Прикладные проблемы оптимизации

УДК 004.021

Васильев И.С.[1], **Жгун Т.В.**[2]

[1] ЗАО «ЭЛСИ», Великий Новгород, Россия
[2]Новгородский государственный университет им. Ярослава Мудрого, Великий Новгород, Россия

ИСПОЛЬЗОВАНИЕ ВЫЧИСЛЕНИЙ С МАНТИССОЙ ПЕРЕМЕННОЙ ДЛИНЫ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ОЦЕНКИ ПАРАМЕТРОВ СИСТЕМ

Аннотация

В работе рассматривается проектирование вычислительного процесса оценки параметров систем по результатам измерений методом максимального правдоподобия с использованием мантиссы переменной длины. Представлена методика определения оптимальной длины мантиссы на каждой итерации решения задачи вычисления параметров конкретной системы, основывающаяся на данных исследования серии итерационных процессов определения параметров смоделированных систем.

Ключевые слова

Мантисса переменной длины; метод максимального правдоподобия; оценка параметров; вычислительная оптимизация; моделирование вычислительных задач; квазиньютоновские методы.

Vassiliev I.S.[1], **Zhgun T.V.**[2]

[1] ZAO ELSI, Velikiy Novgorod, Russia
[2]Novgorod State University, Velikiy Novgorod, Russia

USING THE VARIABLE-LENGTH MANTISSA FOR MODELS' PARAMETERS ESTIMATION

Abstract

The article is dedicated to the problem of computing process design for computing processes used for systems parameters estimation based on measurements using maximum likelihood method, using variable length mantissa calculations. A new method for the optimal mantissa length for each iteration of peculiar system parameters estimation task determination is presented, it is based on a series of iteration processes for modelled systems parameters estimation research.

Keywords

Variable-length mantissa;, maximum likelihood method; parameters estimation; numerical optimization, computational tasks modelling; quasi-Newton methods.G.

Введение

Задачи оценивания параметров и характеристик систем по результатам измерений имеют широкое распространение на практике. Особенно важное место они занимают на всех этапах экспериментальной отработки и эксплуатации объектов ракетно-космической техники, при навигационно-баллистическом обеспечении полетов космических аппаратов, при разработке систем автономной навигации, в ходе летных испытаний, в построении базовых схем промышленной электроники и др.

Методы оценивания параметров нелинейных многооткликовых моделей на основании выборки экспериментальных данных достигли высокого совершенства и подвергались обстоятельным исследованиям [1,2]. Но необходимо отметить, однако, что вопросы улучшения точностных и вычислительных характеристик указанных методов продолжают оставаться актуальными как в целом, так и применительно к конкретным группам задач. Особенно следует отметить, что

авторы исследований в области оптимизационных методов, как правило, ограничиваются описанием теоретической части метода и практически не рассматривают аспекты проектирования вычислительного процесса, что является необходимым этапом при технической реализации. В то же время следует отметить, что при разработке встраиваемых (*Embedded*) специализированных малогабаритных систем проектирование вычислительных процессов имеет ключевое значение, так как оптимальность производства вычислений имеет критичное значение для общей производительности системы.

Можно выделить весьма обширный класс задач, для которых вектор оцениваемых параметров системы неизвестен, однако известна оценка параметров его многомерного распределения, а также области значений его компонентов. Знание параметров распределения позволяет использовать методики оптимизации вычислительных процессов при оценке коэффициентов модели. Область значений, определяемая, как правило, на основе теоретических соображений, подкрепляемых эмпирически, обуславливает применение методов оптимизации с ограничениями в виде неравенств. Это позволяет в некоторой мере защититься от сходимости метода оптимизации к локальному экстремуму.

Целью исследования является построение вычислительно оптимального итерационного процесса нахождения решения задачи оценивания параметров системы по результатам измерений.

Задача оценки коэффициентов моделей радиоэлектронных компонентов рассматривается в качестве типичного примера такой задачи. Эта задача весьма актуальна сама по себе ввиду важности задачи моделирования радиоэлектронных компонентов с учетом разброса их характеристик в целях совершенствования производственного процесса.

Использование метода максимального правдоподобия для оценки коэффициентов *SPICE*-моделей полупроводниковых приборов

В качестве математической модели полупроводниковых приборов на настоящий момент, как правило, используются *SPICE*-модели. Они стали стандартом для систем схемотехнического моделирования и входят в состав почти всех САПР, используемых для проектирования схем [3]. *SPICE*-модель отражает взаимосвязь между задаваемыми и наблюдаемыми электрическими характеристиками электронных компонентов и схем.

Модель представляет собой систему уравнений, получаемых с помощью закона Кирхгофа для тока относительно каждого узла за исключением базового [3]. После выполнения ряда операций система дифференциальных уравнений представляется в общем виде:

$$Y \cdot v = j\,, \qquad (1)$$

где Y – матрица дифференциальных узловых проводимостей, v – вектор узловых напряжений, j – вектор токов. Обычно напряжения используют в качестве задаваемых характеристик (входных параметров), а токи – в качестве наблюдаемых (выходных параметров). Определяемыми параметрами системы (коэффициентами модели) являются физические характеристики изделия, такие как: электрическое сопротивление различных областей, токи насыщения, коэффициенты эмиссии и так далее.

В математическом смысле проблема оценки *SPICE*-параметров является нетривиальной задачей на многофакторную (в современных моделях приборов содержится более сотни *SPICE*-параметров!) условную оптимизацию, успешное решение которой обеспечивается правильным использованием соответствующих оптимизационных методов и алгоритмов их реализации. Также важен правильный подбор параметров оптимизационного процесса и возможность получения экспериментальных данных должного качества.

Методология оценивания параметров в основном базируется на применении метода максимального правдоподобия, который, как показывает практика, является одним из самых эффективных методов определения параметров. Метод максимального правдоподобия (ММП) изначально был разработан Р.Фишером применительно к модели объекта статического типа. Процедура его применения предусматривает составление функции правдоподобия и получение оптимальной оценки путем решения системы нормальных уравнений, выражающей условие максимума этой функции. По смыслу нормальная система представляет необходимое условие оптимальности, характерное для прямых методов оптимизации.

