ГЛАВА I

ВВЕДЕНИЕ

§1. Математическое моделирование. Численные методы и использование ЭВМ в решении прикладных задач

Рассматривая *математический анализ явления* как своего рода *теоретический эксперимент*, из общих и достаточно естественных соображений процесс *математического моделирования* разбивается на несколько этапов:

- **Формулировка математической модели явления.** Математическая модель любого изучаемого явления, по причине его чрезвычайной сложности, должна охватывать важнейшие для рассматриваемой задачи стороны процесса, его существенные характеристики и формализованные связи, подлежащие учёту.

  Как правило, *математическая модель* изучаемого физического явления формулируется в виде *уравнений математической физики.* На этой стадии анализа это существенно нелинейные, многомерные системы уравнений, содержащие большое число неизвестных и параметров.

  Если *математическая модель* выбрана недостаточно тщательно, то какие бы мы не применяли методы для дальнейших расчётов, полученные результаты будут *ненадёжны*, а в отдельных случаях и совершенно *неверны.*

- **Проведение математического исследования** полученной модели и получение соответствующего *решения.*

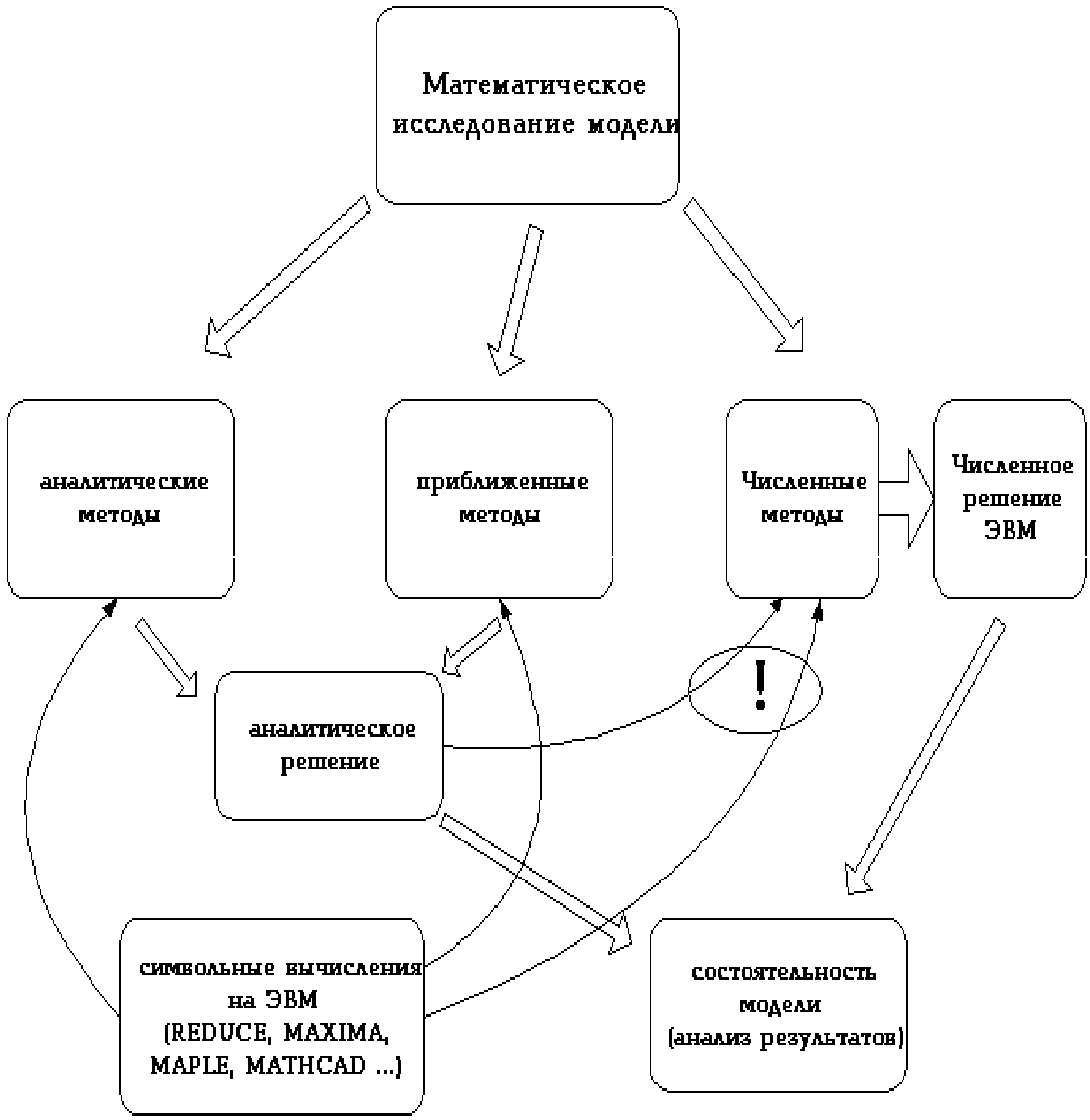
  На этом этапе моделирования, в зависимости от сложности рассматриваемой модели, применяют различные подходы к её исследованию и различный смысл вкладывается в понятие *решения* задачи. Скажем, доказательство теорем *существования и единственности* в определённом смысле *решает* задачу, однако, являясь зачастую неконструктивным, оно не позволяет нам решить проблему изучения качественного поведения решения и оценки его количественных характеристик.

