**Лекція 11. Похибки обчислювальних операцій з фіксованою та плаваючою комою у комп’ютеризованих інформаційно-вимірювальних системах**

**План лекції:**

1. Обчислювальні похибки арифметичних операцій. Загальні положення

2. Похибки округлення арифметичних операцій, що виконуються в режимі з фіксованою комою

3. Похибки округлення арифметичних операцій, що виконуються в режимі з плаваючою комою

4. Обчислювальні похибки арифметичних операцій

**1. Обчислювальні похибки арифметичних операцій. Загальні положення**

*Обчислювальна похибка* — це сумарна похибка арифметичної операції, що складається з трансформованої похибки від похибок операндів, що у операції, і помилки округлення цієї операції. При аналізі обчислювальних похибок арифметичних операцій виходитимемо з того, що кожен операнд, який бере участь в операції, має або деяку похибку, яка виникла як результат округлення при поданні операнда в кінцевій розрядній сітці, або є похибкою, накопиченою в ході виконання деякого ланцюжка операцій при отриманні даного операнда.

Для аналізу обчислювальних похибок введемо такі позначення: ε1, ε2 - похибки операндів xl та х2 відповідно; ε - обчислювальна похибка арифметичної операції; ε0 - помилка округлення арифметичної операції. Аналіз формування обчислювальних похибок операцій порівняно простий і може бути заснований на застосуванні загальної теорії помилок, використовуючи при цьому найпростіший метод лінеаризації, який зручний для аналізу обчислювальних похибок окремих арифметичних операцій та простих обчислювальних процесів.

Для складних обчислювальних процесів загальна теорія точності в більшості випадків не дає рішення через громіздкість обчислень і неврахування деяких принципово важливих факторів, що впливають на поширення похибок під час обчислень.

При аналізі помилок округлення та обчислювальних похибок виходимо з того, що обчислення в ЕОМ ведуться в двійковій системі числення з розрядною сіткою для подання мантиси кожного числа в л-двійкових розрядів. При розгляді похибок приймемо, що після кожної арифметичної операції виконується округлення шляхом додавання одиниці (n+1)-й розряд. При розгляді обчислювальних похибок як дискретних випадкових величин та функцій дискретних випадкових аргументів будемо користуватися методами дослідження, розробленими для безперервних величин та функцій. Такий підхід можливий в одних випадках внаслідок малого кроку дискретності випадкових величин, в інших - внаслідок того, що сама фізична природа не виключає існування вказаних величин у безперервній формі. Нагадаємо кілька визначень.

*Помилка заокруглення.* Під помилкою округлення ε0 розуміється помилка, що виникає при виконанні арифметичної операції в обмеженій розрядній сітці. Операнди, що у операції, вважаються у своїй заданими точно.

*Обчислювальна похибка.* Під обчислювальною похибкою розуміється сумарна похибка, що складається з помилок округлення виконуваної операції і трансформованої (спадкової) похибки від похибок операндів.

*Розрядна сітка ЕОМ.* Під розрядною сіткою ЕОМ *n* розуміється число двійкових розрядів, що відводиться для представлення мантиси числа без розряду знака.

**2. Похибки округления арифметических операций з фиксированою комою**

*2.1. Додавання та віднімання.* Арифметичні операції додавання та віднімання в режимі з фіксованою точкою виконуються без заокруглення, тобто не генерують помилок. Отже, результат цих операцій немає помилки округлення.

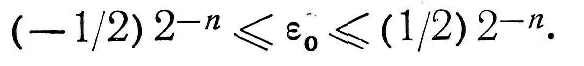
Отже, якщо величина у визначається рівнянням виду y=x1±x2, то результат обчислення матиме той самий вид y=x1±x2.

Хоча операції, що розглядаються, і не мають помилок округлення, результат буде отриманий з похибкою, що визначається похибками операндів. Операції додавання та віднімання, що виконуються в режимі з фіксованою точкою, в ході обчислювального процесу змінюють лише обчислювальну похибку за правилами алгебраїчного складання похибок операндів.

*Переповнення масш коєф зсів похибка*

*2.2. Множення*. При множенні двох n-розрядних чисел у режимі з фіксованою точкою, то добуток у загальному вигляді вимагає 2n розрядів для свого подання. Результат будь-якої операції розміщується в n-розрядах після округлення шляхом додавання одиниці в (n+1)-й розряд. Отже, якщо величина визначається рівнянням y=x1x2, то обчислювальна рівність матиме вигляд: y = x1x2 + ε0, де ε0 - помилка округлення операції множення.

За оптимального округлення маємо

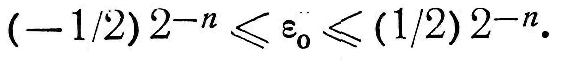


*Адапт відкидання мол розрядів. Ст розряди результату нулі. Або не адап окр*

*детерм підхід макс похибка стат підхід СКЗ похибки. Рівн розподіл у вк діап. СКЗ макс зн / 2 кв к з 3. Квадр підсум СКЗ всіх похибок з переходом у підсумку до довірчого інт заг похибки.*

*2.3. Ділення.* Передбачається, що при діленні двох чисел результат  завжди лежить у певному діапазоні, тобто виконується умова Для отримання округленого результату необхідно його обчислювати з (n+1)-м розрядом. І тут обчислювальна рівність де  — точне приватне; у - обчислене значення частки; ε0 - ошиок округлення операції ділення.