Всюду далее будем предполагать, что наблюдаемый в эксперименте вектор выходных параметров Y – непрерывная случайная величина, имеющая плотность $g(Y; b)$, где b — неизвестный вектор параметров из множества допустимых параметров B. Основная цель метода — максимизация функции правдоподобия $L(b) = g(Y; b)$ как функции аргумента b при фиксированном значении наблюдаемой случайной величины. Так как логарифмическая функция является монотонно возрастающей функцией аргумента, значение b, максимизирующее значение целевой функции $L(b)$, будет максимизировать и $ln(L(B))$. Так как логарифмирование часто упрощает объектную функцию, задача может быть сведена к максимизации натурального логарифма функции правдоподобия (ЛФП).

$$\frac{\delta \ln L(b)}{\delta b} = 0 \ . \qquad (2)$$

Оценка максимума правдоподобия (ОМП) вектора коэффициентов b – это значение b^*, удовлетворяющее всем ограничениям, при котором функция правдоподобия достигает максимального значения, если оно существует.

При относительно слабых ограничениях на вид функции правдоподобия оценки ММП состоятельны и асимптотически эффективны, что особенно справедливо для выборок большого объема. При малых выборках ОМП обычно не обладают какими-либо оптимальными свойствами, оценки не будут ни эффективными, ни несмещенными. Однако многие исследователи показывают, что метод максимума правдоподобия дает приемлемые оценки во многих ситуациях, что является сильным аргументом в пользу применения этого метода для расчета оценок коэффициентов моделей.

Можно привести ещё одно наглядное определение оценки максимального правдоподобия: **оценка максимального правдоподобия *b**** – это такие значения вектора параметров b, которые обеспечивают наибольшую вероятность наблюдения выборки в окрестности действительно наблюдаемой выборки, полученной экспериментально [1].

Если функция правдоподобия (или ее логарифм) имеет профиль по параметрам близкий к квадратичному (например, в случае независимых случайных величин, имеющих нормальное распределение), то ее максимум может быть легко найден посредством решения системы линейных уравнений относительно неизвестных параметров. Однако на практике часто оказывается, что профиль функции правдоподобия весьма далек от квадратичного. Потому весьма обоснованным является применение методов многокритериальной оптимизации для поиска искомой оценки. Объектная функция для таких методов строится на основе логарифма функции правдоподобия.

Методы оптимизации условно можно поделить на методы, не использующие сведений о производной объектной функции (поисковые), методы, использующие сведения о первой производной объектной функции (методы первого порядка) и методы, использующие сведения о второй производной объектной функции (методы второго порядка). Поисковые методы обоснованно применяются в тех случаях, когда производную объектной функции невозможно выразить аналитически либо это слишком сложно. В нашем случае производная в большинстве случаев аналитически выражается и, как правило, не является громоздкой. Потому традиционно для решения подобных задач используются градиентные методы, а именно всевозможные модификации метода Ньютона.

Ньютоновские методы, в частности, метод Ньютона-Рафсона, при некоторых допущениях относительно функции $L(b)$ и достаточно аккуратно выбранном начальном приближении, демонстрирует локальную квадратичную сходимость к решению b^* уравнения (2). Однако они требуют вычисления гессиана функции на каждой итерации, что приводит к дополнительным вычислительным затратам. В качестве альтернативы можно рассмотреть квазиньютоновские методы, которые заменяют гессиан функции некоторым его приближением. Эмпирические исследования показали, что наилучшим соотношением качества получаемого решения и вычислительной сложности обладает метод Ньютона-Гаусса [1, 2, 4], который и применяется далее.

Однако даже квазиньютоновские методы в применении к оценке параметров *SPICE*-моделей требуют производства весьма большого объема вычислений, что может привести к увеличению времени оценки коэффициентов сверх допустимого для данной системы. При использовании явно задаваемой модели в виде некоторой системы уравнений вида $y=F(x,b)$, где y – вектор выходных параметров, x – вектор входных параметров, b – вектор оцениваемых коэффициентов, сложность задачи растет по меньшей мере квадратично при увеличении размерностей используемых данных.

Ещё хуже ситуация обстоит в тех случаях, когда модель задаётся неявно, и вычисление вектора выходных параметров само по себе требует итерационного решения оптимизационной задачи. При вычислении итерационного приближения вектора выходных параметров $b^*(n)$ на каждом шаге n значения компонент вектора вычисляются также с помощью оптимизационного алгоритма поиска корня. Следовательно, и общая вычислительная сложность увеличивается на два порядка. К примеру, оценка параметров явной многоотказовой *SPICE*-модели *npn*-транзистора на персональном компьютере (процессор *IntelCore 2 Duo*, 2,1 *GHz*, *DDR*3 4,0 Гб) требует около 20-30 минут. Следовательно, такие вычислительные схемы не могут быть широко востребованы.

В зависимости от задачи и условий реализации для преодоления проблемы вычислительной сложности могут быть применены экстенсивные методы, такие как построение высокопроизводительных кластеров на месте либо использование облачных вычислений. Первый подход может оказаться неоправданно дорогим и неприемлемым по габаритам установки и энергопотреблению. Второй подход потребует организации надёжного канала связи с облаком, что в ряде случаев представляет неоправданные

технические сложности, кроме того, аренда облака со специализированным программным обеспечением также требует денег. Распараллеливание вычислительного процесса также является вариантом улучшения производительности. Многие задачи весьма успешно могут быть разделены на подзадачи, которые раздаются отдельным процессорам. Во многих случаях именно этот подход является оптимальным, однако он требует усложнения и удорожания аппаратной части, а также увеличивает затраты энергии, требующиеся для выполнения задачи.

