الفصل الأوّل

تمثيل الأعداد ومفهوم الاستقراريّة في التحليل العددي

كان مصنّعيّ الأجهزة الحاسوبيّة، في أعوام الستينات والسبعينات، يطوّرون -كلٌّ على طريقته- نظام تمثيل الأعداد بالنقطة العائمة مما جعل توافقية البرمجيات مع هذه الأجهزة أمراً عسيراً. فعلى سبيل المثال، كانت معظم الآلات الحاسبة تعتمد النظام الثنائي في تمثيل الأعداد في حين كانت حواسيب IBM 360/370 (والتي كانت شائعة في تلك الفترة) تعتمد النظام الست عشري hexadecimal وفيما كانت أنظمة حاسوبيّة أخرى مثل أنظمة HP الحاسبة تعتمد النظام العشري.

ساهمت جهود العديدين من اختصاصييّ الأجهزة الحاسوبيّة، في بداية أعوام الثمانينات من أمثال W. Kahan في وضع نظامٍ قياسي لتمثيل الأعداد بالنقطة العائمة وتلى ذلك دفعٌ كبيّر لهذا النظام بتبنّي كلّ من شركة Intel وشركة Motorola لهذا المعيـــار في تمثيل الأعداد. عُرِفَ هذا النظام القياسي في التمثيل باسم معيـــار IEEE في تمثيل الأعداد بالنقطة العائمة وكان ذلك منذ أن بدأ بتطويره فريقٌ متخصص من معهد الهندسة الكهربائية والالكترونية Institute for Elecrical and Electronics Engineers.

1. التمثيل بطريقة النقطة العائمة

ليكن $2 \le \beta \in \mathbb{N}$ عدداً طبيعياً نسميّه الأساس (غالباً مايكون زوجياً)، عندئذ، أياً كان العدد الحقيقي $x \in \mathbb{R} \setminus \{0\}$ فإنّه يوجد عددٌ صحيحٌ $e \in \mathbb{Z}$ بحيث يكون $\beta^{e} \le |x| < \beta^{e+1}$ وهذا يعني أنّه يوجد عددٌ وحيد $0 \le \lambda < 1$ بحيث يكون

$$|x| = (1-\lambda)\cdot\beta^{e} + \lambda\cdot\beta^{e+1} = \big(1+\lambda\cdot(\beta-1)\big)\cdot\beta^{e}$$

أي أنّ العدد x يكتب بطريقة وحيدة بالشكل

$$x = \pm m\times\beta^{e},\quad 1\le m<\beta,\quad e\in\mathbb{Z}.$$

نسمي العدد m دليل العدد x .

تعتمد طريقة التمثيل بالنقطة العائمة على اختيار عدد صحيح $1\le p$ يسمّى دقّة التمثيل precision إضافةً إلى عددين آخرين $e_{\min}$ و $e_{\max}$ يحدّدان مجال تغيّر الأس e بحيث نستطيع بهذه الطريقة تمثيل الأعداد التالية

$$\pm\underbrace{\left(d_0 + d_1\cdot\beta^{-1}+\cdots+d_{p-1}\cdot\beta^{-(p-1)}\right)}_{m}\times\beta^{e},\qquad 0\le d_i<\beta \tag{1}$$

يعبّر المصطلح عدد النقطة العائمة floating-point number عن العدد الحقيقي الذي يكتب بالشكل (1). لتخزين عدد النقطة العائمة تقسم وحدة الذاكرة computer word المخصّصة لذلك إلى ثلاثة حقول تمثّل على الترتيب الإشارة s، الأس e ومن ثمّ الدليل m. فمثلاً، في نظام حاسوبي يعتمد وحدة ذاكرة قياسها $32-\text{bit}$ يمكن توزيع هذه الحقول كالتالي: 1 bit لتمثيل الإشارة، 8 bits لتمثيل الأس و 23 bits لتمثيل الدليل.

هناك حالتان لا يمكن فيهما تمثيل عدد حقيقي تمثيلاً تاماً كعدد من أعداد النقطة العائمة. الحالة الأولى، وهي الأكثر شيوعاً، يمكن إيضاحها بالعدد العشري 0.1 . فهذا العدد يمتلك تمثيلاً عشرياً منتهياً $(\beta=10)$:

$$0.1 = +1.0\times10^{-1}$$

وفي النظام الثنائي $(\beta=2)$ يقع هذا العدد تماماً بين عددي نقطة عائمة وذلك لكونه لا يقبل تمثيلاً منتهياً:

$$0.1 = 1.10011001100\cdots\times2^{-4}$$

والحالة الثانية، وهي أقلّ حدوثاً وتتعلّق بالأعداد الواقعة خارج مجال التمثيل، أي أن تكون القيمة المطلقة للعدد أكبر من القيمة $\beta\times\beta^{e_{\max}}$ أو أصغر من القيمة $1.0\times\beta^{e_{\min}}$.

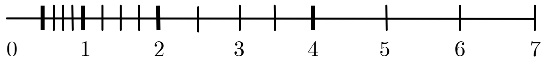
إنّ تمثيل الأعداد بطريقة النقطة العائمة ليس وحيداً، فعلى سبيل المثال، كلاً من الكتابتين: 0.01×10^{1} و 1.00×10^{-1} تمثّل العدد 0.1 . لكي نجعل تمثيل الأعداد الحقيقيّة بطريقة النقطة العائمة **وحيداً** نشترط في هذا التمثيل أن يكون الرقم الأوّل من يسار دليل العدد d_0 **غير معدوم** وعندئذٍ يأخذ تمثيل العدد الشكل التالي:

$$\pm d_0\cdot d_1d_2\cdots d_{p-1}\times\beta^{e},\quad 0\le d_i<\beta,\quad d_0>0. \tag{2}$$

نسميّ طريقة التمثيل (2) طريقة **التمثيل بالنقطة العائمة القياسيّة** Normalized floating-point ، كما نسمّي الأعداد التي تكتب وفق العبارة (2) **أعداداً قياسيّة** Normalized numbers. فعلى سبيل المثال الكتابة 1.00×10^{-1} قياسيّة في حين أنّ الكتابة 0.01×10^{1} ليست كذلك.

مثـــال:

في حالة $\beta = 2$ ، $p = 3$ ، $e_{min} = -1$ و $e_{max} = 2$ لدينا 16 عدداً قياسيّاً موجباً (الإشارة +):



الشكل 1: الأعداد القياسيّة الموجبة عندما $\beta = 2$، $p = 3$،
$e_{min} = -1$ و $e_{max} = 2$.

وهذه الأعداد هي:

$$\bigcup_{i=-1}^{2}\left\{1.00\times 2^{i},\ 1.01\times 2^{i},\ 1.10\times 2^{i},\ 1.11\times 2^{i}\right\}$$

لسوء الحظ، فإنّ الشرط الذي وضعناه ليصبح التمثيل بالنقطة العائمة قياسياً، لا يسمح بتمثيل الصفر!. سنرى في الفقرة التالية كيفيّة تمثيل الصفر وكذلك أعداداً أخرى لا يمكن كتابتها وفق العبارة (2).

2. معيـــار IEEE في تمثيــل الأعداد

في عام 1985 ، جرى اعتماد معيار "IEEE 754" لتمثيل الأعداد بالنقطة العائمة، وهو نظام تمثيلٍ يقوم حصراً على استخدام النظام الثنائي $(\beta = 2)$ في التعبير عن مكوّنات التمثيل: الأس exponent والدليل significand. بعد ذلك بأعوامٍ قليلة، وفي عام 1987 جرى تعميم هذا المعيار إلى "IEEE 854" والذي يسمح باستخدام أحد الأساسين $\beta = 2$ أو $\beta = 10$ في التمثيل على حدٍ سواء.

نناقش في هذا الفصل النموذج الحسابي الذي يحدّده المعيار ANSI/IEEE Standard 754-1985 أو اختصاراً "IEEE 754". ننوّه هنا إلى أنّ جميع معالجات SPARC و x86 تستعمل نظام IEEE الحسابي.

يشتمل معيـــار IEEE في تمثيل الأعداد على ثلاثة أنماط قياسيّة مصنّفةٌ تبعاً للدقّة التي يتيحها كلّ نمط:

- نمط الدقّة البسيطة single precision،
- نمط الدقّة المضاعفة double precision،
- نمط الدقّة الموسّعة extended precision.

- **أولاً: نمـــط الدقّـــة البسيطة**

تتألّف هذه الصيغة من ثلاثة حقول: حقل مكوّن من 23-bit يمثّل الجزء الكسري $0 \le f < 1$؛ حقلٌ مكوّن من 8-bit لتمثيل القيمة المنحازة للأس e وأخيراً، حقلٌ مكوّن من 1-bit لتمثيل الإشارة $s \in \{0,1\}$. تُخزّن هذه الحقول متجاورةً في وحدة ذاكرةٍ قياسها 32-bit كما في الشكل التالي.

	msb	lsb	msb	lsb
s	e [30:23]		f [22:0]	
31	30	23	22	0

الشكل 2: صيغة التخزين البسيطة.

