УДК 513.88

ОПТИМАЛЬНОЕ ЧЕРЕДОВАНИЕ ОСНОВНОГО И МОДИФИЦИРОВАННОГО ПРОЦЕССОВ НЬЮТОНА - КАНТОРОВИЧА

Б.А.Вертгейм

1. Задача об оптимальном управлении вычислительным процессом, например, поиском решения нелинейного уравнения, является весьма актуальной. Эта задача может ставиться по-разному; как правило, речь идёт о достижении максимально возможной точности для некоторого заданного класса уравнений с учётом ограниченности, например, времени счёта, или же речь идет о достижении заданной точности при решении уравнения из рассматриваемого класса таким образом, чтобы в процессе счёта затратить минимальное количество ограниченного ресурса.

В настоящей работе изучаются задачи такого рода в следующей постановке: выбраны класс нелинейных уравнений и семейство приближенных методов для решения уравнений этого класса способом последовательных приближений. В качестве важного конкретного примера будем в основном рассматривать семейство, состоящее из двух наиболее известных способов приближенного решения - основного и модифицированного методов Ньютона - Канторовича (см. [1 - 4]). Считаем известными затраты некоторого ограниченного ресурса (времени) на один шаг каждого из методов; точнее, известна зависимость таких затрат от номера итерации и, возможно, от порядка, в котором чередуются эти методы. Требуется установить оптимальное чередование итераций методов из заданного семейства, то есть применять заданные методы в таком порядке, чтобы получать наибольшую точность при заданных затратах или заданную точность при наименьших затратах. Такая постановка задачи, очевидно, связана с идеями Л.В.Канторовича, предложивнего т. наз. модифицированный метод, который, хотя и обладает меньшей скоростью сходимости, чем основной метод, но зато требует на одну итерацию меньших затрат. В связи с этим в данной статье будет уточнено одно замечание А.М.Островского (см. [4], стр. 171) о модифицированном процессе (в [4] подчёркнута лишь одна сторона дела - сравнительно медленная сходимость упомянутого процесса).

2. Установим необходимые обозначения и определения. Мы рассматриваем решение функционального уравнения

$$P(x) = 0, \tag{1}$$

где $P : X \to Y$ - нелинейное отображение, X и Y - пространства Банаха - с помощью варианта метода Ньютона-Канторовича, описываемого формулой (сравн. [9, 13])

$$x_{k+1} = x_k - [P'(x_{n_i})]^{-1} P(x_k), \quad n_i \le k < n_{i+1}. \tag{2}$$

Здесь $\{n_i\}$, $i = 0,1,2,\ldots$, - заданная возрастающая последовательность целых чисел, $n_0 = 0$, $n_i < n_{i+1}$, составленная из тех номеров n_i, для которых вычисляется производная $P'(x_{n_i})$. Итак, в силу формулы (2), в ходе итераций, начиная от номера n_i до n_{i+1} ($i = 0,1,\ldots$), применяется модифицированный (если $n_{i+1} - n_i > 1$) процесс, затем при $k = n_{i+1}$ вновь вычисляется производная P', на основе которой совершается ещё несколько шагов модифицированного, вообще говоря, процесса и т. д.

Обозначая

$$k_1 + k_2 + \cdots + k_p = m, \tag{3}$$

$$n_{i+1} - n_i = k_{i+1}, \qquad \beta = (k_1, k_2, \cdots, k_p),$$

$$m = k_1 + k_2 + \cdots + k_p = n_p, \qquad x_m = x_m(\beta) = x_m(k_1, \cdots k_p), \tag{4}$$