Если рассматривать аппаратную реализацию в виде встраиваемой системы, то решение проблемы экстенсивными методами (т.е. увеличением вычислительных мощностей) оказывается в ряде случаев неоправданным или невозможным в силу ограничений на энергопотребление и/или на количество процессоров и их мощность. Также для всех реализаций немаловажную роль часто играет и ценовой фактор, что в совокупности определяет ограниченную применимость такого метода решения проблемы вычислительной сложности. В таких случаях наиболее предпочтительным или даже единственным способом реализации требований является интенсивная оптимизация вычислительного процесса, при которой стремятся к получению результата с необходимой точностью при значительном снижении вычислительных затрат.

Задача оптимального выбора точности вычислений

Известно, что вычислительная сложность машинной арифметической операции напрямую зависит от абсолютной погрешности, с которой производятся арифметические операции. Чем ниже эта погрешность, тем ниже и погрешность получаемого результата, но в то же время больше ресурсов требуется на проведение вычислений. Оптимальный подбор этой погрешности и есть один из вариантов оптимизации вычислений.

В качестве рабочего формата чисел для проведения вычислений в таких задачах традиционно используются числа с плавающей запятой. Числа с фиксированной точкой, несмотря на их очевидные достоинства, используются редко в силу нерационального использования памяти при необходимости представления чисел с большим разбросом значений. Следует отметить, впрочем, что соображения, высказанные ниже, могут быть применены и при организации вычислений с фиксированной точкой, и в ряде случаев это будет даже проще.

Машинное число с плавающей точкой представляется в виде: $X_c = M \cdot b^p$, где M – значение мантиссы, p – значение порядка, b – основание системы счисления. В современных ЭВМ повсеместно применяется двоичная арифметика (b=2), но для теоретических соображений удобно работать с десятичным представлением числа (b=10), так как рассматриваемые подходы одинаково работают независимо от выбранного основания.

Относительная погрешность машинных чисел с плавающей точкой не превышает значения, называемого «машинным эпсилоном» [5,6].

$$\varepsilon = \frac{1}{2} b^{1-t}, \quad (3)$$

где b – основание системы счисления, t – длина мантиссы, p – длина поля порядка, $t + p$ – длина используемой разрядной сетки. Для знака числа выделяется ещё один, дополнительный разряд, не входящий в мантиссу. Величина машинного эпсилон является важнейшей константой машинной арифметики, характеризующей точность компьютерных вычислений. Таким образом, обеспечение оптимальной точности вычислений можно свести к подбору значений p и t, то есть длин поля порядка и мантиссы в представлении числа. Как видно из формулы (3), абсолютная точность представления чисел в компьютере достижима при бесконечно больших значениях p и t, что невозможно практически и не требуется для большинства реальных задач.

Общепринято, что $p << t$. К примеру, в стандарте представления чисел IEEE754 для чисел одинарной точности длина мантиссы t = 23, число разрядов, отводимых на порядок числа p = 8, для представления чисел двойной точности t = 52, p = 11 бит. Исходя из этого, разумно положить некоторое постоянное значение p и варьировать только длину мантиссы t. В частности, в большинстве систем принимается $p > 1{,}5 \cdot \log_b t$ [5]. Для размера поля порядка p потребуем далее выполнение условия (4).

$$p \geq \log_b (t+1). \quad (4)$$

Таким образом, задача оптимального выбора точности вычислений сводится к задаче выбора длины мантиссы t. В качестве основных идей оптимизации используются следующие обнаруженные эвристики:

1. Точность вычислений должна быть ненамного больше точности, обусловленной погрешностью метода и/или неустранимой погрешностью исходных данных. Обыкновенно используемые числовые типы данных с плавающей точкой *float* и *double* могут давать избыточную в конкретном случае точность.
2. Поскольку теоретически точность результата растёт с каждой итерацией, то и точность проведения вычислительных операций имеет смысл увеличивать постепенно, с каждой итерацией.

Идея постепенного увеличения мантиссы, с которой производятся вычисления в итерационном процессе, высказана в [6,7], там же доказана эффективность данного подхода для ряда задач. Там же указывается на трудности, возникающие при определении рабочей мантиссы для каждой итерации в общем случае, в частности, необходимость подгонки параметров формулы, которая предложена для определения этой длины, исходя из оценки погрешности вычислений на итерации. Настоящее же исследование предлагает способ регулирования длины мантиссы в итерационном процессе оценки параметров нелинейной многооткликовой модели, основанный на статистическом анализе параметров решения группы задач с близкими исходными данными и определения точности решения на каждой итерации.

Использование мантиссы переменной длины в итерационном процессе

Нетрудно определить выгоду от использования вычислений с мантиссой, увеличивающейся от итерации к итерации. Анализируя итерационный процесс, можно видеть, что вычислительная сложность любой итерации зависит от длины используемой мантиссы, то есть эту сложность можно представить как *F(t).* Очевидно, что эта функция монотонно возрастает при увеличении аргумента, ибо вычислительная сложность увеличивается с увеличением длины рабочей мантиссы. Если процесс требует *N* итераций, то его общая сложность будет равна *N·F(t)* в случае использования мантиссы постоянной длины и $\sum_{i=0}^{N} F(t_i)$ при использовании мантиссы переменной длины, причем $\sum_{i=0}^{N} F(t_i) < N \cdot F(t)$, так как $t_{i+1} \geq t_i$ и только на последней итерации $t_N = t$. Выгода в смысле вычислительных ресурсов будет тем большей, чем больше будет разность между длинами мантисс на разных итерациях и суммарная разность $\Delta = \sum_{i=0}^{N} \Delta t_i$, где $\Delta t_i = t_{i+1} - t_i$.

До начала вычислений задается максимальная длина мантиссы $t_{\max}$, которая определяется требуемой точностью оценки определяемых параметров (числом желаемых верно значащих цифр результата) с запасом в 1-2 разряда. В рамках каждой итерации вычислительного процесса (основной итерации оптимизационного метода), используется постоянная длина мантиссы t_i, которая определяется начальной длиной мантиссы t_0 и номером текущей итерации *i:* $t_i = T(t_0, i)$. Итерационный процесс завершается, если для двух последовательных итераций *s*-1 и *s* относительное изменение всех коэффициентов вектора оцениваемых параметров не превышает заданную точность, определяемую желаемым числом значащих разрядов результата $NSIG \approx t_{\max}$.