  Для наиболее *грубых* и несложных (в некотором смысле) моделей удаётся получить их *аналитическое решение.* Следует оговориться — использование средств *символьных вычислений* на ЭВМ таких как REDUCE, MAXYMA, MAPLE, "интеллектуальных калькуляторов" MATHEMATICA, MathCAD, MatLab и пр. существенно революционизировало это, традиционное для "бумаги и карандаша", поле деятельности.

  Для более точных и сложных моделей *аналитическое решение* удаётся получить сравнительно редко. При теоретическом анализе задачи в такой ситуации пользуются обычно *приближенными* математическими методами, например разложением по малому параметру, осреднением, изучение различных асимптотик и другими. Эти приёмы позволяют опять-таки представить приближенное решение в аналитической форме и с его помощью получить удовлетворительные численные результаты.

Наконец для наиболее точных и сложных моделей основными методами решения являются *численные* методы решения с необходимостью требующие проведения большого объёма вычислений на ЭВМ. Эти методы позволяют добиться хорошего *количественного* и даже *качественного* результата в описании модели. Но, правда, у них есть и принципиальные недостатки — как правило, речь идёт о рассмотрении некоторого *частного* решения.

Приведённая схема частично отражает обсуждаемые взаимосвязи этапов математического моделирования.



Как мы видим, каждый из этапов математического исследования модели связан с использованием *численных методов* и получением *численного решения* задачи.

- **Анализ состоятельности предложенной модели,** т. е. осмысление результатов решения, сопоставление полученного решения с имеющимися данными физического эксперимента. На этом этапе решается вопрос о состоятельности математической модели и проведённого исследования. "Хорошее" согласование с "экспериментом" обычно свидетельствует о правильности выбора модели. В противном случае необходимы дополнительные уточнения, изменения и т. п., повторение предыдущих этапов исследования.

Обсуждая предмет лекционного курса, мы акцентировали наше внимание на двух сторонах предмета "Численные методы": *этапе в математическом моделировании* и на *необходимом моменте в процессе исследования, сопряженном с использованием ЭВМ.*

Использование ЭВМ в процессе математического исследования модели требует специфических, численных методов, т.е. такой "интерпретации" математической модели, которая может быть реализована на ЭВМ – назовём её *дискретной* (или *вычислительной*) моделью. Поскольку ЭВМ выполняет только арифметические и логические операции, то для реализации *вычислительной модели* требуется разработка соответствующего *вычислительного алгоритма.* Дальнейшая последовательность действий — это программирование. расчет на ЭВМ, обработка результатов расчета.