Можна також записати, що при оптимальному заокругленні



Також



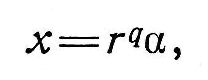
тобто різниця yx2 і х1 завжди менше ніж (1/2)⋅2-n, так як |x2|<1.

Розглядаючи помилку округлення арифметичних операцій, що виконуються в режимі з фіксованою точкою, як випадкову величину, розподілену за рівномірним законом в діапазоні від (—1/2)2-n до (1/2)2-n, можна записати, що диференціальний закон розподілу помилки округлення f(ε0)=2n, математичне очікування М[ε0] = 0, а дисперсія D[ε0] = 2-2n/12.

Таким чином, при виконанні арифметичних онера ім п в режимі з фіксованою точкою джерелами помилок округлення є тільки операції множення поділу.

**3. Похибки округлення арифметичних операцій, що виконуються в режимі з плаваючою комою**

*3.1. Додавання та віднімання.* У ЕОМ з плаваючою комою числа подаються у вигляді



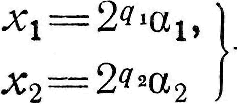
де r - основа системи числення; q - порядок числа х; α - мантіса числа х.

При аналізі похибок у режимі з плаваючою точкою приймемо, що r=2 і тобто числа нормалізовані.

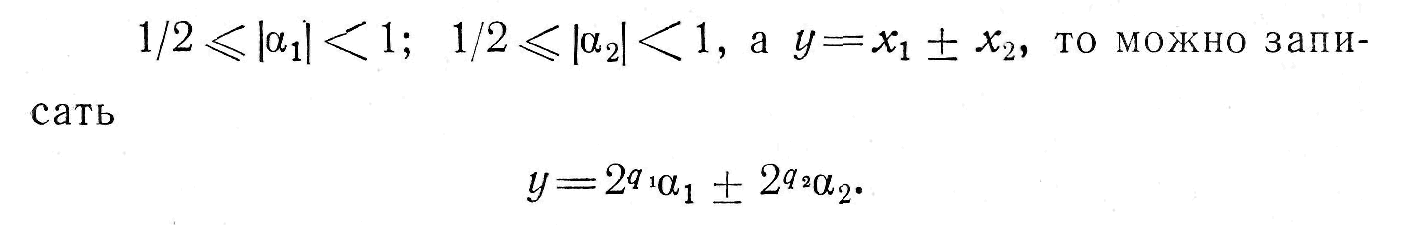
Відомо, що виконання операції алгебраїчного складу в режимі з плаваючою точкою розпадається на ряд етапів: вирівнювання порядків доданків зрушенням праворуч мантиси числа, що має менший порядок; додавання мантис; нормалізація мантиси та корекція порядку результату операції; ок-ругление результату.

Як відомо, при вирівнюванні порядків та нормалізації результатів шляхом зсуву вправо відбувається втрата молодших розрядів, що призводить до появи помилки округлення.

Якщо прийняти

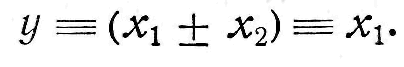


-операнди операції, де

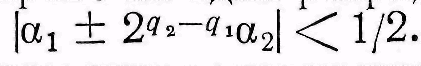


Якщо прийняти, що | x1 | > | x2|, перед додаванням обчислюється різниця (q1—q2).

Якщо різниця q1—q2>n, то х2=0 і обчислювальна рівність

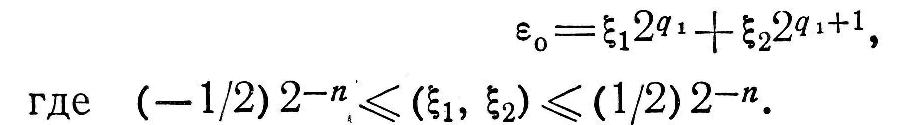


Якщо q1—q2<n, то α2 зсувається праворуч на q1—q2 розрядів. У цьому випадку відбувається втрата значущих цифр, внаслідок чого виникає помилка 

Сума α1±2q2-q1α2 вычисляется точно, далі нормали­зуєтся зсувом праворуч на 1 розряд, якщо  або ліворуч, якщо

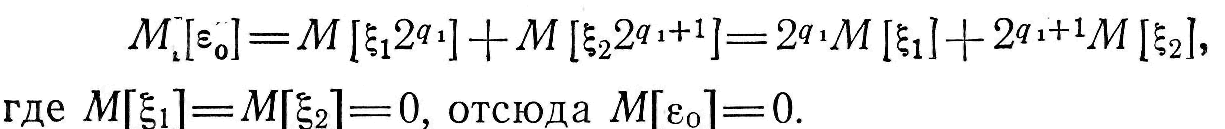
Максимальний зсув вліво за наявності n+l двійкового розряду дорівнює n розрядам. Після закінчення нормалізації проводиться округлення результату додаванням одиниці в n+1 додатковий розряд до прямого коду мантиси. При нормаліції результату зрушенням вправо відбувається також втрата значних цифр, внаслідок чого виникає помилка ξ2.