Для некоторых итерационных методов нахождения минимального значения функции, в частности, для метода Ньютона, мы имеем в распоряжении теоремы о сходимости и точности итерационного процесса. В общем же случае до решения задачи об этих характеристиках нельзя ничего сказать. Однако для весьма широкого класса задач мы обладаем некоторыми сведениями относительно возможных исходных данных.

Зная закон распределения искомых коэффициентов модели, мы можем сгенерировать сколь угодно векторов, принадлежащих тому же распределению. Для каждого из этих векторов, задавшись планами экспериментов, сведениями об ошибках измерения мы можем сгенерировать наборы экспериментальных данных, подавая вектора коэффициентов и значения входных сигналов, указанные в плане эксперимента, на вход моделей и внося возмущения в результат. По полученным наборам смоделированных экспериментальных данных мы можем произвести итерационную оценку параметров для каждого из смоделированных объектов. Совокупность итерационных приближений определяемых параметров на каждом шаге позволяет характеризовать течение итерационного процесса в целом и определить требуемую каждой итерацией длину мантиссы.

Проанализировав порядки величин, характеризующих изменение компонентов оцениваемого вектора для двух последовательных итераций, можно определить наиболее вероятное значение порядка изменений вычисляемых параметров для каждой итерации. Эти значения, в частности, позволят проверить эффективность метода: в случае сильного отклонения от линейной сходимости в худшую сторону можно говорить о слабой эффективности метода, являющегося, возможно, следствием программных ошибок и/или неустранимых погрешностей данных и/или вычислительных ошибок. Можно утверждать, что порядок изменений компонентов вектора оцениваемых коэффициентов на каждой итерации определяет минимальную длину мантиссы, с которой вычисления на данной итерации пройдут без потери точности. Проведённые исследования на многооткликовых *SPICE*-моделях полупроводниковых диода и транзистора с использованием модифицированного метода Ньютона-Гаусса показали, что распределение порядка изменений оцениваемого вектора на каждой итерации характеризуется небольшой дисперсией. Эмпирически выяснилось, что для каждой итерации 90% изменений имеют порядки,

отличающиеся не более, чем на единицу.

По полученному эмпирическому распределению изменений оцениваемого вектора коэффициентов следует определить, какова должна быть на каждой итерации необходимая длина мантиссы. Для этого имеет смысл рассмотреть цену ошибки, то есть неверного определения данной величины:

- в случае использования на каждой итерации избыточной длины мантиссы метод потребует больше времени, и это снизит эффект от применения переменной длины мантиссы в целом, но количество итераций будет таким же, как и для мантиссы постоянной длины, заведомо превышающей необходимую. Назовём эту ошибку ошибкой первого рода;
- в случае недостаточной длины мантиссы на некоторой итерации метод потребует больше итераций, чем мог бы, и сможет сойтись лишь тогда, когда длина мантиссы будет соответствовать требуемой точности. Назовём эту ошибку ошибкой второго рода.

Легко видеть, что ошибка первого рода является наиболее безобидной, ошибка же второго рода в большинстве практических случаев влияет на производительность гораздо сильнее. Можно сделать ещё несколько очевидных соображений:

1. Верхняя граница длины мантиссы устанавливается исходя из потребной точности результата, точности входных данных и эмпирической точности метода, определяемой алгоритмом.
2. При возрастании номера итерации, длина мантиссы может либо увеличиваться, либо оставаться прежней.

Для итерации i задается длина мантиссы t_i, исходя из того, что на этой итерации ожидается порядок изменений коэффициента вектора параметров p_i. Если порядок изменений вычисляемого коэффициента вектора параметров больше ожидаемого, то вычисления на итерации i производятся без потери точности.

Рассмотрим применение мантиссы переменной длины для решения задачи оценки коэффициентов моделей. Таблицу рекомендованных длин мантисс для каждой итерации решения задачи определения параметров системы по экспериментальным данным получаем следующим образом:

1. Получаем достаточно представительную (не менее 200) выборку смоделированных решений задачи. Для каждого компонента вектора оцениваемых коэффициентов формируется таблица итерационных изменений значений компонента, т. е. таблица конечных разностей первого порядка для выполненных итераций;
2. Над разностями производится преобразование порядка, возвращающее значение порядка числа в его стандартном машинном представлении с плавающей точкой;
3. На основе анализа порядков изменений для каждой итерации определяем **рабочее значение порядка изменений (рабочее значение порядка)** – такое значение порядка изменений, что при моделировании выборки решений бОльшая часть порядков первых разностей итерационных приближений (обычно, 90-95%) не менее этой величины;
4. На каждой итерации определяем длину мантиссы в зависимости от рабочего значения порядка изменений для этой итерации (правила пересчёта приведены ниже);
5. Если для разных коэффициентов вектора оцениваемых параметров длины мантиссы для фиксированной итерации различны, то может применяться одна из стратегий выбора длины мантиссы на этой итерации: либо выбор максимальной требуемой мантиссы для всех коэффициентов, либо выбор усреднённой мантиссы, либо иной алгоритм.

Можно говорить о том, что использование мантиссы переменной длины не снижает точности в том случае, если модуль разности результатов, полученных при использовании мантиссы переменной длины и точной арифметики, будет не больше, чем максимально разрешённая погрешность [5]. При этом количество итераций и точность на каждой итерации могут не совпадать, хотя, как правило, они тоже близки. Появление в вычислительном процессе единичных ошибок первого и второго рода, как правило, не оказывает существенного влияния на точность итогового результата и время вычисления. В качестве точного решения используется решение, полученное с использованием заведомо достаточно длинной мантиссы.