будем говорить, что приближение $x_m(\beta)$ получено из x_0 смешанным процессом β, составленным из p модифицированных серий, длины которых равны k_ν, $\nu = 1,2,\ldots p$. Оператор N_k, который заданному начальному элементу $x_0 \in X$ ставит в соответствие x_k - k -ое приближение модифицированного (при $k > 1$) процесса Ньютона-Канторовича (при условии, что этот процесс осуществим, если начать его с заданного x_0) назовем оператором процесса длины κ ($\kappa = 1,2,\dots$). С помощью операторов N_κ процесс (2) с учетом обозначений (3) можно записать следующим образом:

$$x_{n_i} = N_{\kappa_i}(x_{n_{i-1}}) \qquad , \quad i = 1,2,\dots,p \; . \qquad (5)$$

Отметим, что операторы N_κ и N_ℓ при $\kappa \neq \ell$, как правило, не перестановочны, хотя имеется и исключение из этого правила (см. дальше следствие 3 леммы 2.)

Предположим, что осуществление (вычисление) оператора N_κ требует затрат некоторого ресурса в количестве $c + c'(\kappa-1)$, где $0 < c' \leqslant c_1$ - заданные числа; c - цена одного шага основного процесса ($\kappa = 1$); c' - цена одного шага модифицированного процесса.

Мы рассматриваем следующие задачи, см. [18].

ЗАДАЧА A_1. (I) Дано конкретное уравнение (I), его решение x^* и начальное приближение к x^* - элемент x_0, для которого выполнены условия теоремы Л.В.Канторовича, см. [3] теорема 6 ($1XVIII$); (2) Даны также два целых положительных числа p и m, $0<p\leqslant m$. Найти p целых чисел $\kappa_i \geqslant 1$, $i=\overline{1,p}$, удовлетворяющих условию (3), так, чтобы $d(x_m(\beta), x^*) = \|x_m(\beta) - x^*\|$ было минимальным.

ЗАДАЧА A_2. (I) Даны уравнение (I) и x_0, как в задаче A. (П) Дано число $Q>0$ - ресурс.

Найти числа p, m, κ_i ($i=\overline{1,p}$), исходя из требований, подобных задаче A_1 (на этот раз числа p и m не заданы), $cp + \sum_{i=1} c'(\kappa_i - 1) \leqslant Q$.

ЗАДАЧА Б. В условиях задачи А для заданного дополнительно $\Delta > 0$ найти числа p, m, κ_i ($i=\overline{1,p}$), связанные, как в формулах (3)-(4), так, чтобы

$$\|x_m(\beta) - x^*\| \leqslant \Delta, \qquad cp + \sum_{i=1}^{p} c'(\kappa_i - 1) \to \min \; . \qquad (6)$$

В связи с этими задачами отметим, что способы сравнения эффективности разных приближенных процессов изучались в книге [4] и в статьях [9,10], как правило, на основе оценок асимптотической скорости сходимости. В отличие от этого асимптотического подхода мы сравниваем процессы, учитывая конечное, прежде всего, "небольшое" число итераций, а это особенно интересно для практических вычислений. Очевидно, что решение сформулированных задач для конкретно заданного оператора P является, как правило, более трудным делом, чем просто отыскание корня x^*. Поэтому особенно интересен вопрос об устойчивости характеристик оптимальных решений задач А-Б при изменении оператора P. Мы не рассматриваем здесь таких постановок вопроса, в которых речь идёт сразу о целом классе операторов P и применяются минимаксные критерии, подобные применённым для одномерного случая в книге [7], гл. 4.

3. Изложим результат, дающий решение задачи А в вещественном случае ($X = Y = R$) и показывающий, что при обычных ограничениях для всех x_0 из некоторой окрестности корня x^* оптимальное поведение - одно и то же.

ТЕОРЕМА I. Пусть задано вещественное уравнение (I), имеющее корень x^*. Пусть функция P дифференцируема в окрестности корня x^*, $P'(x^*) \neq 0$ и пусть существует конечная вторая производная $P''(x^*)$. Тогда найдется такая окрестность $J = (x^* - \delta, x^* + \delta)$ корня, что, каково бы ни было начальное приближение $x_0 \in J$, задача A_1 имеет следующее решение для системы данных (P, x^*, x_0) :

$$\overline{K_i} = \left[\frac{m+i-1}{p}\right] = \begin{cases} q, & \text{если } 1 \leq i \leq n - r, \\ q+1, & \text{если } n - r < i \leq n, \end{cases} \qquad (7)$$

$$m = q \cdot p + r, \qquad 0 \leq r < p .$$

Здесь q - частное, а r - остаток при делении числа m на p .