В рамках нашего лекционного курса мы остановимся на отдельных проблемах численных методов при анализе сравнительно простых и ставших классическими математических моделей.

Теперь посмотрим на проблему "численных методов" несколько подругому.

§2. Задача "вычисления"

2.1 Задача "вычисления". Анализ постановки

Обычно задачу вычисления величины y по известной величине x записывают, с учётом интересующих нас причинно-следственных связей, в виде

$$y = \mathcal{A}(x), \tag{1}$$

где $y \in \mathcal{Y}, \quad x \in \mathcal{X}$ – элементы соответствующих функциональных пространств [*1)]; $\mathcal{A}$ – оператор (правило), реализующий вычисления.

В первую очередь нас будут интересовать корректно поставлененые задачи вычисления.

Задача вычисления $y = \mathcal{A}(x)$ называется корректно поставленной, если для любых входных данных из некоторого класса решение задачи существует, единственно и устойчиво по входным данным (т.е. непрерывно зависит от входных данных задачи).

[*1)] Если не оговорено особо, то $\mathcal{Y}, \mathcal{X}$ – как правило *линейные, нормированные, полные,* т.е. *банаховы* пространства.

В сформулированное понятии *корректности* поставленной задачи (по Адамару) учтены достаточно естественные требования, действительно: чтобы численно решать задачу нужно быть уверенным, что её решение *существует*. Столь же разумны для конкретных условий и требования *единственности* решения, и, поскольку наши действия носят принципиально приближенный характер, то необходимо требование *устойчивости* решения.

Сделаем несколько замечаний об *устойчивости*. Нас интересует решение y задачи (1) соответствующее входным данным x. Реально мы имеем возмущенные входные данные с погрешностью δx, т.е. $x + \delta x$ и находим возмущенное решение

$$y + \delta y = \mathcal{A}(x + \delta x).$$

Эта погрешность входных данных порождает *неустранимую* погрешность решения

$$\delta y = \mathcal{A}(x + \delta x) - \mathcal{A}(x).$$

Если решение непрерывно зависит от входных данных, то

$$\|\delta y\| \to 0 \quad \text{всегда при} \quad \|\delta x\| \to 0$$

и задача (1) устойчива по входным данным.
Отсутствие устойчивости означает, что даже "небольшим" погрешностям δx могут соответствовать "большие" погрешности δy, т.е. построенное при расчёте решение будет сильно отличаться от истинного.

Применять непосредственно к такой неустойчивой задаче численные методы бессмысленно. Однако и не всякую формально устойчивую задачу удобно решать практически. Пусть имеет место оценка

$$\|\delta y\| \leq C \cdot \|\delta x\|, \quad \text{но} \quad C\text{– велико.}$$

Задача формально устойчива, но *неустранимая ошибка* решения может быть большой. Это случай *плохой обусловленности* или *слабой устойчивости* задачи вычисления.

Приведем несколько примеров постановки *задачи вычисления* (1).

2.2 Примеры постановки задачи вычисления

1°.Задача нахождения корней полинома. Рассмотрим некоторый полином степени n в приведенном виде (старший коэффициент равен единице):

$$p_n(x) \equiv x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1} x + a_n$$

вообще говоря с комплексными коэффициентами $(a_k, x \in C)$.
Требуется определить его корни. Пусть E^n — n–мерное комплексное евклидово пространство. Положим, что компоненты некоторого вектора $\vec{z} = \{z_1, z_2, \ldots, z_n\}$ этого пространства являются корнями полинома $p_n(x)$, т.е.

$$p_n(z_i) = 0, \quad i = \overline{1, n}.$$

Тогда, в силу теоремы Безу, мы можем $p_n(x)$ записать в виде:

$$p_n(x) = x^n + a_1 x^{n-1} + a_2 x^{n-2} + \cdots + a_{n-1}x + a_n = (x - z_1)(x - z_2)\cdots(x - z_n) = \prod_{i=1}^{n}(x - z_i).$$