В [5] описан подход к определению числа верно значащих цифр (ВЗЦ) итерационного решения: «В случаях, когда погрешность проистекает главным образом из-за округления, оценка точности вычисления производится так: результат оценивается с использованием более точной арифметики и сравнивается количество совпадающих знаков с исследуемым вариантом. Разумно предположить, что при более точной арифметике результат точнее».

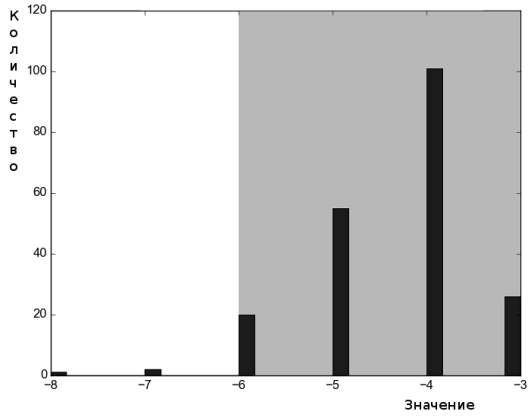
Получение оптимального (рабочего) значения порядка изменений

В основу принципа выбора рабочего значения порядка положены следующие обнаруженные эмпирические закономерности:

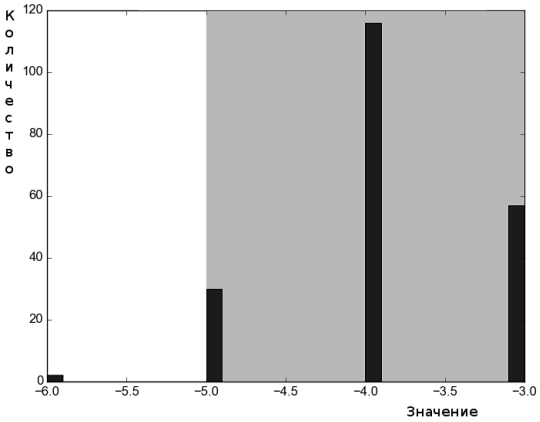
1. в рассматриваемых вычислительных процессах на каждой итерации порядки изменений имеют небольшой разброс;
2. зависимость длины мантиссы от модуля

порядка изменения является убывающей функцией, потому при выборе более низкого рабочего порядка изменений мы получаем увеличенную длину мантиссы.

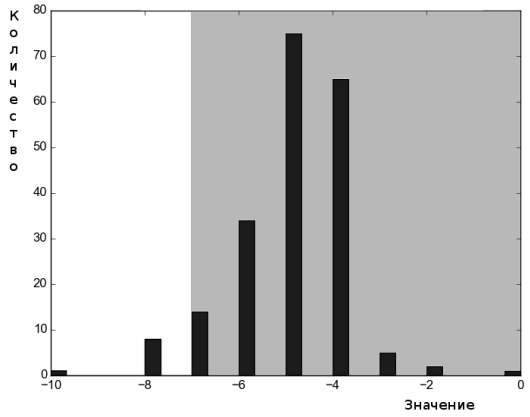
Выберем в качестве рабочего порядка медиану распределения порядков с некоторым запасом. Полученное на основании такого рабочего порядка значение длины мантиссы в большинстве случаев будет оптимальным.



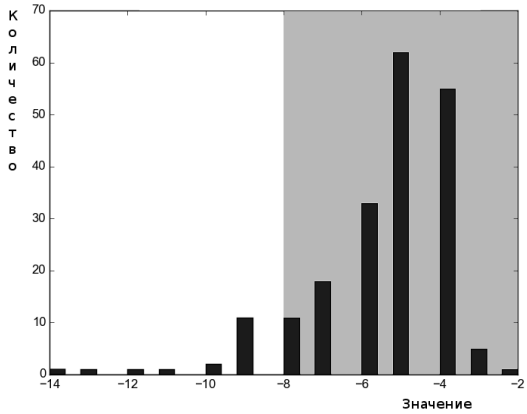
А. Итерация 1



Б. Итерация 2



В. Итерация 3



Г. Итерация 4

Рис. 1. Гистограммы распределения порядков изменений коэффициента 2 на итерациях 1-4

Если выбранная длина мантиссы не оптимальна, в небольшом количестве случаев будем иметь ошибку первого рода, которая не сказывается на производительности слишком сильно, и с крайне малой вероятностью – ошибку второго рода, которая сильно скажется на снижении производительности, но не повлияет на точность результата.

На рис. 1 представлены гистограммы распределения порядков изменений при оценивании второго коэффициента модели полупроводникового диода на первых четырех итерациях. Можно видеть, что распределения значений порядков напоминают нормальное и основная часть наблюдений укладывается в диапазон ±2 от математического ожидания.

Таблица 1 – Выбор рабочего значения порядка изменений (порог) в зависимости от доли покрываемых значений (процент) при определении коэффициента b[1]

Итерация	1					
Порог	-8	-7	-6	-5		
Процент	100	90	75	4		
Итерация	2					
Порог	-8	-7	-6	-5		
Процент	100	99	56	10		
Итерация	3					
Порог	-10	-9	-8	-7	-6	
Процент	100	94	92	52	5	
Итерация	4					
Порог	-11	-10	-9	-8	-7	-6
Процент	100	97	91	60	30	9
Итерация	5					
Порог	-12	-11	-10	-9	-8	-7
Процент	100	98	96	92	65	12
Итерация	6					
Порог	-13	-12	-11	-10	-9	-8
Процент	100	99	93	91	32	8
Итерация	7					
Порог	-15	-13	-12	-11	-12	-13
Процент	100	97	94	54	40	6

Поскольку мы работаем с дискретным распределением, задача анализа весьма сильно упрощается. Для каждой итерации мы составляем

таблицу, в которой каждому значению порядка сопоставляем число элементов распределения, равных или больших этого значения и процент, который это число составляет от общего числа случаев. Такая таблица напоминает таблицу квантилей, но имеет противоположный смысл – для каждого значения выбираются большие, а не меньшие члены распределения. В данном случае количество элементов есть число случаев оценки, для которых мантисса, определяемая данным пороговым значением, будет определена безошибочно либо с ошибкой первого рода. В таблицах 1 и 2 приведем выбор **рабочего значение порядка изменений** в зависимости от доли покрываемых значений. Можно видеть, что для первой итерации при выборе в качестве **рабочего значение порядка изменений** -4 в 96% случаев (таблица 1) оценки мантисса будет определяться безошибочно либо с ошибкой первого рода, при выборе порогового значения -3 – в 65% случаев и так далее.