يبيّن الجدول التالي الحالات المختلفة للحقول الثلاثة s ، e و f وما يقابلها من قيمٍ عدديّة تمثّلها صيغة التمثيل البسيطة.

القيمة العدديّة	صيغة التخزين البسيطة
$(-1)^s \times 1.f \times 2^{e-127}$ normal numbers	$0<e<255$
$(-1)^s \times 0.f \times 2^{-126}$ subnormal numbers	$e=0; f \neq 0$
$(-1)^s \times 0.0$ (signed zero)	$e=0; f=0$
$+$INF $(+\infty)$	$s=0; e=255; f=0$
$-$INF $(-\infty)$	$s=1; e=255; f=0$
NaN (Not-a-Number)	$e=255; f \neq 0$

الجدول 1: القيم العدديّة التي يمثّلها نمط الدقّة البسيطة في معيار IEEE القياسي.

نلاحظ هنا أنّ الدليل m للعدد الممثّل بهذه الصيغة هو عددٌ مركّب يساوي:

- $m=1.f$ عندما $0<e<255$. نسمّي العدد الذي تمثّله هذه الصيغة –في هذه الحالة– **عدداً نظامياً** normal number.
- $m=0.f$ عندما $e=0$ و $f \neq 0$. نسمّي العدد الذي تمثّله هذه الصيغة –في هذه الحالة– **عدداً تحت نظامي** subnormal number.

يتبيّن لنا من ذلك أنّ دليل الصيغة البسيطة m يمثّل بجزءٍ كسري قيمته $0.f$ وجزءٍ صحيح قيمته 1 في حالة الأعداد النظاميّة، و 0 في حالة الأعداد تحت النظاميّة.

يبيّن الجدول التالي بعض صيغ التمثيل البسيطة الهامّة وقيمها بالنظام العشري.

القيمة العدديّة العشرية	صيغة التخزين (Hex)	الاسم الشائع
0.0	00000000	+0
-0.0	80000000	-0
1.0	3F800000	1
2.0	40000000	2
3.40282347e+38	7F7FFFFF	maximum normal number
1.17549435e-38	00800000	minimum positive normal number
1.17549421e-38	007FFFFF	maximum subnormal number
1.40129846e-45	00000001	minimum positive subnormal number
Infinity	7F800000	$+\infty$
-Infinity	FF800000	$-\infty$
NaN	7FC00000	Not-a-Number

الجدول 2: بعض صيغ التمثيل بالدقّة البسيطة وقيمها العدديّة.

- **ثانيـــاً: نمـــط الدقّـــة المضــــاعفة**

تتألّف هذه الصيغة من ثلاثة حقول: حقل مكوّن من 52-bit يمثّل الجزء الكسري $0 \le \mathrm{f} < 1$؛ حقلٌ مكوّن من 11-bit لتمثيل القيمة المنحازة للأس e وأخيراً، حقلٌ مكوّن من 1-bit لتمثيل الإشارة $\mathrm{s} \in \{0,1\}$. تُخَزّن هذه الحقول متجاورةً في وحدة ذاكرةٍ قياسها 64-bit كما في الشكل التالي.

	msb	lsb	msb	lsb
s	e [62:52]		f [51:0]	
63	62	52	51	0

الشكل 3: صيغة التخزين المضاعفة.

يبيّن الجدول التالي الحالات المختلفة للحقول الثلاثة s ، e و f وما يقابلها من قيمٍ عدديّة تمثّلها صيغة التمثيل المضاعفة.

صيغة التخزين المضاعفة	القيمة العدديّة
$0<\mathrm{e}<2047$	$(-1)^{\mathrm{s}} \times 1.\mathrm{f} \times 2^{\mathrm{e}-1023}$ normal numbers
$\mathrm{e}=0;\ \mathrm{f} \neq 0$	$(-1)^{\mathrm{s}} \times 0.\mathrm{f} \times 2^{-1022}$ subnormal numbers
$\mathrm{e}=0;\ \mathrm{f}=0$	$(-1)^{\mathrm{s}} \times 0.0$ (signed zero)
$\mathrm{s}=0;\ \mathrm{e}=2047;\ \mathrm{f}=0$	$+\mathrm{INF}\ (+\infty)$
$\mathrm{s}=1;\ \mathrm{e}=2047;\ \mathrm{f}=0$	$-\mathrm{INF}\ (-\infty)$
$\mathrm{e}=2047;\ \mathrm{f} \neq 0$	NaN (Not-a-Number)

الجدول 3: القيم العدديّة التي يمثّلها نمط الدقّة المضاعفة في معيار IEEE القياسي.

نلاحظ هنا - أيضاً - أنّ الدليل m للعدد الممثّل بهذه الصيغة هو عددٌ مركّب يساوي:

- $m = 1.\mathrm{f}$ عندما $0 < \mathrm{e} < 2047$. نسمّي العدد الذي تمثّله هذه الصيغة –في هذه الحالة– **عدداً نظامياً** normal number.
- $m = 0.\mathrm{f}$ عندما $\mathrm{e} = 0$ و $\mathrm{f} \neq 0$. نسمّي العدد الذي تمثّله هذه الصيغة –في هذه الحالة– **عدداً تحت نظامي** subnormal number.

يتبيّن لنا من ذلك أنّ دليل الصيغة المضاعفة m يمثّل بجزءٍ كسري قيمته $0.\mathrm{f}$ وجزءٍ صحيح قيمته 1 في حالة الأعداد النظاميّة، و 0 في حالة الأعداد تحت النظاميّة.

يبيّن الجدول التالي بعض صيغ التمثيل المضاعفة الهامّة وقيمها بالنظام العشري.

القيمة العدديّة العشرية	صيغة التخزين (Hex)	الاسم الشائع
0.0	00000000 00000000	+0
-0.0	80000000 00000000	-0
1.0	3FF00000 00000000	1
2.0	40000000 00000000	2
1.7976931348623157e+308	7FEFFFFF FFFFFFFF	maximum normal number
2.2250738585072014e-308	00100000 00000000	minimum positive normal number
2.2250738585072009e-308	000FFFFF FFFFFFFF	maximum subnormal number
4.9406564584124654e-324	00000000 00000001	minimum positive subnormal number
Infinity	7FF00000 00000000	$+\infty$
-Infinity	FFF00000 00000000	$-\infty$
NaN	7FF80000 00000000	Not-a-Number

الجدول 4: بعض صيغ التمثيل بالدقّة المضاعفة وقيمها العدديّة.

- **ثالثـــاً: نمـــط الدقّـــة الموسّعة**

تتألّف هذه الصيغة من أربعة حقول: حقل مكوّن من 63-bit يمثّل الجزء الكسري $0 \le f < 1$؛ حقلٌ مكوّن من 1-bit يمثّل الجزء الصحيح من دليل العدد $i \in \{0,1\}$؛حقلٌ مكوّن من 15-bit لتمثيل القيمة المنحازة للأس e وأخيراً، حقلٌ مكوّن من 1-bit لتمثيل الإشارة $s \in \{0,1\}$. تُخَزّن هذه الحقول متجاورةً في وحدة ذاكرة قياسها 80-bit كما في الشكل التالي.

	msb … lsb		msb … lsb
s	e [78:64]	i	f [62:0]
79	78 … 64		62 … 0

الشكل 4: صيغة التخزين الموسّعة.

يبيّن الجدول التالي الحالات المختلفة للحقول الأربعة s ، e ، i و f وما يقابلها من قيمٍ عدديّة تمثّلها صيغة التمثيل المضاعفة.

القيمة العدديّة	صيغة التخزين الموسّعة
Unsupported	0<e<32767; i=0
$(-1)^s \times 1.f \times 2^{e-16383}$ normal numbers	0<e<32767; i=1
$(-1)^s \times 0.f \times 2^{-16382}$ subnormal numbers	$e=0;\ i=0;\ f \neq 0$
$(-1)^s \times 1.f \times 2^{-16382}$ pseudo- denormal numbers	$e=0;\ i=1$
$(-1)^s \times 0.0$ (signed zero)	$e=0;\ i=0;\ f=0$
$+$INF $(+\infty)$	$s=0; e=32767; i=1; f=0$
$-$INF $(-\infty)$	$s=1; e=32767; i=1; f=0$
NaN (Not-a-Number)	$e=32767;\ f \neq 0$

الجدول 5: القيم العدديّة التي يمثّلها نمط الدقّة الموسّعة في معيار IEEE القياسي.

يبيّن الجدول التالي بعض صيغ التمثيل المضاعفة الهامّة وقيمها بالنظام العشري.