Якщо виразити обчислювальну рівність у вигляді у = х1 ± х2 + ε0, де ε0 - помилка округлення операції алгебраїчного становлення, то значення ε0 можна представити у вигляді

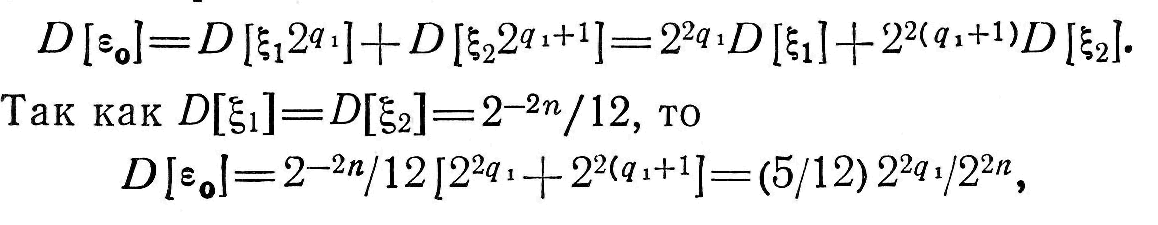


Якщо уявити помилки ξ1 і ξ2 як незалежні випадкові величини, підпорядковані закону рівномірної щільності розподілу ймовірностей, можна знайти закон розподілу випадкової величини ε0 та її моменти.

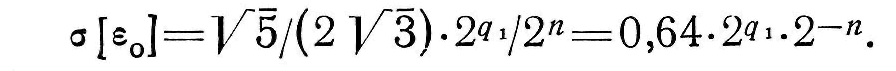
Оскільки зміни порядків (масштабних множників) відбуваються значно повільніше, ніж зміни мантис, будемо розглядати порядки як невипадкові величини. Виходячи з цього, математичне очікування





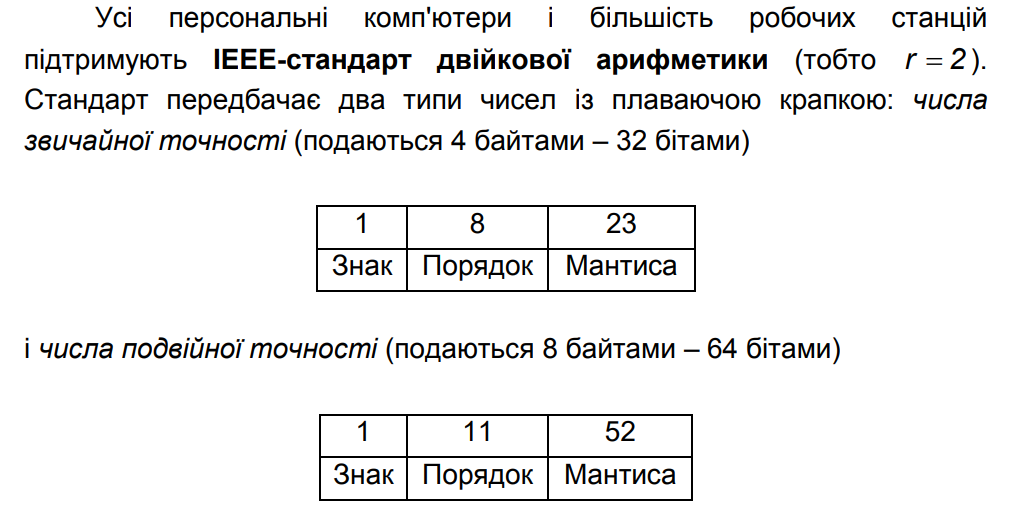


а середньоквадратичне відхилення



*Статистичний підхід*

*Оцінка СКЗ ε0 для q=8, n=32 або 64 од або подв точність дійсних чисел з пл комою. Висновок: окрім прец вим каналів ця похибка дуже мала і її можна взагалі не врахлвувати при оцінки точності результату вимірювань.*

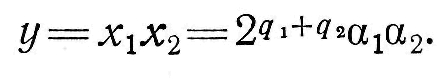


*3.2. Множення.* При множенні чисел у режимі з плаваючою точкою складається з мікрооперацій: визначення знака добуутку; складання порядків множника та множника; множення нормалізованих мантис; нормалізації та корекції порядку результату операції; округлення результату.

Для отримання точного добутку потрібно мати 2n розряди. При множенні в розрядній сітці відбувається втрата молодших розрядів добутку мантис, внаслідок чого з'являється помилка округлення.

При множенні нормалізованих мантис, що мають діапазон змін  може знадобитися нормалізація тільки на один розряд вліво, після чого коригується порядок і округляється результат.

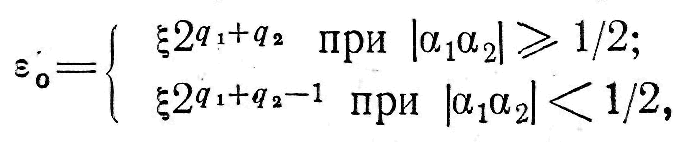
Позначимо —нормализовані мантиси, тоді



Внаслідок виконання операції множення в n-розрядній сітці відбувається втрата молодших розрядів мантиси, що викликає помилку округлення операції множення ε0, і обчислювальна рівність матиме вигляд