ДОКАЗАТЕЛЬСТВО. Используем формулу Тейлора с дополнительным членом в форме Пеано:

$$P(x) = P'(x^*)(x - x^*) + \frac{1}{2} P''(x^*)(x - x^*)^2 + o\left((x - x^*)^2\right) \qquad (8)$$

(напоминаем, что $P(x^*) = 0$). Введем следующие обозначения:

$$[P'(x^*)]^{-1}=\Gamma, \qquad x_n - x^* = \Delta_n, \qquad n=0,1,2,\dots,$$
$$\Gamma P''(x^*)\Delta_n = \theta_n, \qquad 1+K_i = z_i. \tag{9}$$

Как известно (см. [5], теорема II.I), существует такая окрестность точки x^*, что приближения метода Ньютона-Канторовича сходятся к x^*, начиная с любого элемента x_0 этой окрестности. Нетрудно показать, что эту окрестность можно выбрать так, чтобы сходились, начиная с любой её точки, приближения любого процесса (2) при произвольном выборе последовательности (n_i) (сравн. с [II]).

Методом математической индукции докажем, что в случае модифицированного процесса Ньютона, то есть если $x_K = N_K(x_0)$ (оператор N_K определен выше), справедливо равенство

$$\theta_K = \frac{1}{2}\theta_0^{K+1} + o(\Delta_0^{K+1}) \equiv \bar{\theta}_K + o(\Delta_0^{K+1}), \quad \bar{\theta}_K = \frac{1}{2}\theta_0^{K+1}, \quad K \geq 1. \tag{10}$$

При $K=1$ это соотношение получается с помощью следующей выкладки, в которой $P'(x_0)$ заменяется разложением по формуле Тейлора-Пеано, подобной (8):

$$x_1 = x_0 - \frac{P(x_0)}{P'(x_0)}, \quad \Delta_1 = \Delta_0 - \frac{P'(x^*)\Delta_0 + \frac{1}{2}P''(x^*)\Delta_0^2 + o(\Delta_0^2)}{P'(x^*) + P''(x^*)\Delta_0 + o(\Delta_0)} =$$

$$= \frac{1}{2}\Gamma P''(x^*)\Delta_0^2 + o(\Delta_0^2) \Longrightarrow \theta_1 = \frac{1}{2}\theta_0^2 + o(\Delta_0^2)$$

(в формуле для Δ_1 дробь после преобразования оказывается равной $Д = \Delta_0 - 0{,}5\Gamma P''(x^*)\Delta_0^2 + o(\Delta_0^2)$, что проще всего проверяется вычитанием из этой дроби указанного выражения $Д$ и приведением к общему знаменателю). Итак, соотношение (10) доказано для $K=1$. Предполагая его справедливость для некоторого $K \geq 1$, подобной же выкладкой устанавливаем справедливость (10) для $K+1$. Учитывая это соотношение и формулу (5), мы видим, что главные члены разности Δ_{n_i} получатся, если построить последовательность $\{\gamma_i\}: \gamma_i \approx \theta_{n_i}$

$$\gamma_i = \frac{1}{2}\gamma_{i-1}^{K_i+1}, \qquad i=\overline{1,p}.$$

Методом математической индукции доказываем, что

$$\gamma_n = 2^{-\mathcal{T}_n}\ \gamma_0^{z_1 z_2 \dots z_n}, \qquad n=1,2,\dots,p. \tag{II}$$

$$\mathcal{T}_n = 1+\sum_{\nu=2}^{n} z_n z_{n-1}\dots z_\nu \equiv \mathcal{T}_n(z) . \tag{I2}$$

В условиях теоремы I задано общее число итераций, см. формулу (3), и требуется так выбрать числа k_i, сумма которых равна m, чтобы показатель при γ_0 в формуле (II) при $n=p$ был наибольшим. По определению z_i, см. (9), имеем

$$z_1+z_2+\dots+z_p = m+p; \qquad \prod_1^p z_i \to max . \tag{3'}$$

Нам понадобится следующая лемма.