Отсюда мы получаем известные формулы Виетта:

$$a_k = (-1)^k \sigma_k, \quad k = \overline{1,n}. \tag{$*$}$$

Здесь σ_k — элементарные, симметричные относительно $z_1, z_2, \ldots, z_n$ однородные функции k–го порядка

$$\begin{cases} \sigma_1 & = & z_1 + z_2 + \cdots + z_n \\ \sigma_2 & = & z_1z_2 + z_1z_3 + \cdots + z_1z_n + z_2z_3 + \cdots + z_{n-1}z_n \\ & \cdots & \\ \sigma_n & = & z_1z_2\cdots z_n. \end{cases}$$

(каждое σ_k содержит C_n^k слагаемых).

Таким образом формулы Виетта ($*$) сопоставляют каждому вектору $z \in E^n$ вектор $\vec{a} = \{a_1, a_2, \ldots, a_n\} \in E^n$ того же пространства, т.е. определяют отображение $\mathcal{V} : E^n \Rightarrow E^n$ пространства E^n *на себя.* С помощью этого отображения $\mathcal{V}$ задача определения корней полинома $p_n(x)$ формулируется следующим образом:
Для заданного вектора $\vec{a}$ найти вектор $\vec{z} \in E^n$ такой, что

$$\mathcal{V}(\vec{z}) = \vec{a}. \tag{2}$$

В курсе высшей алгебры показано, что отображение $\mathcal{V}$ *взаимнооднозначное* и *взаимнонепрерывное*, т.е. задача (2) *корректна.*

2°.Основная задача линейной алгебры. Пусть дана матрица $A_{(p\times q)} = \|a^i_j\|^p_q$ и два евклидовых пространства E^p и E^q. Тогда определено отображение

$$A: \quad E^q \Rightarrow E^p; \quad \vec{y} = A\vec{x}, \quad \vec{y} \in E^p, \vec{x} \in E^q,$$

($\vec{x}, \vec{y}$ — столбцы соответствующих размерностей).
Основная задача линейной алгебры состоит в том, чтобы по заданному вектору $\vec{f} \in E^p$ найти вектор $\vec{x} \in E^q$ такой, что

$$A\vec{x} = \vec{f}. \tag{3}$$

Задача (3) представляет собой задачу решения *системы линейных алгебраических уравнений* — СЛАУ. Связанная с решением СЛАУ ситуация нами подробно изучена в курсе *линейной алгебры*:

1) если $p = q$ и $det\,A \neq 0$, то задача (3) поставлена *корректно* (её решение дается формулами Крамера);

2) в остальных случаях, если система (3) совместна ($rangA = rangA'$), то решение неединственно. В противном случае решение вовсе отсутствует, т.е. задача (3) в этих случаях *некорректно поставлена.*

3°.Задача Коши для обыкновенного дифференциального уравнения.
Пусть требуется найти решение обыкновенного дифференциального уравнения (ОДУ), отвечающее начальному условию $y(a) = b$

$$\begin{cases} \frac{dy}{dx} = f(x, y), \quad a < x \leq c \\ y(a) = b. \end{cases} \tag{$*$}$$

Здесь a, b — заданные числа; $f(x, y)$ — определена в полосе $\Pi = \{(x, y); \quad a \leq x \leq c; \quad y \in (-\infty; \infty)\}$ и удовлетворяет в Π условиям теоремы о продолжимости решения (*) на отрезок $[a; c]$.

Обозначим через $\mathcal{R}_0$ множество всевозможных решений задачи Коши (*), отвечающих различным значениям начального условия b. Определим отображение $\mathcal{K}: \quad \mathcal{R}_0 \Rightarrow R^1$, полагая

$$\mathcal{K}(y(x)) = y(a), \quad \forall y \in \mathcal{R}_0.$$

Тогда решение задачи Коши для ОДУ (*) можно сформулировать так:
по заданному числу b найти функцию $y(x)$ такую, что

$$\mathcal{K}(y(x)) = b. \tag{4}$$

В курсе *дифференциальных уравнений* доказана *корректность* задачи (4).

