniform limits, not for pointwise limits. The conditions for interchanging limits with differentiation and integration are as follows.

Differentiability: If f_n are differentiable and converges pointwise in $[a, b]$ and $f_n{}'$ converges uniformly to f in $[a, b]$ then f is differentiable with:

$$\mathrm{D}(\lim_{n\to\infty} f_n) = \lim_{n\to\infty}(\mathrm{D}f_n)$$

Integrability: If f_n are Riemann integrable in $[a, b]$ with $f_n \to f$ uniformly then f is Riemann integrable with:

$$\int_a^b \lim_{n\to\infty} f_n dx = \lim_{n\to\infty}\left(\int_a^b f_n dx\right)$$

Weierstrass M-test is a test for uniform convergence of $\sum_{k=1}^{n} f_k$ where f_k has domain A and range in a Banach space, a comple normed vector space, a metric space that is complete for Cauchy sequences. If f_k has bounds $M_k > 0$ s.t. $\forall k \geq 1$, $\forall x \in A$: $\|f_k(x)\| < M_k$ and $\sum_{k=1}^{\infty} M_k < \infty$ then the sequence g_n with $g_n(x) = \sum_{k=1}^{n} f_k(x)$ converges uniformly on A.

$$\boldsymbol{y}_{n+1}(t) - \boldsymbol{y}_0(t) = \sum_{k=0}^{n} \boldsymbol{y}_{k+1}(t) - \boldsymbol{y}_k(t)$$

f continuous and $\bar{I}$ compact $\Rightarrow \sup_{t\in\bar{I}} \|\boldsymbol{f}(t, \boldsymbol{y}_0)\| = L < \infty$

$$\|\boldsymbol{y}_1(t) - \boldsymbol{y}_0(t)\| = \left\| \int_{t_0}^{t} \boldsymbol{f}(s, \boldsymbol{y}_0) ds \right\| \leq L|t - t_0|$$

$$\|\boldsymbol{y}_2(t) - \boldsymbol{y}_1(t)\| = \left\| \int_{t_0}^{t} \boldsymbol{f}(s, \boldsymbol{y}_1) - \boldsymbol{f}(s, \boldsymbol{y}_0) ds \right\| \leq \text{(Lipschitz condition)}$$

$$\leq K \left| \int_{t_0}^{t} \|\boldsymbol{y}_1(s) - \boldsymbol{y}_0(s)\| \, ds \right| \leq$$

$$\leq K \left| \int_{t_0}^{t} L|s - s_0| \, ds \right| = \frac{KL|t - t_0|^2}{2}$$

...

$$\|\boldsymbol{y}_{k+1}(t) - \boldsymbol{y}_k(t)\| \leq \frac{LK^k|t - t_0|^{k+1}}{(k+1)!} \rightarrow \sup_{t\in\bar{I}} \|\boldsymbol{y}_{k+1}(t) - \boldsymbol{y}_k(t)\| \leq \underbrace{\frac{LK^k a^{k+1}}{(k+1)!}}_{M_k}$$

$\sum_{k=1}^{\infty} M_k < \infty$ (Weierstrass Mtest) $\Longrightarrow \boldsymbol{y}_k(t)$ converges uniformly on $\bar{I}$

It remains to show that the limit $\boldsymbol{y}(t) = \lim_{k\to\infty} \boldsymbol{y}_k(t)$ satisfies $T\boldsymbol{y} = \boldsymbol{y}$.

Lipschitz condition $\rightarrow \|f(t, \boldsymbol{y}_n(t) - f(t, \boldsymbol{y}(t))\| \leq K\|\boldsymbol{y}_n(t) - \boldsymbol{y}(t)\|, t \in \bar{I}$

$\sup_{t\in\bar{I}} \|f(t, \boldsymbol{y}_n(t) - f(t, \boldsymbol{y}(t))\| \leq K \sup_{t\in\bar{I}} \|\boldsymbol{y}_n(t) - \boldsymbol{y}(t)\| \Rightarrow$