ЛЕММА I. Пусть заданы положительные целые числа c и $p>0$, и требуется найти p положительных целых чисел $z_1, z_2, z_3, \dots z_p$, в сумме составляющих c, так, чтобы произведение $z_1 z_2 \dots z_n$ принимало наибольшее возможное значение. (Задача (3')). Эта задача при дополнительном условии

$$0<z_1 \leq z_2 \leq z_3 \leq \dots \leq z_p \tag{I 3}$$

имеет единственное решение, которое можно охарактеризовать следующей формулой:

$$z_i^* = \left[\frac{c+i-1}{p}\right], \qquad i=1,2,3,\dots,p , \tag{I4}$$

или, что даёт то же самое, формулой

$$z_i^* = q,\ (i=1,2,\dots,n-r), \qquad z_i^* = q+1\ (i=n-r+1,\dots,n) , \tag{I4'}$$

где q - частное, а r - остаток при делении числа c на p, то есть $c=qp+r$.

ЗАМЕЧАНИЕ. Хорошо известно, что задача отыскания чисел $\{z_i\}$, удовлетворяющих условиям

$$\sum_{i=1}^{p} z_i = c, \qquad z_i>0, \qquad \prod_{i=1}^{n} z_i \to max ,$$

где $c>0$ – задано, числа c и z_i – не обязательно целые, имеет решение

$$x_i^* = \frac{c}{n}, \qquad i=1,2,\dots,n$$

(сравн. с формулой (I4)). В отличие от этого в условии леммы мы имеем дело с задачей целочисленного нелинейного программирования (имеем дополнительное требование целочисленности всех z).

ДОКАЗАТЕЛЬСТВО леммы. Множество наборов $(z_i)\ i=1,\dots,n$ (где z – целые положительные числа, в сумме равные $c>0$), очевидно, к о н е ч н о , и поэтому существование решения $\{z_i^*\}$ поставленной задачи максимизации в условиях леммы не вызывает сомнений. Докажем следующее свойство любого решения задачи (пока без учёта ограничения (I3)): для всех номеров i , j , $1\leq i,j\leq p$, выполнено неравенство

$$|z_i^* - z_j^*| \leq 1 . \tag{15}$$

Предположим противное: пусть существуют такие номера i , j , что $z_i^*-z_j^*\geq 2$ (очевидно, что пропуск знака модуля не ведёт к потере общности). Так как в этом случае

$$(z_i^*-1)(z_j^*+1) = z_i^* z_j^* + z_i^* - z_j^* - 1 > z_i^* z_j^*,$$

то новый допустимый набор $\{z_i'\}$, где

$$z_k' = z_k^* \ (k\neq i,\ k\neq j), \quad z_k' = z_i^*-1, \quad z_j' = z_j^*+1,$$

$$\sum_i z_i' = c, \qquad \prod_i z_i' > \prod_i z_i^*,$$

доставляет большее значение произведению, чем оптимальный (по условию) набор $\{z_i^*\}$, что противоречиво. Итак, в оптимальном решении любые два числа z_i^* и z_j^* либо равны, либо отличаются на единицу, то есть существует такое целое число $q>0$, что $\forall i\ z_i^*=q$ или $z_i^*=q+1$, причем $\exists i\,|\,z_i^*=q$. Введем множество (возможно, пустое) $M=\{i\,|\,z_i^*=q+1\}$, и пусть τ – число элементов в M , то есть число таких z_i^* , что $z_i^*=q+1$; очевидно, $0\leq\tau<n$. Тогда сумма всех z_i^* равна