القيمة العدديّة العشرية	صيغة التخزين x86 (Hex)	الاسم الشائع
0.0	0000 00000000 00000000	+0
-0.0	8000 00000000 00000000	-0
1.0	3FFF 80000000 00000000	1
2.0	4000 80000000 00000000	2
1.18973149535723176505e+4932	7FFE FFFFFFFF FFFFFFFF	maximum normal number
3.36210314311209350626e-4932	0001 80000000 00000000	minimum positive normal number
3.36210314311209350608e-4932	0000 7FFFFFFF FFFFFFFF	maximum subnormal number
3.64519953188247460253e-4951	0000 0000000 00000001	minimum positive subnormal number
Infinity	7FFF 80000000 00000000	$+\infty$
-Infinity	FFFF 80000000 00000000	$-\infty$
NaN	7FFF FFFFFFFF FFFFFFFF	Not-a-Number

3. تدوير الأعداد في نظام IEEE المعياري

يحسب ،في نظام IEEE المعياري، ناتج أيّ عمليّة حسابيّة بدّقة ومن ثّم يجري تدوير الناتج (أي اختيار ممّثل له من بين أعداد الآلة).

بوجه عام، يقصد بعمليّة تدوير عدد حقيقي $x \in \mathbb{R}$ اختيار أقرب أعداد الآلة إليه Round toward nearest. يتطلّب نظام IEEE المعياري تعريف ثلاثة أنماط أخرى من عمليات التدوير:

- التدوير باتجاه الصفر Round toward 0،
- التدوير باتجاه الـ $+\infty$ Round toward،
- التدوير باتجاه الـ $-\infty$ Round toward.

يشتمل نظام IEEE المعياري على مجموعة من المؤشرات التي تدلّ على الحالات الاستثنائيّة في الحساب العددي منها:

- ظاهرة نقص الأسْ underflow،
- ظاهرة زيادة الأسْ overflow،
- القسمة على الصفر division by zero،
- العمليات الحسابيّة غير الصحيحة invalid operation.

بما أنّ تدوير الأعداد يدخل في صميم الحساب بالنقطة العائمة فإنّه من المهمّ لنا ايجاد طريقة لقياس خطأ التدويرعند تنفيذ عمليّة حسابيّة ما. فعلى سبيل المثال، عندما نستخدم نظام تمثيل فيه $\beta = 10$ ، $p = 3$ وكانت نتيجة عمليّة حسابيّة ممّثلة بهذا النظام 3.12×10^{-2} وكان الناتج عندما نجري العمليّة بدقّة لا نهائيّة مساوياً 0.0314، يكون من الواضح أنّ هناك خطأً بمقدار وحدتين من **مرتبة آخر رقم** (هنا 10^{-4}) من أرقام تمثيل الناتج بهذا النظام. بشكلٍ مشابه، يمثّل العدد الحقيقي 0.0314159 في هذا النظام بالعدد 3.14×10^{-2} وعندها يكون خطأ التمثيل مساوياً 0.159 وحدة من **مرتبة آخر رقم** (هنا 10^{-4}).

وبوجهٍ عام، إذا كان $d_0.d_1 \ldots d_{p-1} \times \beta^e$ التمثيل بالنقطة العائمة للعدد الحقيقي x فإنّ الخطأ المرتكب في هذا التمثيل يساوي

$$\left| d_0.d_1 \ldots d_{p-1} - x / \beta^e \right| \cdot \beta^{p-1}$$

وحدةً من **مرتبة آخر رقم** (هنا $\beta^{e-(p-1)}$).

يستعمل عادةً مصطلح **وحدة آخر رقم** unit in the last place واختصاراً *ulp*. فمثلاً: في نظام التمثيل المبيّن بالشكل 1 لدينـــا

$$ulp(1.11 \times 2^{-1}) = 2^{-1-2} = 1/8, \quad ulp(1.00 \times 2^{0}) = 2^{-2} = 1/4,$$
$$ulp(1.10 \times 2^{1}) = 2^{1-2} = 1/2, \quad ulp(1.01 \times 2^{2}) = 2^{2-2} = 1.$$

4. الحالات الخاصّة في نظام IEEE المعياري

رأينا سابقاً أنّ نظام IEEE 754 المعياري يشتمل على مجموعة من القيم الخاصّة (انظر الجدول التالي) التي نحتاجها لتمثيل الناتج عند إجراء العمليّات الحسابيّة. تمتلك جميع هذه القيم تمثيلاً تكون فيه قيمة الأس مساوية لإحدى القيمتين $e_{\min}-1$ أو $e_{\max}+1$.

Exponent	*Fraction*	*Represents*
$e=e_{min}-1$	$f=0$	± 0
$e=e_{min}-1$	$f\neq 0$	$0.f\times 2^{e_{\min}}$
$e_{min}\le e\le e_{\max}$	$-$	$1.f\times 2^{e}$
$e=e_{min}+1$	$f=0$	∞
$e=e_{min}+1$	$f\neq 0$	NaN

القيم الخاصّة في نظام IEEE 754.

- **القيمة NaN**

جرت العادة عند حساب النسبة $0/0$ أو المقدار $\sqrt{-1}$ أنْ يوقف الحساب بخطأ غير قابل للمعالجة unrecoverable error. لكن، هناك العديد من الحالات التي يكون فيها متابعة الحساب مفيداً. يمكن تجنّب هذه المشكلة بتعريف قيمة خاصّة تسمّى NaN واعتبار ناتج كلّ من العمليّتين $0/0$ و $\sqrt{-1}$ مساوياً NaN بدلاً من التوقّف. يلخّص الجدول التالي عدداً من الحالات التي يكون الناتج فيها NaN:

Operation	NaN *Produced By*
$+$	$\infty+(-\infty)$
$\times$	$0\times\infty$
$/$	$0/0\ ,\ \infty/\infty$
$\sqrt{\ }$	$\sqrt{x},\ when\ x<0$

- **القيمة ∞ أو اللانهاية**

يسمح تعريف القيمة ∞ بمتابعة الحساب عند حدوث زيادة في الأسْ overflow. يعتبر هذا الخيار أفضل من إرجاع القيمة الموافقة لأكبر عدد قابل للتمثيل. فمثلاً على سبيل المثال، عند حساب قيمة المقدار $\sqrt{x^2+y^2}$ في نظام تمثيل للأعداد فيه

$$\beta=10,\quad p=3,\quad e_{\max}=98.$$

فإذا كان $x=3\times 10^{70}$ و $y=4\times 10^{70}$ فإنّ حساب قيمة x^2 سوف تحدث زيادةً في الأسْ ولنفترض أنّ الناتج قد استبدل بالقيمة العظمى في هذا النظام وهي 9.99×10^{98}. بوجه مشابه تحدث زيادةً في الأسْ عند حساب y^2. وكذلك

تحدث زيادةً في الأسْ عند حساب المجموع x^2+y^2 ومن ثمّ يكون الناتج أيضاً 9.99×10^{98}. وأخيراً، تكون نتيجة حساب المقدار $\sqrt{x^2+y^2}$ مساويةً

$$\sqrt{9.99\times10^{98}}=3.16\times10^{49}$$

ومن الواضح أنّ الخطأ في هذه النتيجة فظيعٌ جداً !!! والجواب الصحيح هو وضوحاً 5.00×10^{70}.

في نظام IEEE الحسابي تكون نتيجة حساب x^2 مساويةً للقيمة ∞ وكذلك الأمر بالنسبة لنتيجة حساب y^2 و $\sqrt{x^2+y^2}$. ومن ثمّ تكون النتيجة النهائيّة ∞. من الواضح أنّ إعطاء هذه القيمة لنتيجة الحساب آمن من سابقتها.

من الأمثلة على الحالات التي يكون الناتج فيها مساوياً ∞ ما يلـــي:

$$1/0=\infty,\quad -1/0=-\infty,\quad \sqrt{\infty}=\infty.$$

- **الصفر مع الإشارة ±0**

يمثّل الصفر في نظام IEEE بجزءٍ كسري معدوم $f\equiv0$ وأسٍ قيمته $e_{\min}-1$. وبما أنّ إشارة التمثيل sign يمكن أن تأخذ قيمتين فإنّه يكون هناك تمثيلين للصفر $+0$ و -0. في مثل هذا الوضع يجب الحذر من طريقة التعامل مع هذا التمثيل، فمثلاً اختبار المقارنة $if\,(x=0)$ يعتمد على إشارة x ولا يمكن في هذه الحالة إعطاء نتيجة دون وضع قاعدة توضّح العلاقة بين تمثيلي الصفر. فنظام IEEE المعياري يعرّف المقارنة بحيث يكون $+0=-0$ بدلاً من $+0>-0$.

من جهة أخرى، يمكن للمرء أنْ يفكّر بإهمال الإشارة عند تمثيل الصفر، ولكنّ هذا أيضاً ينطوي على إشكالات، ونظام IEEE المعياري لا يفعل ذلك، لأنّ إشارة الصفر تفيد في تحديد إشارة ناتج حساب المقادير العدديّة. فمثلاً، تبطل صلاحيّة المطابقة $x=1/(1/x)$ عندما $x=\pm\infty$. والسبب في ذلك يعود إلى كون ناتج كلّ من $1/-\infty$ و $1/+\infty$ هو الصفر وناتج $1/0$ هو $+\infty$. وبذا تضيع معلومة إشارة الناتج.