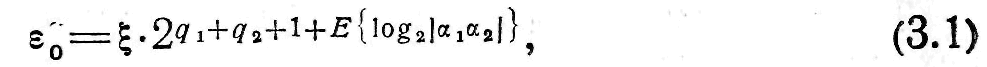
Значення ε0



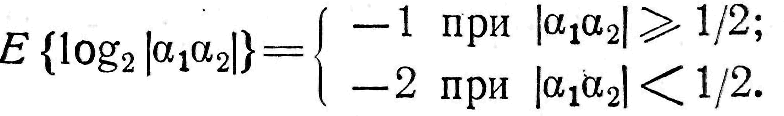
де ξ — ошибка округлення мантиси результату:



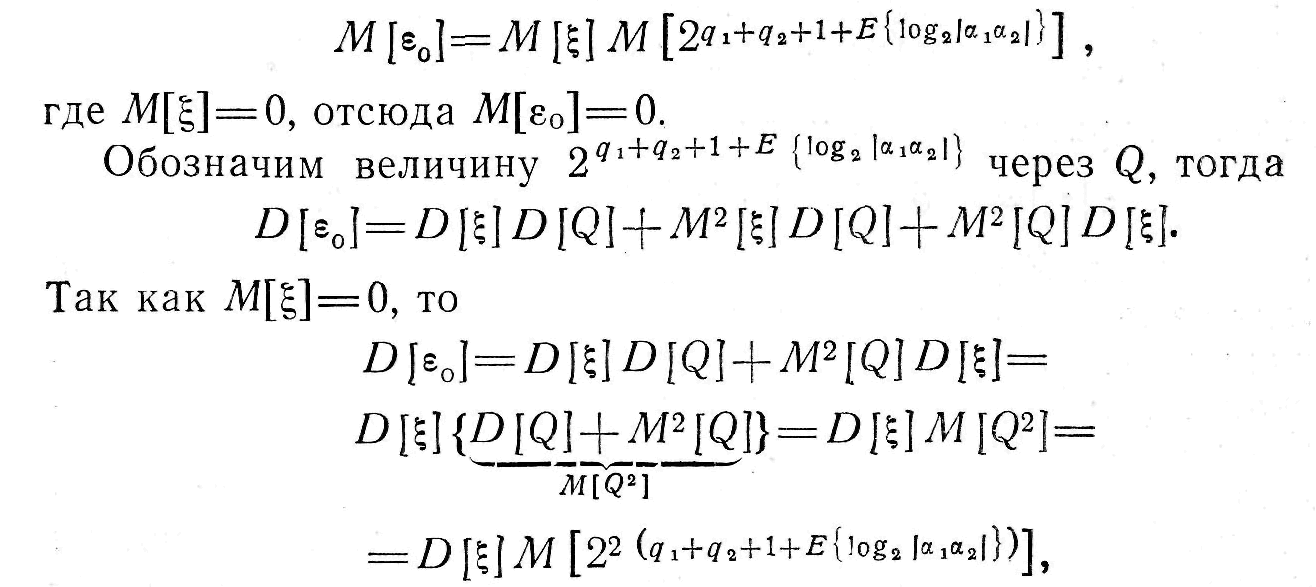
Помилка округлення операції множення, що виконується в режимі з точкою, можна представити у вигляді



де Е{х} — ціла частина числа х, що принимає значенния:



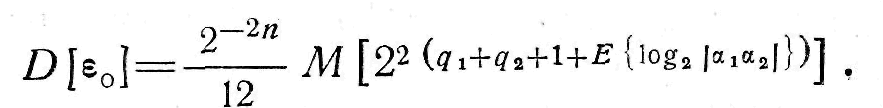
З виразу для ε0 можна визначити моменти закону розподілу помилки округлення операції множення, що виконується в режимі з плаваючою точкою:

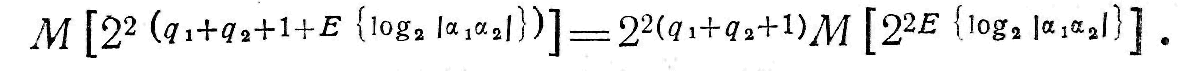


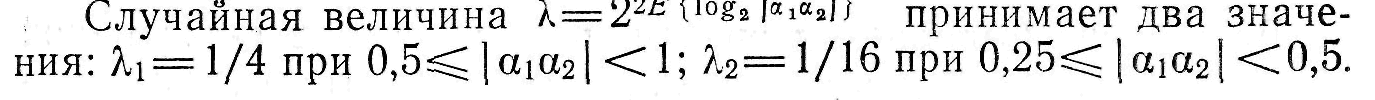
– другий початковий момент;

