Электронный журнал «Труды МАИ». Выпуск № 49 www.mai.ru/science/trudy/

УДК 004.27

К обоснованию оптимизации итерационных процессов в неоднородных вычислительных системах

А.С.Марков

В работе излагаются общие теоретические обоснования целесообразности использования в системном программировании нового подхода к оптимальной организации итерационных вычислительных процессов в неоднородных вычислительных системах за счет рационального динамического подбора точности вычислений.

В обоснование подхода предлагаются некоторые математические зависимости, позволяющие количественно оценивать возможности ускорения решения задач без потери точности результата с сохранением свойств устойчивости и сходимости вычислительного метода. Для частного случая метода Ньютона полученные формулы пригодны для практического применения.

системное программирование; итерационный вычислительный процесс; неоднородные вычислительные системы; точность вычислений.

Основные исходные предпосылки предлагаемой работы заключаются в следующем.

Вычислительно ёмкие программы никогда не бывают линейными и почти всегда реализуют тот или иной итерационный вычислительный процесс. В теории и практике оптимизации таких процессов по времени внимание специалистов традиционно сосредоточено на машинонезависимой проблеме рационального динамического подбора сеточного шага метода. Если же при переходе к другой разрядной сетке (другой ЭВМ) характеристики метода неприемлемо ухудшаются, то, как правило, предпринимаются дорогостоящие и не связанные с методом непосредственно меры: перепрограммирование, использование другого вычислительного оборудования, аппаратурное или программное удвоение мантиссы и т. п.

С другой стороны, при выполнении трудоемких вычислений вручную варьирование точности вычислений от шага к шагу и от итерации к итерации применяется почти всегда, хотя делается это лишь на интуитивных и, как правило, справедливых основаниях в естественном стремлении уменьшить трудозатраты.

Третьей исходной предпосылкой предлагаемых в настоящей статье исследований явилось стремление представить в более явном виде неиспользуемые резервы инструментально-прикладных систем, в которых совмещаются черты языков программирования высокого и низкого уровней.

Таким образом, нашей целью является обоснование возможности создания новых и модификации существующих вычислительно ёмких вычислительных методов, обладающих информацией о тех еще не утилизованных запасах сходимости и устойчивости на каждом этапе проведения расчётов, которые могут быть использованы при программировании для оптимизации вычислительного процесса по времени за счёт рационального подбора точности вычислений.

Пусть мы обозреваем корректное применение некоторого известного итерационного численного метода М , заданного устойчивым вычислительным процессом P с исходными данными x_1 , промежуточными результатами $x_i(i=1,2, \ldots, N\text{-}1)$ и с результатом x_Nпо схеме

$$\tilde{x}_{i+1} = P_i(\tilde{x}_1, \tilde{x}_2 \ \ldots, \tilde{x}_i) + \delta_i,$$

реализующейпрограммно с некоторыми погрешностями округлений (вообще – вычислений) δ_i непрерывный в евклидовом пространстве $\Re^n$ оператор

$P_i{:}X_1{\times}X_2{\times}\ldots{\times}X_i{\to}X_{i+1}$

Где $X_1 = \Re^n$, $x_{i+1}{\in}X_{i+1}$ $(i=1,2, \ldots, N\text{-}1)$ и $\delta_i{\in}X_{i+1}$

так, что при малых $\|\delta_j\|$ $(j = 1, 2, \ldots, i)$ норма $\|\tilde{x}_j - x_j\|$ также мала и не сильно зависит от i , а Nзадается заранее либо устанавливается в самом вычислительном процессе.

При алгоритмизации и программировании таких процессов предлагается, не ограничиваясь установлением принципиально важного свойства вычислительного процесса – слабой зависимости δ_iот номера iитерации, предпринимать попытки исследования запаса устойчивости. Этот запас устойчивости в некоторых случаях может и сильно зависеть от i как в бỏльшую, так и в мẻньшую сторону, индицируя тем самым возможность оптимизации в той или иной мере вычислительного процесса по времени его выполнения. Нередко, как это обычно и бывает в экстремальных задачах, дополнительным источником такого запаса сходимости может явиться свойство стационарности оператора P_i $(P_i{=}\ P)$ при $i=1,2, \ldots, N\text{-}1$.