مثالاً آخر على فائدة إشارة الصفر في حالة نقص الأسْ underflow والتوابع التي لها انقطاعٌ عند الصفر مثل التابع اللوغاريتمي log. في نظام IEEE المعياري، يعرّف $\log0=-\infty$ و $\log x=\mathrm{NaN}$ عندما $x<0$. لنفترض x يمثّل عدداً سالباً صغيراً بما يكفي ليحدث نقصاً في الأسْ فيدوّر إلى الصفر. عندها ولحسن الحظ أنّ إشارة x سالبة ويكون ناتج حساب قيمة تابع اللوغاريتم NaN.

- **الأعداد تحت النظاميّة Subnormal**

لننظر في حالة العددين $x=6.87\times10^{-97}$ و $y=6.81\times10^{-97}$ في نظام تمثيل بالنقطة العائمة وسطاؤه: $\beta=10$، $p=3$ و $e_{\min}=-98$. من الواضح أنّ هذين العددين نظاميين تماماً وكلّ منهما يكبر أصغر الأعداد النظاميّة الممثّلة بهذا النظام 1.00×10^{-98}. مع ذلك، نلاحظ وجود ظاهرة غريبة يتمتّع بها هذان العددان : $x-^*y=0$ في حين أنّ $x\neq y$. والسبب في ذلك يعود إلى كون $x-y=0.06\times10^{-97}=6.0\times10^{-99}$ وهذا الفرق يمثّل عدداً صغيراً لا يمكن تمثيله بعددٍ نظامي normal number ولهذا يدوّر هذا الفرق إلى الصفر.

يتبيّن من ذلك أهميّة أنْ يحافظ نظام التمثيل على صحّة الخاصّة التالية

$$x = y \quad \Leftrightarrow \quad x - y = 0 \tag{3}$$

يضمن نظام التمثيل المعياري IEEE 754 وكذلك IEEE 854 الخاصّة (3) إضافةً إلى العديد من الخواص الأخرى من خلال توسيع مجموعة الأعداد النظاميّة الممثّلة بهذا النظام لتشمل أعداداً واقعة بين الصفر وأصغر الأعداد النظاميّة. تسمّى هذه الأعداد denormalized numbers في نظام IEEE 754 وأعيد تسميتها في نظام IEEE 854 لتصبح subnormal.

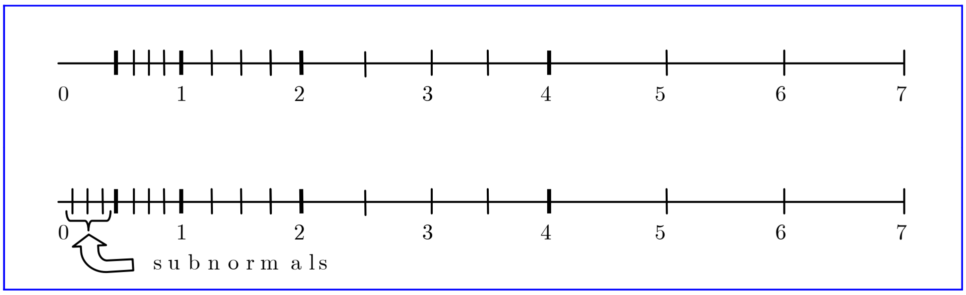
مثـــال: في حالة نظام تمثيل بالنقطة العائمة القياسيّة وسطاؤه $\beta = 2$، $p = 3$ ، $e_{\min} = -1$ و $\mathrm{e}_{\max} = 2$ يكون لدينا (انظر الشكل 5):

- مجموعة الأعداد النظامـــيّة: وتشمل

$$\mathcal{N} = \bigcup_{i=-1}^{2} \left\{1.00 \times 2^i,\ 1.01 \times 2^i,\ 1.10 \times 2^i,\ 1.11 \times 2^i\right\}$$

- مجموعة الأعداد تحت النظامـــيّة وهي

$$\mathcal{S} = \left\{0.01 \times 2^{-1},\ 0.10 \times 2^{-1},\ 0.11 \times 2^{-1}\right\}$$



الشكل 5: الأعداد القياسيّة الموجبة: النظاميّة وتحت النظاميّة في حالة نظام التمثيل IEEE المعياري الموافق للوسطاء
$\beta = 2$، $p = 3$، $\mathrm{e}_{\min} = -1$ و $\mathrm{e}_{\max} = 2$.

5. الأخطاء من وجهة النظر العدديّة

تعتبر دراسة الخطأ موضوعاً جوهريّاً في التحليل العددي، ذلك أنّ معظم الطرائق العددية توفّر حلولاً تقريبيّة للحل المنشود، ومن المهم هنا أن نستطيع تقدير الخطأ الناشئ عن اعتماد هذه الحلول، وأن نحصره ضمن حدودٍ معيّنة. نصنّف في

هذه الفقرة الأخطاء المتعلّقة بمسألة معيّنة، كما نقدّم عرضاً موجزاً لنتائج أوليّة تتعلق بانتشار الأخطاء في أنواعٍ مختلفةٍ من الحسابات العددية.

نعرّف الخطأ في تقريبٍ لقيمة عددٍ ما بأنّه الفرق بين قيمته الحقيقيّة والقيمة التقريبيّة. نرمز بــ x_T إلى القيمة الحقيقيّة وبــ x_A إلى القيمة التقريبيّة، ونضع

$$\mathrm{Err}\,(x_A) = x_T - x_A \tag{4}$$

وفي حالة $x_T \neq 0$ نعرّف الخطأ النسبي بالعلاقة

$$\mathrm{Rel}\,(x_A) = \frac{x_T - x_A}{x_T} \tag{5}$$

مثـــال: إذا كان

$$\begin{cases} x_T = e = 2.7182818\cdots \\ x_A = \dfrac{19}{7} = 2.7142857\cdots \end{cases}$$

فإنّ

$$\begin{cases} \mathrm{Err}\,(x_A) = 0.03996 \\ \mathrm{Rel}\,(x_A) = 0.00147 \end{cases}$$

1.5 مصادر الأخطاء

تصنّف مصادر الأخطاء الأساسيّة كما يلي:

- **أولاً: أخطاء ناجمة عن النمذجة الرياضيّة للمسائل الفيزيائيّة:**

يعتبر النموذج الرياضي لظاهرةٍ فيزيائية محاولةً لإبراز علاقةٍ رياضيّة بين بعض المقادير الفيزيائيّـــة المتعلّقـــة بهذه الظاهرة. وبسبب تعقيد الواقع الفيزيائي، نعمد إلى استخدام فرضيّات تبسيطيّة، بغية الحصول على نموذج رياضي بسيط يحكم الظاهرة الفيزيائيّة. وبحكم التبسيط فإنّ النموذج الرياضيّ الموضوع سيكون محدود الدقّة، وقد تكون هذه المحدوديّة في بعض الحالات عائقاً دون الاستفادة منه، وقد يعطي النموذج نتائج مقبولة في حالاتٍ أخرى ، وهذا يتعلّق باستخدامات النموذج. وفي الحالة التي يكون فيها النموذج غير دقيق فإنّ الحلّ العددي للنموذج لا يمكنه أن يحسّن من نقص الدقّة في النموذج.

- **ثانياً: أخطاء ناجمة عن عدم الدقّة في المعطيات الفيزيائيّة:**

تشتمل معظم المعطيات في المسائل الفيزيائيّة على أخطاء، وهذا ما يؤثّر في دقّة الحسابات التي تُجرى على هذه المعطيات، ومن ثمَّ تحدّ من دقّة النتائج التي نحصل عليها.

- **ثالثاً: أخطاء ناجمة عن تدوير الأعداد:**

إنّ تمثيل الأعداد في الأنظمة الحاسوبيّة بعدد محدود من الأرقام، يجعل عمليّة تقريب الأعداد الحقيقيّة بأعداد الآلة أمراً لا بدَّ منه، للتمكّن من تمثيل الأعداد والتعامل معها في الحسابات. إنّ دقّة التقريب في تمثيل الأعداد في النظام الحاسوبي، تتعلّق

كما رأينا بالمواصفات التي يتسم بها النظام الحاسوبي للتعامل مع الأعداد. في حالة التمثيل بالنقطة العائمة، رأينا أنّ الدقّة في التمثيل تتعلّق بعدد الأرقام p المستخدمة في التعبير عن الجزء الكسري وبطبيعة الحال بأساس نظام التمثيل β المستخدم.

2.5 انتشار الأخطاء

ندرس في هذه الفقرة، الآثار الناجمة عن العمليّات الحسابيّة على أعداد تشوبها أخطاء. لهذا نرمز بـــ ω إلى عمليّة حسابيّة أساسيّة مثل $+$ ، $-$ ، $\times$ ، $/$ ؛ ونرمز بـــ ω^* إلى العمليّة الحسابيّة المقابلة في النظام الحاسوبي، والتي غالباً ما تتضمّن عمليّة تدوير rounding نرمز إليها بالتطبيق $\mathrm{rd}:\mathbb{R}\to\mathcal{A}$ حيث تمثّل $\mathcal{A}$ مجموعة الأعداد القابلة للتمثيل في النظام الحاسوبي.