$\boldsymbol{f}\big(t, \boldsymbol{y}_n(t)\big) \rightarrow \boldsymbol{f}(t, \boldsymbol{y}(t))$ uniformly on $\bar{I}$. (Order of lim and $\int$ can be changed)

$$\boldsymbol{y}_{n+1}(t) = \boldsymbol{y}_0 + \int_{t_0}^{t} \boldsymbol{f}\big(s, \boldsymbol{y}_n(s)\big) ds$$

$$\lim_{n\to\infty} \boldsymbol{y}_{n+1}(t) = \boldsymbol{y}_0 + \lim_{n\to\infty} \int_{t_0}^{t} \boldsymbol{f}\big(s, \boldsymbol{y}_n(s)\big) ds$$

$$\boldsymbol{y}(t) = \boldsymbol{y}_0 + \int_{t_0}^{t} \underbrace{\lim_{n\to\infty} \boldsymbol{f}\big(s, \boldsymbol{y}_n(s)\big)}_{\boldsymbol{f}(s, \boldsymbol{y}(s))} ds$$

$\therefore \dot{\boldsymbol{y}} = \boldsymbol{f}(t, \boldsymbol{y}), \boldsymbol{y}(t_0) = \boldsymbol{y}_0$ has a solution given by $\boldsymbol{y}(t) = \lim_{n\to\infty} \boldsymbol{y}_n(t)$

Assume $\dot{\boldsymbol{y}} = \boldsymbol{f}(t,\boldsymbol{y}), \boldsymbol{y}(t_0) = \boldsymbol{y}_0$ has two solutions, $\boldsymbol{y}(t)$ and $\boldsymbol{z}(t)$ and $t \geq t_0$.

$$\|\boldsymbol{y}(t) - \boldsymbol{z}(t)\| = \left\|\int_{t_0}^{t} \boldsymbol{f}(s,\boldsymbol{y}(s) - \boldsymbol{f}(s,\boldsymbol{z}(s))\right\| \leq \underbrace{K\int_{t_0}^{t} \|\boldsymbol{y}(s) - \boldsymbol{z}(s)\| ds}_{w(t)}$$

$$\left.\begin{array}{l} w'(t) = K\|\boldsymbol{y}(t) - \boldsymbol{z}(t)\| \\ \|\boldsymbol{y}(t) - \boldsymbol{z}(t)\| \leq w(t) \end{array}\right. \Rightarrow w'(t) \leq Kw(t) \Rightarrow \frac{d}{dt}\left(e^{-Kt}w(t)\right) \leq 0 \Rightarrow$$

$e^{-Kt}w(t)$ decreasing $\Rightarrow 0 \leq$

$e^{-Kt}w(t) \leq e^{-Kt_0}w(t_0) = 0 \Rightarrow$

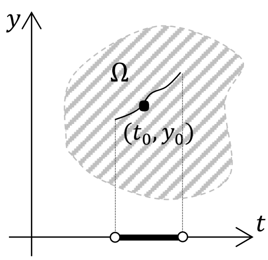
$w(t) \equiv 0$ for $t_0 \leq t < t_0 + a$ $\Rightarrow$
same goes for $t_0 - a < t \leq t_0$

$\boldsymbol{y}(t) = \boldsymbol{z}(t)$ for every $t \in I$ ■

Theorem 2.

If f is continuous in a neighborhood Ω of $(t_0, \boldsymbol{y}_0)$ and fulfills a Lipschitz condition in Ω then there is an open interval around t_0 with a unique solution to:

$$\begin{array}{l} \dot{\boldsymbol{y}} = \boldsymbol{f}(t,\boldsymbol{y}) \\ \boldsymbol{y}(t_0) = \boldsymbol{y}_0 \end{array}$$



Proof.

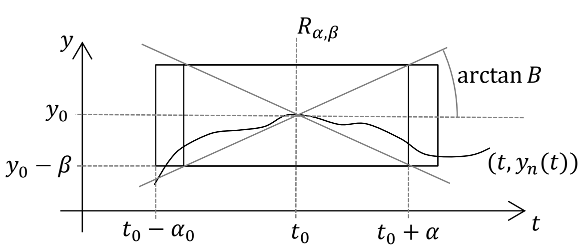
Choose an area $R_{\alpha_0,\beta} = \{(t,\boldsymbol{y}) \in \mathbb{R} \times \mathbb{R}^n : |t - t_0| \leq \alpha_0, \|\boldsymbol{y} - \boldsymbol{y}_0\| \leq \beta\}$ s.t. f satisfies a Lipschitz condition in $R_{\alpha_0,\beta}$.
Let $B = \sup\|\boldsymbol{f}(t,\boldsymbol{y}(t))\|$ over $R_{\alpha_0,\beta}$ and $\alpha = \min(\alpha_0, \beta/B)$ then the proof of theorem 1 works as long as all curves $(t,\boldsymbol{y}_n(t))_{n=1}^{\infty}$ are within $R_{\alpha,\beta}$.