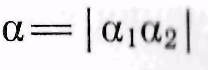


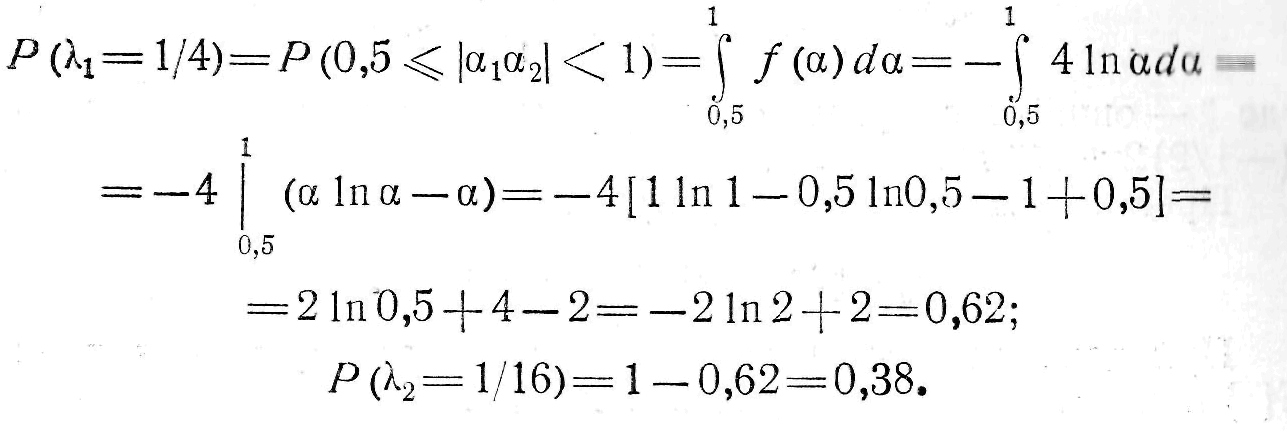
Тоді



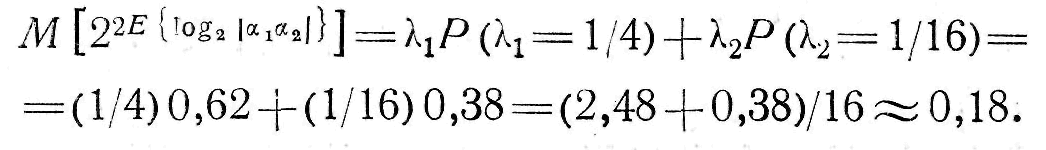




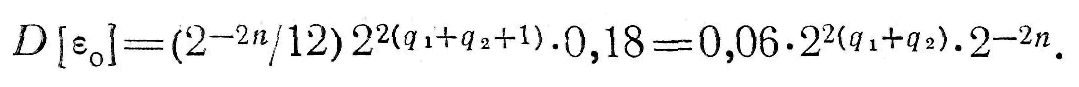
Знайдемо їх ймовірності, враховуючи, що для  закон розподілу 



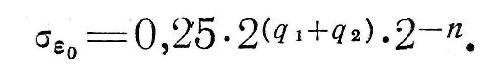
Тоді



Звідси випливає, що дисперсія помилки округлення операції множення двох нормалізованих мантис, представлених з плаваючою точкою, дорівнює



Звідсі



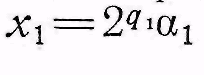
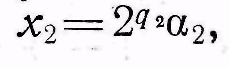
*Статистичний підхід*

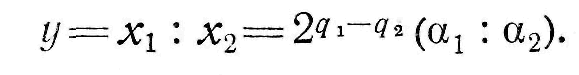
*Оцінка СКЗ ε0 для q=8, n=32 або 64 од або подв точність дійсних чисел з пл комою. Висновок: окрім прец вим каналів ця похибка дуже мала і її можна взагалі не врахлвувати при оцінки точності результату вимірювань.*

*3.3. Ділення*. При аналізі похибок можна уявити поділ нормалізованих чисел як з плаваючою точкою як низка мікрооперацій: визначення знака частки; віднімання з порядку ділимого порядку дільника; розподіл нормалізованих мантис; нормалізація та корекція порядку результату операції; округлення результату.

При розподілі точних чисел можна отримати будь-яке число вірних цифр. Дослідження показали, якщо є n+ 1 розряд, можна отримати п вірних значущих цифр. При розподілі нормалізованих мантис може знадобитися нормалізація мантиси приватного тільки на один розряд вправо, після чого коригується порядок і визначається результат.

Якщо є ділене і дільник, то запишемо рівність

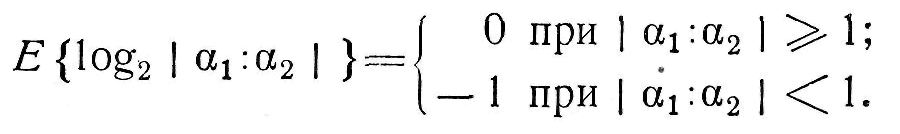


Виконання операції поділу в n-розрядній сітці викликає помилку округлення і обчислювальну рівність можна у вигляді