ليكن x_A و y_A العددين اللذين ستجرى عليهما العمليّة الحسابيّة ، ونفترض أنّهما مشوبان بخطأ مقداره ε و η على الترتيب، أي أنّ القيم الصحيحة هي:

$$x_T = x_A + \varepsilon\ ,\ \ y_T = y_A + \eta.$$

عندها يكون $x_A\cdot\omega^*\cdot y_A$ هو العدد الناتج من إجراء العمليّة ω^*، ويكون الخطأ المرتكب في النتيجة:

$$x_T\cdot\omega\cdot y_T - x_A\cdot\omega^*\cdot y_A = \left(x_T\cdot\omega\cdot y_T - x_A\cdot\omega\cdot y_A\right) + \left(x_A\cdot\omega\cdot y_A - x_A\cdot\omega^*\cdot y_A\right). \tag{6}$$

نسمّي المقدار

$$\left(x_T\cdot\omega\cdot y_T - x_A\cdot\omega\cdot y_A\right)$$

خطأ الانتشار ، وبالمثـــل نسمّي المقدار

$$\left(x_A\cdot\omega\cdot y_A - x_A\cdot\omega^*\cdot y_A\right)$$

خطأ التدوير. وفيما يخصّ خطأ التدوير فإنّه غالباً مايكون:

$$x_A\cdot\omega^*\cdot y_A = \mathrm{rd}\left(x_A\cdot\omega\cdot y_A\right). \tag{7}$$

وهذا يعني أنّ المقدار $x_A\cdot\omega\cdot y_A$ **يحسب تماماً** ومن ثَمَّ يجري تدوير الناتج. وينتج من ذلك أنّ خطأ التدوير يحقق المتراجحة:

$$\left|x_A\cdot\omega\cdot y_A - x_A\cdot\omega^*\cdot y_A\right| \le \frac{\beta}{2}\cdot\left|x_A\cdot\omega\cdot y_A\right|\cdot\beta^{-p} \tag{8}$$

وفيما يخصّ انتشار الأخطاء، نتفحّص الحالات الخاصّة التالية:

☒ **حالة عمليّة الضرب**

$$\begin{aligned} x_T\cdot y_T - x_A\cdot y_A &= x_T\cdot y_T - (x_T-\varepsilon)\cdot(y_T-\eta)\\ &= x_T\cdot\eta + y_T\cdot\varepsilon - \varepsilon\cdot\eta \end{aligned}$$

ويكون الخطأ النسبي:

$$\mathrm{Rel}\,(x_A \cdot y_A) = \frac{x_T y_T - x_A y_A}{x_T y_T} = \frac{\eta}{y_T} + \frac{\varepsilon}{x_T} - \frac{\varepsilon}{x_T} \cdot \frac{\eta}{y_T}$$

ومنه

$$\mathrm{Rel}\,(x_A \cdot y_A) = \mathrm{Rel}\,(x_A) + \mathrm{Rel}\,(y_A) - \mathrm{Rel}\,(x_A) \cdot \mathrm{Rel}\,(y_A) \tag{9}$$

وفي حالة كون $\left|\mathrm{Rel}\,(x_A)\right| \ll 1$ و $\left|\mathrm{Rel}\,(y_A)\right| \ll 1$ يكون

$$\mathrm{Rel}\,(x_A \cdot y_A) \approx \mathrm{Rel}\,(x_A) + \mathrm{Rel}\,(y_A) \tag{10}$$

الرمز $\ll$ يعني "أصغر بكثير من".

☒ حالة عمليّة القسمة

بطريقةٍ مشابهةٍ للحالة السابقة نحصل على:

$$\mathrm{Rel}\,(x_A \,/\, y_A) = \frac{\mathrm{Rel}\,(x_A) - \mathrm{Rel}\,(y_A)}{1 - \mathrm{Rel}\,(y_A)} \tag{11}$$

عندما يكون $\left|\mathrm{Rel}\,(y_A)\right| \ll 1$ فإنّ

$$\mathrm{Rel}\,(x_A \,/\, y_A) \approx \mathrm{Rel}\,(x_A) - \mathrm{Rel}\,(y_A) \tag{12}$$

نلاحظ في عمليّتي الضرب والقسمة، أنّ الأخطاء النسبيّة لا تنتشر بسرعة.

☒ حالة عمليّتي الجمع والطرح

لدينا هنا

$$(x_T \pm y_T) - (x_A \pm y_A) = (x_T - x_A) \pm (y_T - y_A) = \varepsilon \pm \eta$$

ومنه

$$\mathrm{Err}\,(x_A \pm y_A) = \mathrm{Err}\,(x_A) \pm \mathrm{Err}\,(y_A) \tag{13}$$

☒ حالة حساب القيمة العدديّة لتابع

لتكن $f(x_A)$ القيمة التقريبيّة لقيمة التابع $f(x_T)$ عند النقطة x_T. باستخدام مبرهنة القيمة الوسطيّة نكتب

$$f(x_T) - f(x_A) \approx f'(x_T) \cdot (x_T - x_A) \tag{14}$$

وذلك بفَرْض أنّ x_A و x_T عددان قريبان نسبياً وأنّ $f'(x)$ لا يتغيَّر كثيراً بين العددين x_A و x_T. على سبيل المثال:

$$\begin{aligned}\sin(\pi/5)-\sin(0.628) &\approx \cos(\pi/5)\cdot(\pi/5-0.628)\\ &\approx 0.00026\end{aligned}$$

وهذا تقريبٌ ممتاز للخطأ.

☒ حالة حساب مجموع

لندرس حساب مجموع من النمط

$$S=\sum_{i=1}^{m} x_i \tag{15}$$

باستخدام نظامٍ حاسوبي تمثّل فيه الأعداد $x_1,\ldots,x_m$ مجموعة من أعداد الآلة، أي إنّ $x_i=\mathrm{rd}(x_i)$ أيّا كان $i\in\{1,\ldots,m\}$. نعرِّف

$$S_2=\mathrm{rd}(x_1+x_2)=(x_1+x_2)\cdot(1+\varepsilon_2)$$

حيث

$$|\varepsilon_2|\le\frac{\beta}{2}\beta^{-p}$$

نعرِّف بالتدريج، المجموع

$$S_{r+1}=\mathrm{rd}(S_r+x_{r+1})\ ,\qquad 1\le r\le m-1.$$

عندها يكون

$$S_{r+1}=\ (S_r+x_{r+1})\cdot(1+\varepsilon_{r+1})\ ,\ |\varepsilon_{r+1}|\le\frac{\beta}{2}\beta^{-p}.$$

ونحصل أخيراً على

$$\begin{aligned}S_2-(x_1+x_2)&=(x_1+x_2)\cdot\varepsilon_2\\ S_3-(x_1+x_2+x_3)&=(x_1+x_2)\cdot\varepsilon_2+(x_1+x_2)\cdot(1+\varepsilon_2)\cdot\varepsilon_3+x_3\cdot\varepsilon_3\\ &\approx(x_1+x_2)\cdot\varepsilon_2+(x_1+x_2+x_3)\cdot\varepsilon_3\\ S_4-(x_1+x_2+x_3+x_4)&\approx(x_1+x_2)\cdot\varepsilon_2+(x_1+x_2+x_3)\cdot\varepsilon_3+\\ &\quad(x_1+x_2+x_3+x_4)\cdot\varepsilon_4\end{aligned}$$

ومن ثَمَّ، فإنّ

$$\begin{aligned}S_m-\sum_{i=1}^{m}x_i&\approx(x_1+x_2)\cdot\varepsilon_2+\cdots+(x_1+\cdots+x_m)\cdot\varepsilon_m\\ &=x_1\cdot(\varepsilon_2+\cdots+\varepsilon_m)+x_2\cdot(\varepsilon_2+\cdots+\varepsilon_m)+\\ &\quad x_3\cdot(\varepsilon_3+\cdots+\varepsilon_m)+\cdots+x_m\cdot\varepsilon_m\end{aligned}\tag{16}$$

من هذه العلاقة نستنتج أنّ أفضل استراتيجيّة لجمع الأعداد هي أن يجري الجمع **من الأصغر إلى الأكبر**.