Таблица 2 – Выбор рабочего значение порядка изменений (порог) в зависимости от процента покрываемых значений при определении коэффициента b[2]

Итерация	1							
Порог	-5		-4		-3		-2	
Процент	100		96		65		3	
Итерация	2							
Порог	-6		-5		-4		-3	
Процент	100		99		84		27	
Итерация	3							
Порог	-8		-7	-6	-5	-4	-3	
Процент	100		99	98	88	61	12	
Итерация	4							
Порог	-10	-8	-7	-6	-5	-4	-3	-2
Процент	100	99	95	88	72	35	3	1

Далее выбираем **рабочий пороговый процент** — желаемая доля покрываемых значений. Здесь нельзя рекомендовать единый подход, выбор стратегии зависит от конкретных условий. Чем выше пороговый процент, тем медленнее будут идти вычисления в целом, но меньше будет разница между вычислительными процессами решения разных задач. Снижение порогового процента (порога) ускоряет вычисления в среднем, но в некоторых случаях сильно снижает скорость вычислений, и чем порог ниже, тем таких случаев будет больше (так как вероятность ошибок второго рода будет расти). Увеличение порогового процента замедляет вычисления в среднем, но делает их более предсказуемыми в смысле времени, снижая вероятность ошибок второго рода, но увеличивая вероятность ошибок первого рода.

Это даёт основания применить следующую стратегию: чем ниже производительность используемых вычислительных средств, тем менее величина выбираемого порогового процента и наоборот.

В целом задача выбора оптимальной длины мантиссы на каждой итерации может быть рассмотрена как оптимизационная. Предлагаемый подход позволяет от решения задачи оптимизации длины мантиссы на каждой итерации перейти к задаче оптимизации всего итерационного процесса путём поиска оптимального рабочего процента закрываемых значений. Выбрав рабочий пороговый процент, мы получаем таблицу рабочих значений порядка изменений. В таблице 3 представлены такие значения для рабочего порогового процента 90% для четырёх итераций

Таблица 3 – Рабочие значения порядка изменений для рабочего порогового процента 90% для первых двух коэффициентов

Итерация	Рабочее значение порядка изменений для коэффициента *b*[1]	Рабочее значение порядка изменений для коэффициента *b*[2]
1	-7	-4
2	-7	-5
3	-8	-6
4	-9	-7
5	-9	-
6	-10	-
7	-12	-

Определение длины мантиссы на основе рабочих значений порядка изменений

В стандартизованной арифметике сложение и вычитание чисел с плавающей точкой выполняются в предположении, что число с меньшей длиной мантиссы определяет точность итогового результата. При выполнении арифметических операций, если числу с меньшей длиной мантиссы соответствует и меньшее значение порядка, то при выравнивании порядков в результате мантисса числа с меньшим порядком сдвигается вправо на разницу порядков операндов и младшие выталкиваемые разряды мантиссы при этом теряются.

Пусть t_1, t_2 – длины мантисс операндов, p_1, p_2 – их порядки. Пусть при этом $t_1 > t_2$ и $p_1 > p_2$. В этом случае мантисса первого операнда будет сдвинута вправо на p_1-p_2 разрядов, и младшие разряды будут при этом потеряны. Кроме того, длина мантиссы результата принимается равной наименьшей длине мантиссы операндов (это актуально и при одинаковых порядках чисел). При выполнении предположений длина мантиссы результата $t_{res} = min\ (t_1,t_2)$, и ещё t_1-t_2 разрядов будут отброшены.

Порядок результата $p_{res} = max\ (p_1,p_2)$.

Для выбора оптимальной длины мантиссы необходимо определить её минимальную длину, такую, что при сложении двух чисел не будет происходить потери точности. Для этого рассматриваем число с меньшей длиной мантиссы (значение, получаемое на предыдущей итерации) как точное число, но с нулями в младших разрядах. Длина мантиссы, необходимая для исключения потери точности при проведении сложения/вычитания, будет определяться формулами:

$$t = \max(t_1, t_2 + |p_1 - p_2|) \quad (5)$$

При наступлении переполнения мантиссу надлежит нарастить ещё на 1 разряд, дабы избежать потери младшего либо старшего разряда. Компонент $|p_1 - p_2|$ – число разрядов, на которые сдвигается мантисса меньшего числа вправо. Этот компонент добавлен, чтобы избежать потери этих разрядов, округляемых при обычной операции сложения.

Приведём пример: Пусть имеются два числа, 1.001 и 1.0001. Мантисса первого числа составляет 4 разряда, второго – 5 разрядов, порядки одинаковы. Пользуясь формулой (5) устанавливаем мантиссу в 5 разрядов и получаем итоговый результат 2.0011. Стандартный же подход к вычислениям с учётом погрешности дал бы нам результат 2.001.

На основании формулы (5) составляем формулу пересчёта порогового значения порядка в значение длины мантиссы в итерационном процессе:

$$t_s = \max(t_{s-1}, t_{delta} + |p_{s-1} - p_{delta}|) \quad (6)$$

В этом случае t_{s-1} – это длина мантиссы на предыдущей итерации, t_{delta} – длина мантиссы изменения вычисляемого значения коэффициента на текущей итерации, при которой не происходит существенной потери точности, p_{s-1} – порядок значения на предыдущей итерации, как правило, равный порядку начального значения, p_{delta} – это порядок изменений. Длина мантиссы изменения, обычно, принимается равной 1-3. Для начальной итерации $t_{s-1} = t_0$ – это число ВЗЦ начального значения плюс некоторый запас (обычно 1-2 разряда).