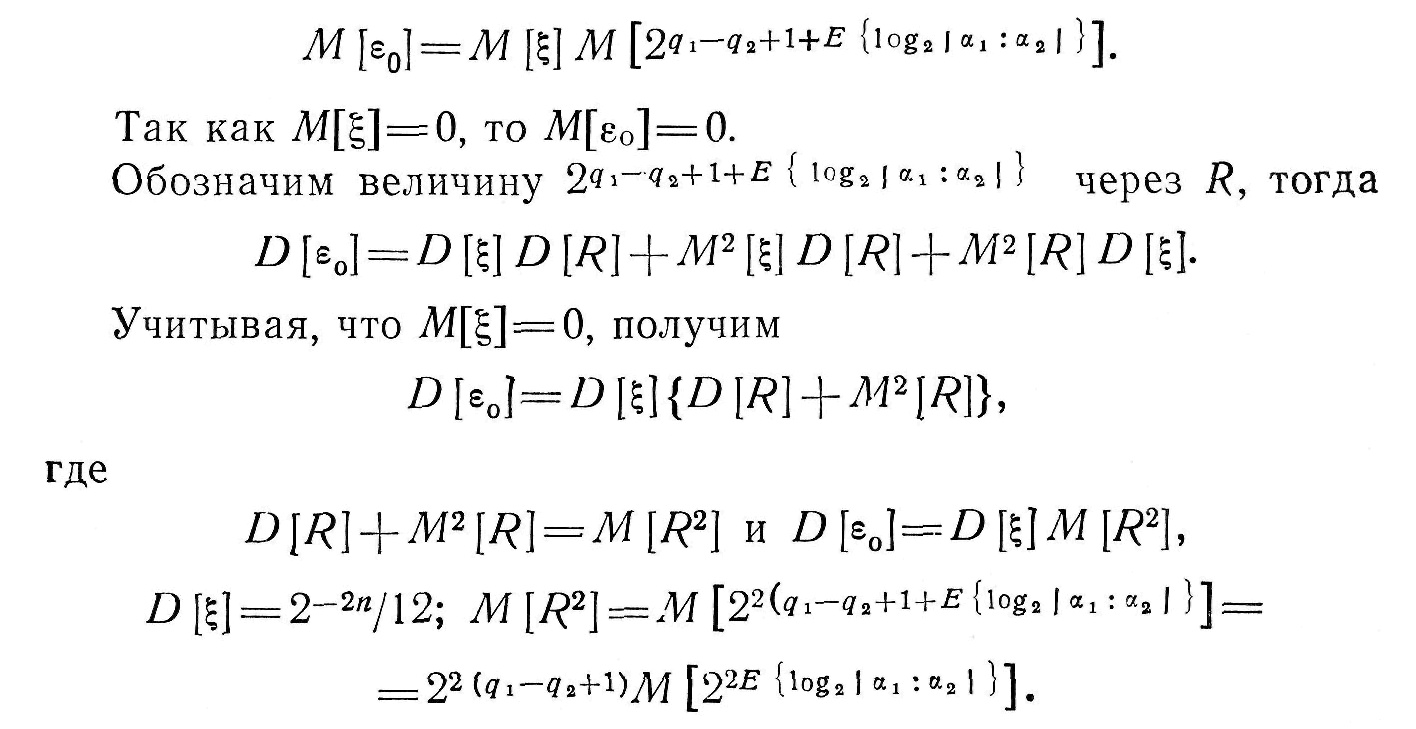


де ε0 — помилка округлення операції ділення.

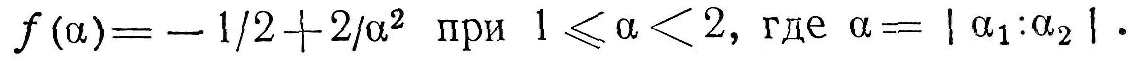
При діленні нормалізованих мантис α1 та α2 величина Е{} приймає цілочисельні значення:



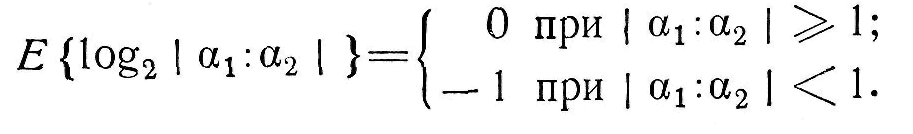
Розглядаючи помилку округлення як випадкову величину, підпорядковану деякому закону розподілу, можна визначити моменти закону розподілу помилки округлення операції розподілу. Справді,

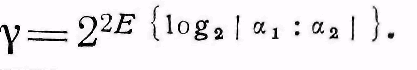


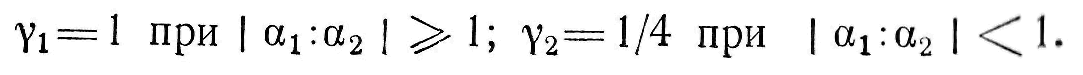
Для визначення дисперсії помилки округлення D[ε0] необхідно розкрити значення математичного очікування М [2 2Е {log2|α1^α2|}]. Для цього скористаємося законом розподілу модуля мантиси частного:



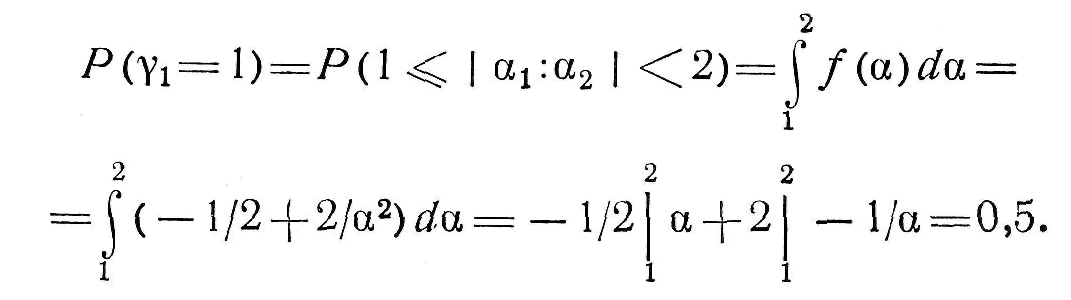
Величина



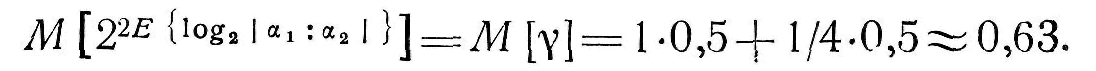
Позначимо через , що приймає два значення:



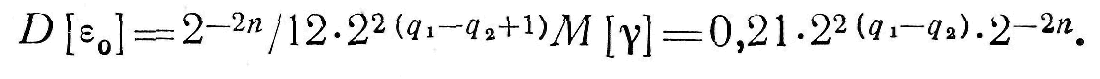
Знайдемо їх ймовірності



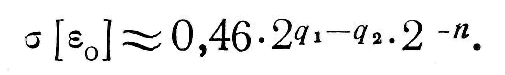
Тоді математичне очікування



Дисперсію помилки округлення операції поділу двох нормалізованих чисел, представлених з плаваючою точкою, можна визначити із співвідношення



Звідсі



*Статистичний підхід*

*Оцінка СКЗ ε0 для q=8, n=32 або 64 од або подв точність дійсних чисел з пл комою. Висновок: окрім прец вим каналів ця похибка дуже мала і її можна взагалі не врахлвувати при оцінки точності результату вимірювань.*

Виведені співвідношення показують, що єдиним способом зменшення помилок округлення арифметичних операцій при виконанні в режимах з фіксованою і плаваючою точкою є збільшення числа двійкових розрядів для подання як мантис операндів, так і мантис результатів операцій. Зазначене можна досягти або схемним способом або програмним.

*Процесор та матем співпр. Ап реаліз та прогр емул оп пл комою. Вартість та швидкодія. Директиви компілятора у програмах.*

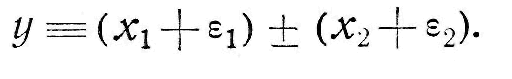
*Сучасні засоби пл кома. Блок обч з пл комою.*

*Висновок: окрім прец вим каналів похибка заокр в форматі з пл комою дуже мала і її можна взагалі не врахлвувати при оцінки точності результату вимірювань.*

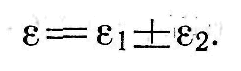
**4. Обчислювальні похибки арифметичних операцій**

Загалом обчислювальні похибки складаються з трансформованої похибки та похибки округлення результатів операцій. Так як в сучасних обчислювальних пристроях (комп’ютерах, промислових комп’ютерах, мікроконтролерах) зазвичай використовується формат чисел з плаваючою комою, то похибки округлення результатів операцій малі та їх можна не враховувати (див. оцінки п.3). Тому розглянемо трансформовані похибки як основну складову обчислювальних похибок арифметичних операцій.

*4.1. Додавання та віднімання.* При виконанні операцій додавання та віднімання двох операндів, що мають похибки, обчислювальну рівність можна записати у вигляді



Звідси випливає, що трансформовану похибку операції складання та віднімання можна визначити із співвідношення

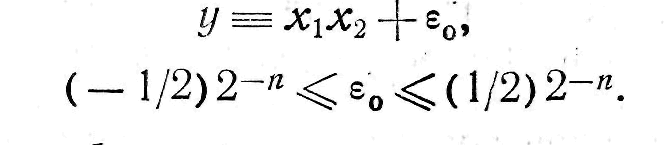


Похибки ε1 та ε2 не можна розглядати як випадкові величини, підпорядковані рівномірному закону розподілу, оскільки ці похибки у випадку виникли результаті виконання деяких послідовних операцій. Припущення про рівномірний розподіл обчислювальних похибок часто при імовірнісному аналізі призводить до підвищених результатів.

Для детермінованого підходу згідно наведеної формули арифметично підсумовуємо максимальні значення похибок. Така оцінка є завищеною.

Для ймовірнісного підходу квадратично підсумовуємо СКЗ похибок операндів. Довірчі межі трансформованої похибки визначаємо для заданої довірчої ймовірності та для нормального розподілу або розподілу Стьюдента даної похибки.

*4.2. Множення.* Обчислювальна рівність операції множення має вигляд



Вважаємо, що ε0=0.

Якщо операнди мають похибки ε1 та ε2, що виникли в ході обчислювального процесу при отриманні операндів х1 та х2, то обчислювальну похибку операції множення можна визначити зі співвідношення



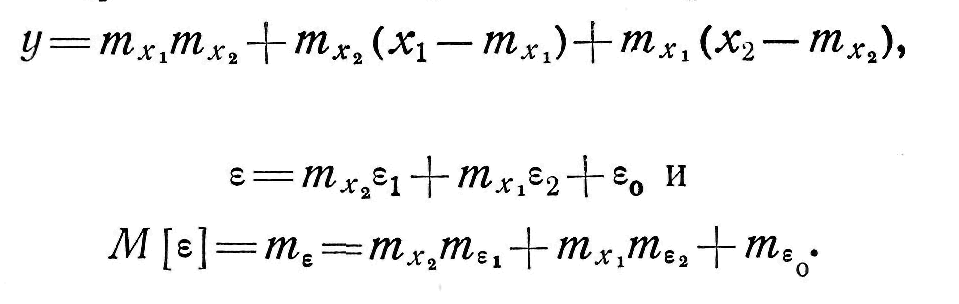
Вважаємо, що ε0=0.

Для детермінованого підходу згідно наведеної формули арифметично підсумовуємо максимальні значення похибок при максимально можливих значеннях операндів. Така оцінка є завищеною.