☒ تأثير المعطيات غير الدقيقة

إنّ المعطيات ذات المصدر التجريبي (ناجمة عن قياسات فيزيائيّة) غالباً ماتكون ذات **دقّة محدودة**، لنقل إنّها مقاديرٌ مقيسةٌ بدقّة r رقماً. إنّ الأخطاء المرتكبة في القياس (نتيجة نقص الدقّة في أجهزة القياس) سوف تنتشر بالحسابات التي تجرى على هذه المعطيات، ومن ثَمَّ تتسبب في الحصول على نتائج حسابيّة لها عددٌ من أرقام الدقَّة يَقِلُّ عن r رقماً. فمثلاً، لننظر في مسألة حساب المجموع

$$\sum_{i=1}^{m} x_i \cdot y_i$$

الأعداد x_i و y_i تقع في المجال $[0.1,1]$ وتتمتع بدقّة r رقماً. باستخدام الحسابات العشريّة $(\beta = 10)$، نحسب المجموع بطريقتين ونحلل الخطأ. في الحالتين تمثّل X_i و Y_i القيم الصحيحة:

$$\left.\begin{aligned} X_i &= x_i + \varepsilon_i \\ Y_i &= y_i + \eta_i \end{aligned}\right\} , \ |\varepsilon_i| \quad , \quad |\eta_i| \le 5 \times 10^{-(r+1)}.$$

حالة أولى: نكوّن الجداء $x_i \cdot y_i$ وندوّر إلى r رقماً ثُمَّ نجمع. إنّ الخطأ المرتكب في المجموع هو

$$\begin{aligned} E_1 &= \sum_{i=1}^{m} X_i Y_i - \text{Computed value of} \left(\sum_{i=1}^{m} x_i y_i \right) \\ &= \sum_{i=1}^{m} X_i Y_i - \sum_{i=1}^{m} \left[(X_i - \varepsilon_i) \cdot (Y_i - \eta_i) + \gamma_i \right] \end{aligned}$$

حيث $|\gamma_i| \le 5 \times 10^{-(r+1)}$ ، γ_i هو خطأ التدوير في الجداء $x_i \cdot y_i$. لدينا في هذه الحالة

$$E_1 \approx \sum_{i=1}^{m} \left[\varepsilon_i \cdot Y_i + \eta_i \cdot X_i - \gamma_i \right]$$

ومنه نجد

$$|E_1| \le 3m \left(5 \times 10^{-(r+1)} \right) \tag{17}$$

حالة ثانية: نكوّن الجداءات $(i = 1, \dots, m)$. $x_i \cdot y_i$ ، بدقّة $2r$ ونجري عمليّة جمعها محتفظين بكلّ أرقام الدقّة. ثُمَّ نقوم بتدوير ناتج الجمع الأخير إلى r رقماً. عندها يعطى الخطأ المرتكب في هذه الحالة

$$\begin{aligned} E_2 &= \sum_{i=1}^{m} X_i Y_i - \text{Computed value of} \left(\sum_{i=1}^{m} x_i y_i \right) \\ &= \sum_{i=1}^{m} X_i Y_i - \left\{ \sum_{i=1}^{m} \left[(X_i - \varepsilon_i) \cdot (Y_i - \eta_i) \right] + \gamma \right\} \\ &\approx \left(\sum_{i=1}^{m} \left[\varepsilon_i \cdot X_i + \eta_i \cdot Y_i \right] \right) - \gamma \end{aligned}$$

حيث $|\gamma| \le 5 \times 10^{-(r+1)}$. ومنه نستنتج

$$|E_2| \le (2m + 1) \left(5 \times 10^{-(r+1)} \right) \tag{18}$$

من الواضح أنّ تحديد الخطأ في المتراجحة (15) أفضل مما هو عليه في المتراجحة (14). إنّ الخطأ E_2 ناتجٌ من انتشار أخطاء المعطيات الأصليّة ؛ في حين يشمل E_1، إضافةً إلى ذلك أخطاء تدوير النتائج المتوسطة إلى r رقماً. ومن ثَمَّ تَتَبيّن أهميّة إجراء الحسابات بدقّة تتجاوز الـــ r رقماً لتجنّب حدوث انتشار أوسع للأخطاء.

6. مفهوم الاستقراريّة في التحليل العددي

تتأثر حلول العديد من المسائل الرياضيّة بدرجةٍ ملحوظة ببعض الأخطاء الحسابيّة مثل أخطاء التدوير. ولكي نستطيع دراسة وتحليل مثل هذه الظواهر، نشرح مفهوم "الاستقراريّة" Stability ومفهوم "الحساسيّة" Sensitivity. إنّ حساسيّة مسألة ما مرتبطةٌ إلى حدٍّ بعيد بالدقّة العظمى التي يمكننا الوصول إليها في حلِّ هذه المسألة، عندما نستخدم نظام تمثيلٍ منتهٍ للأعداد ونجري العمليّات الحسابيّة فيه لإيجاد الحل. نوسّع في مرحلةٍ لاحقة هذه المفاهيم لتشمل الطرائق العددية المستخدمة في حلّ المسائل الرياضيّة على أنواعها. وبوجهٍ عام، نعتمد الطرائق العدديّة التي لا تتمتّع بحساسيّة عالية للأخطاء الصغيرة التي قد تشوب معطيات المسألة، أو قد تنجم عن أخطاء تمثيل الأعداد وأخطاء التدوير عند إجراء الحسابات.

لتبسيط العرض، نقتصر في نقاشنا على دراسة المسائل التي لها شكل المعادلة:

$$F(x,y)=0 \tag{19}$$

حيث يمثّل المتحوّل x مجهول المسألة الذي نودُّ تعيينه بحلّ هذه المعادلة، ويمثّل المتحوّل y المعطيات التي تتعلّق بها نتيجة الحل. إنّ شكل المعادلة (19) يمكن أن يمثّل العديد من المسائل الرياضيّة. فعلى سبيل المثال:

- تابعٌ حقيقيّ بمتحوّل حقيقي x ومتحوّل شعاعي y، تساهم مركّباته في تعريف العلاقة التابعيّة F ؛
- المعادلة (19) يمكن أن تكون معادلة تكامليّة أو معادلة تفاضليّة يمثّل فيها x التابع المجهول ويمثّل y تابعاً معطى أو قيماً محيطيّة معطاة.

نقول عن المسألة (19) إنّها مستقرّة Stable إذا كان الحلّ x يتعلّق باستمرار بالمتحوّل y. بمعنى أنّه إذا كانت $\{y_n\}$ متتالية من القيم التي تقرِّب قيمة y بمعنًى من المعاني، فإنّ متتالية الحلول الموافقة $\{x_n\}$ يجب أن تقرِّب الحلّ x على نحوٍ موافق. نسمّي أيضاً المسائل المستقرّة، مسائلَ جيِّدة الطّرح well-posed problems كما نسمّي المسائل غير المستقرّة Unstable، مسائلَ سيِّئة الطرح ill-posed problems.

أمثـــلة:

❶ لننظر في مسألة حل المعادلة الجبريّة

$$a\,x^2+b\,x+c=0\ ,\ \ a\neq 0.$$

المعطيات هنا هي الثلاثيّة $y=(a,b,c)$. نعلم أنّه في حالة تحقق الشرط $b^2-4ac\geq 0$ تقبل هذه المعادلة حلولاً حقيقيّة تعطى بالقانون:

$$x = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$$

ومن الواضح أنّ هذين الحليّن مستمرّان كتوابع بالمتحولات a، b و c.

❷ لتكن المعادلة التكامليّة التالية:

$$\int_0^1 \frac{0.75\, x(t)\, dt}{1.25 - \cos\left[2\pi\,(s+t)\right]} = y(s)\ , \qquad 0 \le s \le 1. \tag{20}$$

إنّ هذه المسألة غير مستقرّة. لأنّه توجد متتالية من التشويشات $\{\delta_n\}_{n\in\mathbb{N}}$:

$$n \in \mathbb{N},\ \ \delta_n(s) = y_n(s) - y(s)$$

تحقق

$$\lim_{n\to\infty}\ \max_{0\le s\le 1}\ \left|\delta_n(s)\right| = 0. \tag{21}$$

وتحقق متتالية الحلول الموافقة $\{x_n\}_{n\in\mathbb{N}}$ الشرط:

$$\forall n \in \mathbb{N},\ \ \max_{0\le s\le 1}\ \left|x_n(s) - x(s)\right| = 1. \tag{22}$$

فعلى سبيل المثال، نعرّف $y_n(s) = y(s) + \delta_n(s)$ حيث

$$\forall n \in \mathbb{N},\ \ \delta_n(s) = \frac{1}{2^n}\cos(2n\pi s)\ , \qquad 0 \le s \le 1.$$

ويكون في هذه الحالة

$$x_n(s) - x\,(s) = \cos(2n\pi\, s).$$

وهذا يحقق الشرط (19).

إذا كانت مسألة من النمط (19) غير مستقرّة، فستواجهنا صعوبات جديّة إذا ما حاولنا حلَّها. ومن غير الممكن عادةً أن نحلَّ مثل هذه المسائل بدون محاولتنا فهم جوانبَ عديدة تتعلّق بخصائص الحل. وهذا الجانب من المعالجة، للمسائل غير المستقرّة هو مجال بحثٍ حثيث في الرياضيّات التطبيقيّة عموماً، وفي مسائل التحليل العددي بوجهٍ خاص.

في الواقع العملي، هناك العديد من المسائل المستقرّة بالمعنى الذي أوردناه آنفاً، ولكن تبقى هناك صعوبات وعوائق تنجم عن الحسابات العددية المتعلّقة بطرائق حلّ هذه المسائل. ولكي نستطيع التعامل مع هذه الصعوبات، نستخدم مفهوماً لقياس استقراريّة المسألة نسمّيه "عدد الحساسيّة" ويقابل هذه التسمية المصطلح اللاتيني Condition number.