Proof by induction over n:
$\boldsymbol{y}_0(t) \equiv \boldsymbol{y}_0 \Rightarrow \boldsymbol{y}_0(t)$ is inside $R_{\alpha,\beta}$

$$\|\boldsymbol{y}_{n+1}(t) - \boldsymbol{y}_0(t)\| = \left\|\int_{t_0}^{t} \boldsymbol{f}(s,\boldsymbol{y}_n(s))ds\right\| \leq \left|\int_{t_0}^{t} Bds\right| = B|t - t_0| \leq \beta$$

(Induction assumption $\boldsymbol{y}_n(s)$ within $R_{\alpha,\beta}$ is used in the first inequality)

Uniqueness of the solution is shown in the same way as for theorem 1. ■



Maximally extended solutions

Let $\boldsymbol{f}(t, \boldsymbol{y})$ be continuous and satisfy a Lipschitz condition in a neighborhood of every point in a domain Ω with $(t_0, y_0) \in \Omega$. Theorem 2 gives an interval $I = |t - t_0| < \alpha$ with a unique solution curve inside Ω. By picking two new points within I close to its endpoints and using the method in the proof of theorem 2 extends the solution to the left and right.

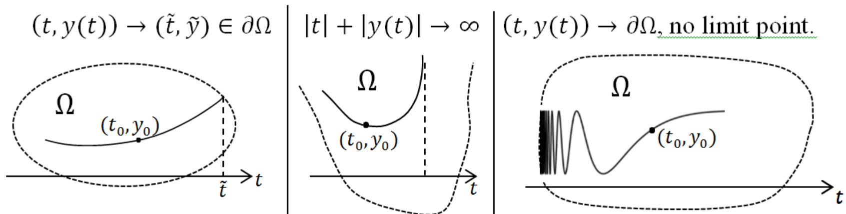
$t^+ \equiv \sup\{t|\text{solution curve can be extended to } [\mathrm{t}_0, t)\}$, possibly $t^+ = +\infty$
$t^- \equiv \inf\{t|\text{solution curve can be extended to } (\mathrm{t}, \mathrm{t}_0]\ \}$, possibly $t^- = -\infty$

A solution extended to (t^-, t^+) is called **maximal.**

Theorem 3.

If K is a compact subset of Ω then a maximal solution $(t, \boldsymbol{y}(t))$ will leave K when t approaches t^- or t^+.

Three cases can occur:
I. $(t, \boldsymbol{y}(t))$ approaches a point on $\partial\Omega$.
II. $|t| + \|\boldsymbol{y}(t)\| \to \infty$.
III. $\big(t, \boldsymbol{y}(t)\big) \to \partial\Omega$ without approaching a specific point on $\partial\Omega$.



Proof.

Assume the opposite, then there is a compact subset $K \subset \Omega$ and a sequence $t_1, t_2, \ldots$ with t^+ as limit (or t^-) s.t. $\forall j: P_j = (t_j, \boldsymbol{y}(t_j)) \in K$. Bolzano-Weierstrass theorem gives this sequence a limit point $P = (t^+, \boldsymbol{y}^+) \in K$. P is an inner point of Ω and it is the center of an area $R_{\alpha_0,\beta}$ contained in Ω. Pick a subsequence P_{j_k} of P_j that converges toward P then if k is big enough $P_{j_k} \in R_{\alpha_0,\beta}$. From P_{j_k} the solution curve can be extended to at least t^* where the cone centered at P_{j_k} with opening angle $2\arctan(B)$ first crosses $R_{\alpha_0,\beta}$. The angle doesn't change $B = \sup\{\|f(t, \boldsymbol{y})\| : (t, \boldsymbol{y}) \in R_{\alpha_0,\beta}\}$. As $P_{j_k} \to P$, the point t^* will extend beyond t^+. The assumption must be wrong. ■