Формула (6) предлагается в качестве основной формулы пересчёта порогового значения порядка в значение длины мантиссы в итерационном процессе. Очевидно, что чем больше разница между порядками чисел, тем больше длина мантиссы результата, что следует из (5). Таким образом, при условии $p_1 > p_2$ функция $t(p_2)$ будет убывающей.

Исследование эффективности метода

С целью определения эффективности метода был произведён ряд практических исследований применения вычислений с переменной точностью – с регулируемой длиной мантиссы, определяемой на каждой итерации вышеописанным способом. Ниже приводятся краткие результаты исследования применения предлагаемой методики для оценки коэффициентов модели полупроводникового диода [8, 9].

Для каждой группы задач с близкими исходными данными моделировалось 205 решений с фиксированием всех основных сведений по каждой итерации с мантиссой фиксированной длины, обусловленной итоговой точностью результата. На основе вычисляемых итерационных значений была составлена таблица конечных разностей до четвертого порядка включительно. Затем для всех значений конечных разностей было выполнено преобразование порядка *p(x)*, возвращающее значение порядка.

Выше, на рис. 1 приведены диаграммы для порядков изменений на разных итерациях для второго коэффициента модели. Можно видеть малый разброс порядка изменений на начальных итерациях и увеличивающийся разброс порядка – на дальнейших итерациях. Распределение порядков значений дискретно, что значительно упрощает его анализ.

Далее составляем таблицу, где каждому значению порядка изменения сопоставляем число элементов, не превосходящих это значение, и процент, который такие элементы составляют от общего количества, как было описано выше (табл. 1 и 2) . В качестве рабочего порогового процента возьмём 90%. Затем рассчитаем длины мантисс по формуле (6) (табл. 4 и 5).

Таблица 4 – Расчёт длины мантисс для коэффициента b[2]

Итерация	Рабочее значение порядка изменений на итерации	Расчёт длины мантиссы	Длина мантиссы
1	-4	*max* (2,1+|1 - 4|)	4
2	-5	*max* (4,1+|1 - 5|)	5
3	-6	*max* (5,1+|1 - 6|)	6
4 и далее	-7	*max* (6,1+|1 - 7|)	7

Далее задаём стартовую длину мантиссы (мантиссу для начальных значений) в 2 разряда, порядок 1-го коэффициента – 1, второго – минус 7. Точность входных данных в данном эксперименте – 6 верно значащих цифр, определена точностью измерений. Исходя из этого, длину мантиссы можно ограничить семью разрядами. Следует отметить, что в программной реализации длина мантиссы может автоматически увеличиться в случае возникновения переполнения.

Таблица 5 – Расчёт длины мантисс для коэффициента b[1]

Итерация	Рабочее значение порядка изменений на итерации	Расчёт длины мантиссы	Длина мантиссы
1	-7	*max*(2,1+|-7 +7|)	2
2	-7	*max*(2,1+|-7 +7|)	2
3	-8	*max*(2,1+|-7 +8|)	2
4	-9	*max*(2,1+|-7 +9|)	3
5	-9	*max*(3,1+|-7 +9|)	3
6	-10	*max*(3,1+|-7 +10|)	4
7 и далее	-12	*max*(4,1+|-7 +12|)	6

Для первой итерации длину мантиссы начальных значений мы полагаем 2. Длину мантиссы изменений (число ВЗЦ) полагаем 1 (этот параметр рекомендуется увеличивать при увеличении порядка сходимости). Увеличение длины мантиссы прекращается, когда эта длина станет равной количеству ВЗЦ входных данных.

Для анализа эффективности метода было произведено 45 решений задачи определения параметров полупроводникового диода с применением мантисс в 20 разрядов, 10 разрядов, 7 разрядов, а также с мантиссой переменной длины, определяемой предложенным алгоритмом. Для компьютера с конфигурацией *IntelPentiumCore i5, 4Gb RAM* были получены решения задачи определения параметров полупроводникового диода, характеристики которых отражены в таблице 6.

Таблица 6 – Сравнительный анализ эффективности вычислений с использованием разных длин мантисс

Мантисса	Число итераций	Затраченное время, с
$t = 20$	7	141
$t = 10$	7	79
$t = 7$	7	64
Мантисса переменной длины	7	54

Точность полученных решений совпадает (определяется точностью исходных данных, однако время, затраченное на решение задачи, существенно отличается. Это говорит о существенной большей скорости вычислений при использовании мантиссы переменной длины. Следует учесть, что расчёты производились для весьма простой модели. Чем сложнее будет оцениваемая модель, тем больше будет и получаемый выигрыш во времени. Кроме этого, в качестве программных средств использовалась библиотека GMP и процессор общего назначения. При использовании же специализированных математических сопроцессоров во встраиваемых решениях, поддерживающих вычисления с мантиссой с переменной длины аппаратно, выигрыш во времени должен возрасти ещё больше.

Следует отметить наличие специализированных программных библиотек, позволяющих производить вычисления с мантиссой, меняющейся в ходе работы программы. Это, в первую очередь, библиотека GMP, которая используется в проектах на C++ и Python. Для встраиваемых решений в ряде случаев может быть рекомендована разработка специализированных высокоэффективных библиотек, оптимально реализующих вычисления с мантиссой переменной длины

Заключение

Проведенное исследование продемонстрировало, что применение в итерационном процессе мантиссы, длина которой определяется по предложенному алгоритму, не ухудшает точности решения, но дает весомый выигрыш в части экономии вычислительных ресурсов. Реализация таких вычислений имеет широкие возможности как в случае единичного и мелкосерийного производства целевых изделий в программируемых логических интегральных схемах, так и в случае массового производства устройств в интегральных схемах специального назначения — математических сопроцессорах, поддерживающих вычисления с мантиссой переменной длины. Общая цена устройства при этом снизится, что может быть важным фактором для таких областей, как, например, разработка самонаводящихся реактивных снарядов, беспилотных летательных аппаратов и т. п.