نحاول بواسطة عدد الحساسيّة أن نقيس أسوأ آثار المسألة (19) في الحل x، عندما يتعرّض المتحول y لتشويش بمقدارٍ بسيط. ليكن δy التشويش الذي تعرَّض له المتحوّل y وليكن $x + \delta x$ هو حلّ المعادلة المشوّشة:

$$F(x+\delta x, y+\delta y)=0 \tag{23}$$

نعرِّف

$$\kappa(x)=\sup_{\delta y}\frac{\|\delta x\|/\|x\|}{\|\delta y\|/\|y\|} \tag{24}$$

نستخدم هنا الرمز $\|\cdot\|$ لندلَّ على قياسٍ للمقدار الموضوع داخله. يحسب الحـــدّ الأعلى sup في العبارة (24) على مجموعة التشويشات الصغيرة δy التي تكون المسألة (23) عندها ذات معنىً، أي قابلةً للحل. وبطبيعة الحال فإنّ المسائل غير المستقرّة تقودنا إلى $\kappa(x)=\infty$.

إنّ عدد الحساسيّة $\kappa(x)$ للمسألة (19) يقيس حساسيّة الحل x للتغيّرات الصغيرة في المعطيات y. فإذا كان $\kappa(x)$ كبيراً نسبيّاً، فإنّ هذا يعني وجود تغيير صغير نسبيّاً δy للمعطيات y يحدث تغييراً كبيراً نسبيّاً δx في الحل x. ولكن إذا كـــان $\kappa(x)$ صغيراً، ولنقل $\kappa(x)\le 10$، عندها لا يكون للتغييرات الصغيرة في المعطيات δy الأثر الكبير في تغيّر الحل δx.

فالتغيرات الصغيرة نسبيّاً في y تقودنا إلى تغيّـــرات صغيرة نسبيّاً في الحل x. ولمّا كانت الحسابات العدديّة يشوبها عادةً أخطاءٌ صغيرةٌ ناجمة عن التدوير وتقريب الأعداد، فإننا لا نرغب أن يكون عدد الحساسيّة لمسألة نودّ حلّها بالطرائق العدديّة كبيراً. إنّ مسألة كهذه نسميّها سيّئة الحساسيّة ill-conditioned (ذات حساسيّة عالية)، وتكون غالباً صعبة الحل بدقّة.

أمثـــلة:

❶ لنطرح مسألة إيجاد حل للمعادلة

$$x-a^{y}=0\ ,\ \ a>0 \tag{25}$$

نشوش y بـــ δy فنحصل على

$$\frac{\delta x}{x}=\frac{a^{y+\delta y}-a^{y}}{a^{y}}=a^{\delta y}-1$$

ومنه نجد عبارة عدد الحساسيّة

$$\kappa(x)=\sup_{\delta y}\frac{|\delta x/x|}{|\delta y/y|}=\sup_{\delta y}\left|y\left(\frac{a^{\delta y}-1}{\delta y}\right)\right|$$

فإذا كان المقدار δy صغيراً نحصل على

$$\kappa(x)\approx|y\ln a| \tag{26}$$

وبصرف النظر عن طريقـــة حساب x في المعادلة (25)، إذا كان $\kappa(x)$ كبيراً فإنّ تغيّرات صغيرة في y تقود إلى تغيّرات أكبر بكثير نسبيّاً في الحل x. فإذا كان $\mathcal{K}(x)=10^{4}$ وكان الخطأ النسبي في قيمة y المستخدمة هو 10^{-7} وهو

خطأ ناجم عن التمثيل بدقّة منتهية في النظام الحاسوبي، عندها نتوقع أن يكون الخطأ النسبي في قيمة x الناتجة قُرابة 10^{-3}. ويمثّل هذا هبوطاً كبيراً في الدقّة.

❷ لتكن المعادلة

$$\begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix} \cdot \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}$$

التي تكتب بالشكل

$$A \cdot x = b \tag{27}$$

حيث

$$A = \begin{pmatrix} 0.780 & 0.563 \\ 0.913 & 0.659 \end{pmatrix}, \ b = \begin{pmatrix} 0.217 \\ 0.254 \end{pmatrix}, \ x = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}.$$

وهذه مسألة من النمط (19) معطياتها y والمصفوفة A والشعاع b. نبحث عن الحل x الذي يحقق الشرط(27).
تتمتع هذه المسألة بحساسية عالية، فأيّ تغيير بسيط في المعطيات، أي عناصر المصفوفة A وعناصر الشعاع b، ينتج عنه تغيير كبيرٌ في الحل. تقبل المعادلة (27) حلاً وحيداً هو

$$x = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

فإذا أجرينا تغييراً بسيطاً في عناصر الشعاع b وليكن $\Delta b = \begin{pmatrix} 10^{-3} \\ -10^{-3} \end{pmatrix}$ حصلنا على المعادلة المشوشة

$$A \cdot \tilde{x} = b + \Delta b \tag{28}$$

تقبل هذه المعادلة حلاً هو

$$\tilde{x} = A^{-1} \cdot \tilde{b} = \begin{pmatrix} 659000 & -563000 \\ -913000 & 780000 \end{pmatrix} \cdot \begin{pmatrix} 0.218 \\ 0.253 \end{pmatrix} = \begin{pmatrix} 1223 \\ -1694 \end{pmatrix}.$$

نلاحظ هنا أنّ

$$\begin{aligned} \|\delta x\|_\infty &= 1693 \ , \\ \|x\|_\infty &= 1, \\ \|\delta b\|_\infty &= 10^{-3} \ , \\ \|b\|_\infty &= 0.254. \end{aligned}$$

وهذا ما يبرر **السلوك الحسّاس** جداً لهذه الجملة الخطيّة.

7. استقراريّة الطرائق العددية

نقول عن طريقة عددية لحل مسألةٍ رياضيّة إنّها مستقرّة إذا كانت حساسيّة ناتجها العددي للمعطيات لا تزيد عن حساسيّة المسألة الرياضيّة الأصليّة. يعالج المثال التالي مسألة حساب تكامل محدود باتباع طرائق عددية، وبه نوضّح مفهوم الاستقراريّة هذا.

ليكن

$$E_n = \int_0^1 x^n e^{x-1} dx, \quad n \in \mathbb{N}. \tag{29}$$

إذا كان $n \geq 1$ فإنّ

$$\left(x^n e^{x-1}\right)' = n \cdot x^{n-1} e^{x-1} + x^n e^{x-1}.$$

بمكاملة الطرفين على المجال $[0,1]$ نجد

$$\int_0^1 x^n e^{x-1} dx = \left[x^n e^{x-1}\right]_0^1 - n \int_0^1 x^{n-1} e^{x-1} dx.$$

ومنه نحصل على علاقة التدريج

$$\begin{cases} E_n = 1 - n \cdot E_{n-1} \ , \ n \geq 1 \\ E_0 = \int_0^1 e^{x-1} dx = 1 - e^{-1} \end{cases} \tag{30}$$

باستخدام نظام تمثيلٍ حاسوبي فيه $n = 6$ و $\beta = 10$، وبتطبيق علاقة التدريج (27) يمكننا حساب قيمٍ تقريبيّة لـ E_n عندما $n = 1, 2, \ldots, 9$ فنجد:

$E_1 \cong 0.367879$	$E_6 \cong 0.127120$
$E_2 \cong 0.264242$	$E_7 \cong 0.110160$
$E_3 \cong 0.207274$	$E_8 \cong 0.118720$
$E_4 \cong 0.170904$	$E_9 \cong -0.068480 < 0$
$E_5 \cong 0.145480$	

نلاحظ هنا أنّ القيمة العدديّة التي نحصل عليها بهذه الطريقة كقيمةٍ تقريبيّة للتكامل E_9 هي قيمة سالبة، على حين يجب أن يكون تكامل التابع $x^9 e^{x-1}$ على المجال $[0,1]$ موجباً، فما هو سبب هذه المشكلة؟

في الواقع، إذا أردنا تحليل ما حدث ومعرفة مصدر الخطأ الذي تسبب في هذه النتيجة، نحلل الأخطاء الحاصلة في النواتج العدديّة للتكاملات E_1 , E_2 , ..., E_9 فنجد:

في حساب $E_1 = 1 - E_0 = e^{-1}$ نرتكب خطأ تقريب من مرتبة دقّة الآلة مقداره 4.412×10^{-7} . هذا الخطأ ينتشر في حساب E_2 ليصبح $(-2) \times 4.412 \times 10^{-7}$ ، وهكذا يكون خطأ التقريب الناتج في حساب E_9 من مرتبة

$$(-2)(-3)\cdots(-9) \times 4.412 \times 10^{-7} = 9! \times 4.412 \times 10^{-7} \approx 0.1601.$$

إنّ القيمة الدقيقة لـ E_9 في هذه الحالة (بدقّة ثلاثة أرقام) هي:

$$-0.06848 + 0.1601 = 0.0916.$$

وهنا نطرح التساؤل عن وجود طريقة أخرى تحلّ مشكلة عدم الاستقرار هذا ؟ في الواقع، إذا أعدنا كتابة علاقة التدريج (30) بالشكل:

$$E_{n-1} = \frac{1 - E_n}{n} \;,\; n \geq 1 \tag{31}$$

وإذا كنّا نعرف قيمةً تقريبيّة لـ E_n فإن استخدام العلاقة (31) لحساب E_{n-1} يسمح لنا بالحصول على قيمةٍ تقريبيّة يتناقص فيها خطأ التقريب n مرّة عمّا هو عليه في E_n.