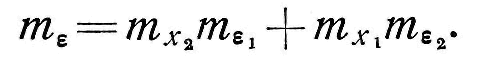
Дане співвідношення вказує на закон формування обчислювальної похибки при виконанні операції множення. Звідси випливає, що у формуванні обчислювальної похибки беруть участь як похибки операндів, а й самі операнди. Тому, для розгляду обчислювальної похибки операції множення необхідно дослідити розподіл старших значущих розрядів операндів, які, головним чином, впливають на утворення обчислювальної похибки цієї операції. Аналіз експериментальних даних, є результатами моделювання на ЕОМ, показує, що цифри старших розрядів розподілені нерівномірно; у яких переважають цифри, рівні 0. Отже, з погляду формування обчислювальної похибки і під час операції множення розподіл старших розрядів вважатимуться сприятливим.

Такий розподіл старших розрядів зміщуватиме складові обчислювальної похибки х1ε2 та х2ε1 у бік молодших розрядів значно далі, ніж за рівномірному розподілі цифр старших розрядів операндов.

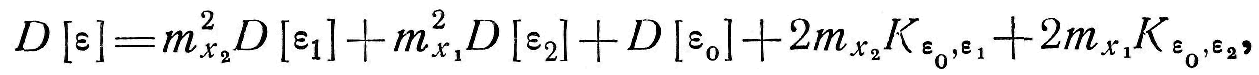
Розглядаючи обчислювальну похибку операції як випадкову величину, підпорядковану деякому закону розподілу, що залежить не тільки від законів розподілу похибок операндів, а й законів розподілу самих операндів, для визначення числових характеристик похибок можна скористатися методом лінеаризації нелінійних функцій. Справді, якщо маємо y=x1x2, де х1, х2 — значення випадкових величин з математичними очікуваннями mx1, mx2 та дисперсіями Dxl і Dx2 відповідно, то можна уявити







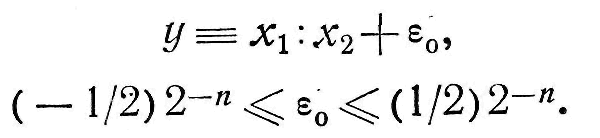
Так як при визначенні дисперсії обчислювальної похибки не можна вважати помилку округлення операції множення незалежної від похибок операндів, то вираз для дисперсії обчислювальної похибки матиме вигляд



де Kε0,ε1 та Kε0,ε2 - кореляційні моменти величин ε0,ε1 та ε0,ε2 відповідно.

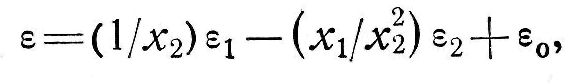
Для ймовірнісного підходу квадратично підсумовуємо СКЗ похибок операндів. Довірчі межі трансформованої похибки визначаємо для заданої довірчої ймовірності та для нормального розподілу або розподілу Стьюдента даної похибки.

*4.3. Ділення.* Обчислювальну рівність операції ділення подаємо у вигляді



Вважаємо, що ε0=0.

Якщо операнди мають похибки ε1, ε2, то обчислювальну похибку операції поділу можна визначити зі співвідношення

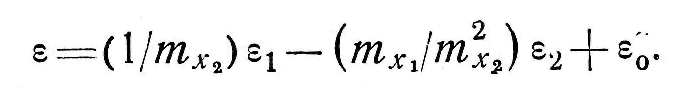


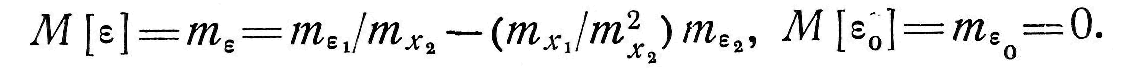
яке показує, як формується обчислювальна похибка і під час операції поділу.

Вважаємо, що ε0=0.

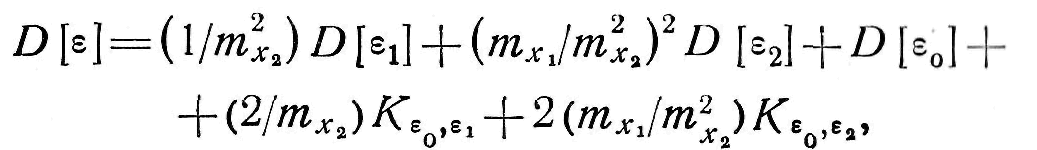
Зі співвідношення випливає, що сумарна обчислювальна похибка операції поділу залежить від операндів, що беруть участь в операції. Отже, для формування обчислювальної похибки операції поділу також важливо, яке розподіл мають старші розряди операндов.

При ймовірнісному аналізі похибок





Дисперсія величини ε



де Kε0,ε1 та Kε0,ε2 - кореляційні моменти величин ε0,ε1 та ε0,ε2 відповідно.

Звідси випливає, що для визначення числових характеристик обчислювальних похибок операцій множення та поділу в режимі з фіксованою точкою необхідно знати чисельні характеристики похибок операндів і кореляційні моменти величин ε0,ε1 та ε0,ε2, приймаючи при цьому з деяким припущенням величини ε1 та ε2 незалежними.

Для ймовірнісного підходу квадратично підсумовуємо СКЗ похибок операндів. Довірчі межі трансформованої похибки визначаємо для заданої довірчої ймовірності та для нормального розподілу або розподілу Стьюдента даної похибки.