Число рассмотренных примеров *задачи вычисления*можно было бы множить, но мы ограничимся рассмотренными примерами постановки *задачи вычисления*.

§3. Численное решение корректных задач Структура погрешности решения

3.1 Задача "вычисления". Погрешности

Обратимся снова к *задаче вычисления* (1)

$$y = \mathcal{A}(x).$$

В рассмотренных примерах (2)–(4) соответствующее *правило* $\mathcal{A}$ реализующее "вычисление" задано явно неконструктивно. Речь идёт по сути об обращении операторов $\mathcal{V}^{-1}, A^{-1}, \mathcal{K}^{-1}$, точнее о численной реализации обратного отображения для (2)–(4). Такая ситуация типична и лишний раз показывает, что, как правило, *вычисление* $\mathcal{A}$ не может быть "просто " реализовано. Чтобы преодолеть эти сложности задачу (1) заменяют другой, "близкой" к ней задачей, но уже которая "легко" решается численно. При этом в первую очередь анализируют вопрос о вносимых в решение погрешностях.

Есть четыре основных источника погрешности результата вычислений: *математическая модель*; *исходные данные* задачи; *приближнный метод* и *погрешность при реализации вычислений* (в частности *погрешность округления*):

$\delta_1 y$ –*погрешность математической модели*, связана с физическими допущениями при выборе математической модели и на анализе этой погрешности мы останавливаться не будем;

$\delta_2 y$ –*погрешность исходных данных*, порождает *неустранимую погрешность решения*

$$\delta_2 y = \mathcal{A}(x+\delta x) - \mathcal{A}(x);$$

$\delta_3 y$ –*погрешность метода.* Выражение $\mathcal{A}(x)$, вообще говоря, не может быть "просто" численно реализовано. Задачу $y = \mathcal{A}(x)$ заменяют "близкой" задачей

$$\bar{y} = \bar{\mathcal{A}}(\bar{x}), \qquad (1')$$

Мы переходим к другим функциональным пространствам $\mathcal{X}, \mathcal{Y} \Rightarrow \bar{\mathcal{X}}, \bar{\mathcal{Y}}$ элементы которых допускают сравнительно "простую" численную реализацию. Соответствующим образом меняется и отображение $\mathcal{A} \Rightarrow \bar{\mathcal{A}}$.

При этом естественно требовать, чтобы задача $(1')$ была *корректна* и чтобы решение $\bar{y}$ было близко к решению y. Величина

$$\delta_3 y = y - \bar{y} = \mathcal{A}(x) - \bar{\mathcal{A}}(\bar{x})$$

и представляет собой *погрешность метода.*

$\delta_4 y$ –*вычислительная погрешность.* При численной реализации $\bar{y}$, которая уже, по предположению, возможна получают элемент $\tilde{y}$, поскольку промежуточные результаты округлялись и т.п. Таким образом *вычислительная погрешность метода* может быть записана в виде

$$\delta_4 y = \bar{y} - \tilde{y} = \bar{\mathcal{A}}(\bar{x}) - \tilde{y}.$$

Полезно сразу же сформулировать некоторые эмпирические правила, которых придерживаются при реализации задачи вычисления:

$$\|\delta_2 y\| \sim (2 \div 5)\|\delta_3 y\| \gg \|\delta_4 y\|.$$

1) При проведении вычислений нужно стремиться, чтобы погрешность метода $\delta_3 y$ была бы в несколько раз меньше *неустранимой погрешности* решения $\delta_2 y$:

2) *Вычислительная погрешность* $\delta_4 y$ должна быть существенно меньше всех остальных погрешностей решения, т.е. расчёт нужно вести с таким количеством значащих цифр, чтобы погрешность округления была существенно меньше всех остальных погрешностей.