Большинство известных вычислительных методов минимизации функций многих переменных являются итерационными и в общем виде могут быть охвачены схемой

$$x^{k+1} = x^k + P\,(\alpha(F(x^k)g(F'(x^k)))), \beta(F'(x^k))) \qquad (1) \quad ,$$

Где зависимость β от матриц вторых производных F'' нередко лишь косвенная и при собственно вычислении P не участвует вовсе: требуется, например, для обоснования метода, их существование с ограничениями на положительную определенность матрицы. Во многих случаях в некоторой области G при определенных довольно жестких условиях, смягчение которых не входит в нашу задачу (например, в методах наискорейшего спуска требуется при любых $x,\ y\in G$ выполнения неравенства $m\|y\|^2 \le (F''(x)y, y) \le M\|y\|^2$ с константами $0 < m \le M$, метод сходится со скоростью геометрической прогрессии со знаменателем $q < 1$. Ошибки техники программирования, вносимые заменой абстрактных процессов P и F машинными программами $\approx P$ и $\approx F$, аппроксимирующими алгоритмические $\tilde{P}$ и $\tilde{F}$, мы также выносим за рамки рассмотрения.

Сохраняя возможность вернуться к анализу более общего процесса (1), обратим внимание на схему

$$x^{k+1} = x^k + P(F(x^k)),$$

охватывающую весьма широкий круг методов – особенно, если учитывать тот факт, что зависимость от $F'(x)$ в (1) часто остается лишь в обосновании метода, а при его алгоритмизации дифференцирование сложных $F(x)$ выполняется численно с приемлемой для метода точностью.

Разумеется, в тех вычислительных итерационных (рекуррентных) процессах, для которых накопление вычислительной погрешности уже тщательно исследовано и получены удобные для практического применения в расчетах количественные оценки зависимости роста погрешности от i, использование соответствующего математического методического аппарата может лишь ещё более упростить предлагаемый нами анализ.

Для супервычислений (подразумеваются вычисления с критическим для возможностей некоторой используемой ЭВМ объёмом арифметических операций с плавающей точкой) нас особенно должен интересовать довольно обширный круг задач, в которых

$$t(P) \ll t(F(x)),$$

где функция t(γ) определяет временную оценку машинной реализации γ.

Отправляясь от некоторой точки $\tilde{x}$ исходного машино-запрограммированного процесса, будем считать, что вычисления как $\tilde{P}$, так и $\approx P$ (как мы уже условились, двойная тильда относит помечаемый символ к модифицированному в смысле динамического изменения погрешности способу алгоритмизации исходного процесса) вносит пренебрежительно малые в сравнении с вычислением как

$\tilde{F}$, так и $\approx F$

погрешности за счет существенно большего объёма вычисления последних. В обоснование этого допущения достаточно напомнить, что, например, в линейном методе Ньютона вычисление Pозначает единственную арифметическую операцию. Это позволяет предположить, что в G оператор P удовлетворяет условию Гёльдера

$$\omega\left(\delta_{\mathrm{P},}\mathrm{P}_{\sim F(\sim x)}\right) = \max_{|h|<\delta P}\max_{F(x)}\left|P\left(\sim F(\sim x)\right) - P(F(x))\right| \leq p(\delta p)^{\alpha}$$

(где символом ω обозначен модуль непрерывности функции), и, аналогично,

$$\omega\left(\delta_{\mathrm{P},}\mathrm{P}_{\approx F(\sim x)}\right) = \max_{|h|<\delta P}\max_{\approx F(x)}\left|P\left(\approx F(\sim x)\right) - P\left(F(x)\right)\right| \leq p(\delta p)^{\alpha}$$

с одной и той же константой p и с $\alpha < 1$.

Пусть $t(P) = Tt(F)$ с константой T (напоминаем, что P и F предполагаются стационарными по i), а запас геометрической сходимости $q \leq Q < 1$или $1 - q \geq \beta > 0$. Обозначая как h^* шаг, следующий за шагом h от рассматриваемой точки пространства $\tilde{x}$, перейдем к количественным критериям подбора точности вычислений:

$\sim h/\sim h^* = q \leq \beta < 1$,

$\approx h/\sim h^* = (\sim h + \varepsilon(h))/\sim h^* = \varepsilon(h)/\sim hq$

Пусть k_q– та доля от $(1 - q)$, которой мы соглашаемся оплачивать оптимизацию расчета за счет погрешности отклонения от шага исходного процесса, а k_h – относительная доля для рассматриваемого шага итерации, на которую вследствие этого будет возникать ограничение на изменение погрешности расчета. Для любых вычислительных процессов со степенной сходимостью

$$k_{\approx h} < \frac{q + k_q(1-q) - q}{q + k_q(1-q)} = \frac{k_q(1-q)}{q + k_q(1-q)},$$