Статистический анализ порядков изменений итерационных приближений при определении параметров смоделированных систем может дать информацию не только для реализации вычислений с мантиссой переменной длины при определении параметров конкретной системы. Такой анализ, в частности, позволяет также увидеть разброс изменений значений вектора оцениваемых параметров при итерационном процессе, выявить количество случаев, приводящих к слишком медленной сходимости, и помочь в грамотном выборе параметров итерационного процесса.

Литература

1. *Bard J.* Nonlinear Parameter Estimation. New York: Academic Press, 1974, 341p.
2. *Englezos P., Kalogerakis N.* Applied Parameter Estimation for Chemical Engineers. Boka-Raton: CRC Press, 2000, 460 p.
3. *Петров М. Н., Гудков Г. В.* Моделирование компонентов и элементов СБИС: Учебное пособие. Великий Новгород: НовГУ им. Ярослава Мудрого, 2006, 584с.
4. *McLachlan G. J., Peel D.* Finite Mixture Models. New York: Wiley, 2000, 419p.
5. *Higham N. J.* Accuracy and stability of numerical algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 1996, 675 p.

6. *Jonathan Richard Shewchuk* Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates // Discrete & Computational Geometry. 1997. №18(3). Pp. 305–363.
7. *Гриневич А. И.* Метод оценки погрешности округлений значений вычисляемой функции, основанный на варьировании длины мантиссы в арифметике с плавающей запятой. Диссертация на соискание учёной степени кандидата физико-математических наук. Москва: МФТИ, 2013.
8. *Васильев И. С., Попов С. А.* Оценивание SPICE-параметров с заданной точностью // Вестник Новгородского Государственного Университета. 2014. №1.81. С. 20–22.
9. *Васильев И. С., Попов С. А.* Метод экстракции параметров ИС и расчёт их статистических характеристик // IV Международная заочная научно-техническая конференция «Информационные технологии. Радиоэлектроника. Телекоммуникации (ITRT-2014). Тольятти, 2014.
10. *Cramer H.* Mathematical Methods of Statistics. Princeton, New Jersey: Princeton University Press, 1946, 575c.
11. *Воеводин В. В.* Вычислительные основы линейной алгебры. М.: Наука, 1977. 304 с.
12. *Годунов С. К. и др.* Гарантированная точность решения систем линейных уравнений в евклидовых пространствах. Новосибирск: Наука, 1988, 456с.

References

1. *Bard J.* Nonlinear Parameter Estimation. New York: Academic Press, 1974, 341p.
2. *Englezos P., Kalogerakis N.* Applied Parameter Estimation for Chemical Engineers. Boka-Raton: CRC Press, 2000, 460 p.
3. Petrov M. N., Gudkov G. V. Modelirovanie komponentov i elementov SBIS: Uchebnoe posobie [IC components and elements modelling: coursebook]. Novgorod State university, 2006, 584p..
4. *McLachlan G. J., Peel D.* Finite Mixture Models. New York: Wiley, 2000, 419p.
5. *Higham N. J.* Accuracy and stability of numerical algorithms. Philadelphia: Society for Industrial and Applied Mathematics, 1996, 675 p.
6. *Jonathan Richard Shewchuk* Adaptive Precision Floating-Point Arithmetic and Fast Robust Geometric Predicates // Discrete & Computational Geometry. 1997. №18(3). Pp. 305–363.
7. Grinevich A. I. *Metod otsenki pogreshnosti okruglenii znachenii vychislyaemoi funktsii, osnovannyi na var'irovanii dliny mantissy v arifmetike s plavayushchei zapyatoi. Diss. kand. fiziko-matematicheskikh nauk* [Method of counted function rounding error estimation, based on mantissa length changing using floating-point arithmetics. PHd in Physics & Mathematics diss.]. Moscow, MFTI, 2013.
8. Vassiliev I. S. Popov S. A. *Otsenivanie SPICE-parametrov s zadannoi tochnost'yu* [SPICE-parameter estimation with given accuracy]. *Vestnik Novgorodskogo Gosudarstvennogo Universiteta* 2014. №1.81. С. 20–22.
9. Vassiliev I. S. Popov S. A. *Metod ekstraktsii parametrov IS i raschet ikh statisticheskikh kharakteristik* [Method of IS parameter extraction and their statistical characteristics estimation]. *IV Mezhdunarodnaya zaochnaya nauchno-tekhnicheskaya konferentsiya «Informatsionnye tekhnologii. Radioelektronika. Telekommunikatsii (ITRT-2014)* [Proc. of IV International Extramural science-technic conference "Informational technologies. Radioelectronics. Telecommunications. (ITRT-2014)"], Togliatti, 2014.
10. Cramer H. Mathematical Methods of Statistics. Princeton, New Jersey: Princeton University Press, 1946, 575p.
11. Voevodin V. V. Vychislitel'nye osnovy linejnoj algebry [Calculation basis of linear algebra].Moscow, Nauka Publ., 1977. 304p.
12. Godunov S. K. et al. Garantirovannaja tochnost' reshenija sistem linejnyh uravnenij v evklidovyh prostranstvah [Guaranteed accuracy of linear equation systems solving in Euclidean fields]. Novosibirsk, Nauka Publ., 1988, 456p.

Поступила: 24.03.2017

Сведения об авторах:

Васильев Илья Сергеевич, дипломированный инженер, аспирант, ЗАО «ЭЛСИ», Великий Новгород, mksd@list.ru

Жгун Татьяна Валентиновна, кандидат физико-математических наук, доцент, Новгородский государственный университет им. Ярослава Мудрого, Великий Новгород, zhtv@mail.ru

Note on the authors:

Vassiliev Ilia, Sophisticated Engineer, PhD student, ZAO ELSI, mksd@list.ru

Zhgun Tatjana, Candidate of Physical and Mathematical Sciences, Associate Professor, Novgorod State University, zhtv@mail.ru