Теперь мы можем ещё раз очертить круг вопросов, рассматриваемых в рамках нашего лекционного курса "Численных методов" — это [*1)]

1) конструирование *дискретной* (или *вычислительной*) модели $\{\bar{\mathcal{X}}, \bar{x}, \bar{\mathcal{A}}\}$:

2) разработка на её основе соответствующих алгоритмов решения редуцированной задачи вычисления

$$\bar{y} = \bar{\mathcal{A}}(\bar{x});$$

3) анализ погрешности метода $\delta_3 y$ и частично вычислительной погрешности $\delta_4 y$ алгоритма, реализующего вычисления $\bar{\mathcal{A}}$.

[*1)] Предмет лекционного курса мог бы быть и более содержательным и обширным, но, как всегда, здесь есть свои, не зависящие от нашего желания, ограничения, определяемые спецификой учебного плана факультета.

3.2 Погрешность округления на t–разрядной ЭВМ

Остановимся несколько подробнее в рамках этого параграфа, но кратко, на анализе *вычислительной погрешности* $\delta_4 y$, обязанной погрешностям округления при реализации численного алгоритма.

1°.Погрешность единичного округления. В современных ЭВМ действительные числа представляются в т.н. форме с *плавающей запятой*, т.е. если само число a в позиционной системе счисления с основанием r записано в виде r-ичной дроби

$$a = \operatorname{sign} a(a_n a_{n-1} \dots a_1 a_0{,}a_{-1}a_{-2}\dots)_r = \operatorname{sign} a(a_n r^n + \dots + a_1 r + a_0 + \frac{a_{-1}}{r} + \frac{a_{-2}}{r^2} + \dots),$$

то такую форму записи числа a называют *представлением с фиксированной запятой.* Здесь $a_k \in \{0; 1; \dots; "(r-1)"\}$ — r–ичные цифры.
Представление числа a в форме с *плавающей запятой* или *нормализованное представление* означает его запись в виде

$$a = \operatorname{sign} a M r^p = \operatorname{sign} a \cdot r^p \cdot (\frac{b_1}{r} + \frac{b_2}{r^2} + \dots),$$

где p — порядок числа (целое); M — мантисса числа a, причем $1/r \le M < 1$, т.е. превая r–ичная цифра в записи мантиссы b_1 неравна нулю.

В современных ЭВМ в качестве основания системы счисления r выбирается двойка — $r = 2$. Тогда, если для записи мантиссы отводится только t двоичных разрядов, то это позволяет из диапазона $[M_0; M_\infty = M_0{}^{-1}]$ (для положительных чисел) записать лишь конечное число рациональных чисел, а все остальные вещественные числа подвергаются округлению при их представлении в ЭВМ.
Точность представления числа a с помощью округлённого числа $\tilde{a}$ характеризуется относительной погрешностью округления

$$\delta_a = \frac{|a - \tilde{a}|}{|a|}.$$

При простейшем способе округления *усечением*, когда все лишние разряды мантиссы просто отбрасываются, можно легко получить оценку величины относительной погрешности δ_a единичного округления. Действительно [*1)]

$$|a - \tilde{a}| = 2^p \left| \frac{b_{t+1}}{2^{t+1}} + \cdots \right| \le 2^p \cdot \frac{1}{2^{t+1}} \left(1 + \frac{1}{2} + \cdots \right) = 2^{p-t}.$$

С другой стороны[*2)] $|a| \ge 2^p \cdot (1/2)$. Таким образом для погрешности единичного округления получаем

$$\delta_a = \frac{|a - \tilde{a}|}{|a|} \le \frac{2^{p-t}}{2^{p-1}} = 2^{-(t-1)}.$$

Более точный способ округления дает для погрешности единичного округления вдвое меньшую оценку через *машинное эпсилон*

$$\delta_a = 2^{-t} \le \varepsilon_M. \tag{5}$$

[*1)] Здесь при оценке все двоичные цифры в остатке заменены единицей $b_k \le 1, k \ge t+1$.
[*2)] Мы полагаем $a_i = 0,$ при $i \ge 2$; $\quad a_1 = 1$ всегда.