لتكن $n \gg 1$، لدينا

$$E_n = \int_0^1 x^n e^{x-1} dx \leq \int_0^1 x^n dx = \left. \frac{x^{n+1}}{n+1} \right|_0^1 = \frac{1}{n+1} \cdot$$

نستنتج من ذلك أنّ $\lim_{n \to \infty} E_n = 0$.

فعلى سبيل المثال، يمكننا اعتبار القيمة 0 تقريباً لـ E_{20} ونكون في هذه الحالة قد ارتكبنا خطأً لا يتجاوز الـ $\frac{1}{21}$.

وباستخدام العلاقة (28) يمكننا حساب قيمة تقريبيّة لـ E_{19} يكون الخطأ فيها محدوداً بـ

$$\frac{1}{20} \times \frac{1}{21} \cong 0.0024.$$

تسمح علاقة التدريج (31) تدريجيّاً، بإيجاد قيمةٍ تقريبيّة لـ E_{15} بخطأٍ لا يتجاوز 4×10^{-8} وهذا الخطأ أقلّ من دقّة الآلة التي نجري بها الحسابات (هنا $eps = 5 \times 10^{-6}$). باستخدام هذه الطريقة نحصل على القيم التقريبيّة التالية:

$E_{20} \cong 0.0$	$E_{14} \cong 0.062732\underline{2}$
$E_{19} \cong 0.0500000$	$E_{13} \cong 0.066947\underline{7}$
$E_{18} \cong 0.0500000$	$E_{12} \cong 0.071773\underline{3}$
$E_{17} \cong 0.0527778$	$E_{11} \cong 0.077352\underline{3}$
$E_{16} \cong 0.0557190$	$E_{10} \cong 0.083877\underline{1}$
$E_{15} \cong 0.059017\underline{6}$	$E_9 \cong 0.091612\underline{3}$

نلاحظ هنا أنّ الخطأ الابتدائي المرتكب في تقريب قيمة التكامل E_{20} قد اختفى نهائيّاً اعتباراً من الحد E_{15} ؛ وهذا يعود إلى استقراريّة هذه الطريقة، ومن ثَمَّ نكون قد حصلنا على قيمٍ تقريبيّة للتكاملات $E_{15}, \ldots, E_9$ بدقّةٍ تفوق دقّة الآلة التي نجري بها الحسابات، ومن هذا نستنتج أن جميع أرقام الجزء الكسري في تمثيل القيم $E_{15}, \ldots, E_9$ هي أرقامٌ **معنويّة** significant digit باستثناء الرقم الأخير الذي يمكن أن يكون مدوَّراً.

8 مسائل وتمارين

1.8. يعتمد تمثيل النقطة العائمة القياسيّة في نظامٍ حاسوبي المواصفات

$$e_{\max} = 9 \text{ ، } e_{\min} = -9 \text{ ، } p = 3 \text{ ، } \beta = 10$$

كم يبلغ عدد الأعداد القياسيّة Normalized number في هذه النظام؟

2.8. ليكن نظام التمثيل بالنقطة العائمة القياسيّة المحدّد بالوسطاء $\beta = 10$ ، $p = 3$. ما هو الخطأ النسبي المرتكب في حساب ناتج العمليات الحسابيّة التالية:

$$(3.28\ 10^{-2}) \times^* (6.98\ 10^3)$$
$$[(3.28\ 10^{-2}) \times^* (6.98\ 10^3)] /^* (4.82\ 10^{-8})$$

إذا كان هذا النظام يعتمد التدوير إلى الأقرب في تمثيل النواتج.

3.8. بيّن أنّه إذا كان x عدداً حقيقياً صغيراً بالقدر الكافي فإنّ حساب القيمة العددية للمقدار

$$1 - \cos x$$

يكون مصحوباً بخطأٍ كبيرٍ ناجمٍ عن عمليّة الطرح. اقترحْ صيغة مكافئة لحساب هذه القيمة بحيث نتجنّب هذا الخطأ.

4.8. نعلم أنّ أحد جذور المعادلة $a \cdot x^2 + b \cdot x + c = 0$ يعطى بالعلاقة

$$r_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a}$$

(*a*) عيّن قيمة r_1 بأكبر دقّة ممكنة وذلك في حالة الأمثال $a = 1$ ، $b = 111.11$ و $c = 1.2121$.

(*b*) احسب قيمة هذا الجذر مستخدماً نظام تمثيل بخمسة أرقام $p = 5$. ما هو تعليقك على النتيجة؟

(*c*) اقترح صيغة مكافئة أكثر ملائمةً للحساب في نظام تمثيلٍ بخمسة أرقام.

5.8. نجد في الكتب التي تعرض جداول تكاملية علاقة التدريج التالية

$$\int \frac{\mathrm{d}\,x}{x(a+x^2)^{m+1}} = \frac{1}{2am(a+x^2)^m} + \frac{1}{a}\int \frac{\mathrm{d}\,x}{x(a+x^2)^m} \qquad (m \neq 0)$$

(a) نعرّف

$$f(m) = \int_1^2 \frac{\mathrm{d}\,x}{x(a+x^2)^m}$$

استخدم علاقة التدريج السابقة في تعيين المقدار c_m الذي يحقق العلاقة

$$(M) \qquad f(m+1) = c_m + \frac{1}{a} f(m)$$

(b) ادرس استقراريّة الطريقة (M).

6.8. لتكن المتتالية

$$a_0 = 11/2,$$
$$a_1 = 61/11,$$
$$a_{n+1} = 111 - \frac{1130}{a_n} + \frac{3000}{a_n a_{n-1}}, \qquad n \geq 1.$$

(a) ادرس تجريبياً تقارب هذه المتتالية مستخدماً نظام تمثيل الأعداد في برنامج مثل Mathmatica أو Maple وذلك بإجراء الحسابات بالدقّة البسيط (ثمانية أرقام عشرية) وبالدقّة المضاعفة (ست عشرة رقماً عشرياً). بما في ذلك القيم الابتدائيّة $\hat{a}_0 = \mathrm{rd}(a_0)$ و $\hat{a}_1 = \mathrm{rd}(a_1)$.

(b) ادرس التقارب النظري لهذه المتتالية بمساعدة أحد البرنامجين Mathmatica أو Maple. ماذا تلاحظ؟

7.8. ليكن $x_A = 0.937$ قيمة تقريبيّة بثلاثة أرقام معنويّة للقيمة الحقيقيّة x_T.

أولاً: عيّن تحديداً من الأعلى للخطأ النسبي في القيمة التقريبيّة x_A.

ثانياً: ليكن التابع $x \mapsto f(x) = \sqrt{1-x}$.

عيّن تحديداً من الأعلى للخطأ النسبي في حساب القيمة التقريبيّة $f(x_A)$ بالمقارنة مع القيمة الافتراضيّة $f(x_T)$.

8.8. لتكن المعادلة

$$x^2 - 40 \quad x + 1 = 0$$

أوجد جذور هذه المعادلة بدقّة خمسة أرقام، علماً أنّ $\sqrt{399} \approx 19.975$ هو تقريبٌ ناجمٌ عن تدوير القيمة الصحيحة للجذر التربيعي $\sqrt{399}$ إلى خمسة أرقام.

9.8. ليكن التابع f المعرّف بالعلاقة

$$x \mapsto f(x) = \frac{\ln(1-x) + x \cdot e^{x/2}}{x^3}$$

1. احسب قيم التابع $f(x)$ عند النقاط $\left\{ x = 10^{-m} : m = 1, \cdots, 7 \right\}$.
2. ماهي القيمة المتوقّعة (بمقتضى الدراسة النظريّة) للنهاية $\lim_{x\to 0} f(x)$ ؟
3. عندما تكون x قريبة من الصفر، ماهي أفضل طريقة لحساب $f(x)$ ؟
   (توجيـــه: استخدم منشور تايلور بجوار الصفر).

10.8. نريد أن نستخدم المتسلسلة الصحيحة $e^x = \sum_{n=0}^{\infty} x^n / n!$ لحساب قيمة تقريبيّة للعدد e^{-5}. ما هو عدد حدود التقريب $\sum_{n=0}^{m} x^n / n!$ لكي نحصل على قيمة تقريبيّـــة بخطأٍ نسبي لا يتجاوز 10^{-3} ؟