или

$$k_{\approx h} < \frac{k_q\beta}{q + k_q\beta} = \frac{1}{1 + q/k_q\beta} = k_q\beta\left(1 + \frac{1-(q+k_q\beta)}{q + k_q\beta}\right),$$

Тогда

$$\sim h^*/\approx h = \sim hq/(\sim h + k_{\approx h}\sim h) = \frac{q}{1 + k_{\approx h}},$$

откуда может быть получено неравенство

$k_{\approx h} \le 1 - \frac{q}{q+k_q(1-q)}$,

Например, для группы процессов минимизации функций методом Ньютона

$P(x) = P\big(f(x), f'(x)\big) = f(x)/f'(x)$,

имеем:

$P(x+\delta) = P(x) + P'(x)\delta + O(\delta^2)$,

$$P'(x) = \frac{f'(x)\big/f'(x) + f'(x)\big/f''(x)}{f'(x)} = \frac{f'(x)\big/f''(x)}{f'(x)^2},$$

$$P(x+\delta) = f(x)\big/f'(x)\left[1 - \left(f''(x)\big/f'(x)\right)\delta\right] + \delta + O(\delta^2),$$

Поэтому (обозначая далее K=k_h) ,

$$K < \frac{f'(x)\big/f(x) - f''(x)\big/f'(x)}{1 - \frac{q}{q+k_q(1-q)}},$$

Таким образом,

$$K < \frac{\left(f'(x)\right)^2 - f(x)f''(x)}{f(x)\big/f'(x)}\left(\frac{q}{k_q(1-q)} + 1\right),$$

Последнее неравенство уже пригодно для практического использования при алгоритмизации и программировании. Способ его вывода от общего к частному позволяет надеяться на возможность проведения аналогичного анализа и других вычислительных процессов. Если способы машинных и программных округлений позволяют определять несмещенные оценки погрешностей, то из интуитивных соображений можно надеяться, что количество шагов в модифицированном процессе на интересующих нас участках корректного понижения точности вычислений в среднем должно оставаться тем же самым, что и в исходном процессе.

В развитии подобных исследований могут быть заинтересованы не только специалисты в области методов вычислений, прикладные и системные программисты, но и специалисты по разработке нового вычислительного оборудования, ориентированного на решение вычислительно емких задач.

В многопроцессорных вычислительных системах со специализированными арифметическими процессорами применение предлагаемых здесь подходов может позволить во многих случаях более эффективно распределять загрузку узлов и устройств в системе.

При проектировании новых программно-технических комплексов (ПТК) и, в частности, ПТК с массовым параллелизмом, изложенные соображения могут повлиять на состав вычислительного оборудования при оптимизации соответственных количеств, включаемых в ПКТ дорогостоящих полноразрядных и более дешевых малоразрядных процессоров с целью улучшения показателя производительность/стоимость ПКТ.

В любом случае, по нашему мнению, представляется интересным в научном плане и полезным практически при исследовании и алгоритмизации итерационных методов проводить дополнительный анализ итерационных вычислительно емких процессов в предлагаемом нами направлении, а именно: (1) стремиться выявить и при возможности выразить в явной форме количественную оценку неиспользуемого в вычислительном алгоритме динамически переменного запаса сходимости метода, и там, где это будет признаноцелесообразным, (2) выработать алгоритм рационального динамического подбора точности, (3) не нарушающий устойчивости метода и (4) не ухудшающий заметно скорости его сходимости. Выполненное на основании подобного анализа практическое программирование модифицированного метода может позволить, как показывают предварительные проведенные нами имитационные эксперименты, (5) сократить на десятки, а иногда и на сотни процентов загрузку наиболее дорогостоящего вычислительного оборудования и существенно повысить показатель производительность/стоимость использования ПТК при решении задач описываемого в статье класса итерационных методов.

Исследования в указанном направлении могут быть отнесены к области перспективных информационных технологий и имеют целью разработку и внедрение новых базовых средств информационного и математического обеспечения вычислительных систем и комплексов, позволяющих заметно повысить для них показатель производительность/стоимость в вычислительно емких приложениях.

Автор сердечно признателен академикам М.Р. Шура-Буре, Л.Н. Королеву и профессору И.В. Пузынину за неизменный интерес к теме и столь ценные дискуссии, без которых не была бы выполнена эта работа.

Литература

1)Марчук Г.И.Методы вычислительной математики. Наука. –М.: 1989.

2)Вопросы математической эксплуатации вычислительных машин. –Вып. 2, №1. –М: 1965.

Марков Александр Сергеевич,главный научный сотрудникОАО «Научно-исследовательский центр электроннойвычислительной техники», д.ф.-м.н,e-mail:markov@nicevt.ru