Относительная погрешность представления числа с плавающей запятой в ЭВМ определяется числом разрядов мантиссы и не превышает машинного эпсилон $\varepsilon_M = 2^{-t}(\sim 10^{-12})$.
Опираясь на оценку (5) мы можем считать, что само число a и его округлённое значение $\tilde{a}$ связаны соотношением

$$\tilde{a} = \mathrm{fl}\,(a) = a(1+\varepsilon_a),$$

где $|\varepsilon_a| \le \varepsilon_M = 2^{-t}$. Однако отметим, что для чисел $|a| < M_0$ в результате округления получим $\tilde{a} = 0$ и тем самым для этих чисел $\varepsilon_a = -1$ (!).

Арифметическое Устройство (АУ) современных ЭВМ сконструировано таким образом, что любая арифметическая операция при последующем округлении даёт относительную ошибку не более ε_M.
Для оценки влияния погрешности округлений на результат того или иного вычислительного алгоритма пользуются предположением о том, что *результат вычислений, искажнный погрешностью округления совпадает с результатом точного вычисления по тому же алгоритму, но с иными —* $\tilde{x}$*, входными данными.*
Таким образом

$$\tilde{y} = \bar{\mathcal{A}}(\tilde{x}) \quad \text{и} \quad \delta_4 y = \bar{y} - \tilde{y} = \bar{\mathcal{A}}(\bar{x}) - \bar{\mathcal{A}}(\tilde{x}).$$

Это допущение позволяет связать анализ *вычислительной погрешности* $\delta_4 y$ с анализом *устойчивости* алгоритма $\bar{\mathcal{A}}$ по *входным данным.* Ограничимся рассмотрением
Пример. Рассмотрим задачу о нахождении произведения n сомножителей

$$z_n = \prod_{k=1}^{n} y_k.$$

Пусть вычисления реализованы по алгоритму $\bar{\mathcal{A}}$ следующим образом:

$$\begin{cases} z_k &= y_k \cdot z_{k-1}, \quad k = 1, 2, \ldots, n \\ z_0 &= 1. \end{cases}$$

Предположим, что в результате округлений вместо точного значения z_{k-1} получено знаgение $\tilde{z}_{k-1}$. Тогда вместо величины $y_k \tilde{z}_{k-1}$ получим величину

$$\tilde{z}_k = \mathrm{fl}\,(y_k \cdot \tilde{z}_{k-1}) = y_k \cdot \tilde{z}_{k-1}(1+\varepsilon_k); \quad |\varepsilon_k| \le \varepsilon_M.$$

Таким образом мы получили алгоритм $\tilde{\mathcal{A}}^{*1)}$

$$\begin{cases} \tilde{z}_k &= \tilde{y}_k \cdot \tilde{z}_{k-1}, \quad k = 1, 2, \ldots, n \\ \tilde{z}_0 &= 1. \end{cases}$$

Оценим результирующую относительную погрешность

$$\delta_{z_n} = \left| \frac{z_n - \tilde{z}_n}{z_n} \right| = \frac{\left| \prod\limits_{k=1}^{n} y_k - \prod\limits_{k=1}^{n} (1+\varepsilon_k) y_k \right|}{\left| \prod\limits_{k=1}^{n} y_k \right|} \le (1+\varepsilon_M)^n - 1 = n\varepsilon_M + O(\varepsilon_M^2).$$

или, пренебрегая слагаемыми второго и больших порядков по ε_M получим окончательно

$$\delta_{z_n} \le n\varepsilon_M.$$

$^{*1)}$Структура полученного алгоритма $\bar{\mathcal{A}}$ подтверждает сформулированное допущение.