$$\sum_{i=1}^{n} z_i^* = q(p-\tau)+(q+1)\tau = qp+\tau ,$$

и из условия $qp+r=c$ видно, что интересующие нас числа q и r определяются простым делением данных целых чисел и p. При этом q - частное, а r - остаток. Нетрудно видеть, что условие (13) приводит, с учетом свойства (15), к единственности решения, причём без этого условия решение единственно тогда и только тогда, когда p является делителем числа c (например, при $p=2$ и $c=5$ имеется два решения: $z_1^*=2$, $z_2^*=3$ и $z_1^*=3$, $z_2^*=2$; из них лишь первая пара удовлетворяет условию (13)). Лемма I доказана.

Продолжим доказательство теоремы. Мы пришли к задаче типа (3′), где $c=m+p$. Эта задача имеет, вообще говоря, много решений; их число равно C_p^k, см. доказательство леммы I. Как показывают формулы (11-12), мы должны выбрать такое решение задачи (3′), для которого показатель $\mathcal{K}_n$ имеет наибольшее возможное значение. Покажем, что это решение должно удовлетворять условиям (13).

В самом деле, пусть некоторое решение $\{\bar{z}_i\}$ задачи (3′) таково, что найдутся номера i, j, $i<j\leq p$, такие, что $\bar{z}_i>\bar{z}_j$. Построим другое решение $\{z_i^*\}$, переставляя числа, отвечающие номерам i и j, то есть

$$z_q^*=\bar{z}_{t(q)}, \quad t(i)=j, \quad t(j)=i, \quad t(q)=q \ (q\neq i,\ q\neq j)$$

(t - это перестановка, точнее - цикл ($i\,j$)).

Оказывается, что $\mathcal{K}_n(z^*)>\mathcal{K}_n(\bar{z})$. Действительно, применение транспозиции t оставляет без изменения все слагаемые сумм в формуле (12), отвечающие значениям $v\leq i$ (при $i=1$ множество таких слагаемых пусто) и значениям $v>j$; при этом слагаемые, соответствующие v таким, что $i<v\leq j$, увеличиваются вследствие того, что $z_j^*=\bar{z}_i>\bar{z}_j$. Итак, если $\bar{z}_i>\bar{z}_j$, то показатель $\mathcal{K}_n(\bar{z})$ может быть увеличен; это показывает, что из множества решений задачи (3′) надо выбрать то, которое удовлетворяет условию (13). Остаётся применить лемму I и, сравнивая порядки малости величин $\Delta_n=x_n-x^*$, см. формулы (9,10,11), получить утверждение теоремы I. При этом сначала устанавливается существование при заданном m промежутка $J_m=(x^*-\delta_m,\ x^*+\delta_m)$, радиус которого $\delta_m>0$ зависит, вообще говоря, от m, а затем очевидным дополнительным рассуждением устанавливается возможность выбора $\delta>0$ сразу для всех целых m так, что будут выполняться все утверждения теоремы I.

Теорема I доказана.

4. Подойдём к вопросу с несколько иной точки зрения. Рассмотрим задачу об оптимальном чередовании приближенных методов для вещественного уравнения, содержащего параметры:

$$f \equiv a - x + \lambda \varphi(x) = 0 , \tag{16}$$

где $\varphi(x)$ – вещественная функция, аналитическая в окрестности точки $x=a$, $a>0$ и λ – вещественные параметры. Решение этого уравнения $x^* = x^*(a,\lambda)$ может быть записано в удобном виде с помощью ряда Лагранжа, см. [14,5]. Успешное изучение приближенных методов для уравнений с параметром проводилось в ряде работ, см. [3,5], однако задача об оптимальном чередовании, по-видимому, ранее не рассматривалась.

Заслуживает внимания следующий частный случай уравнения (16):

$$f \equiv 1 - x + \frac{\lambda}{2} x^2 = 0 . \tag{17}$$

Это уравнение часто встречается в качестве модельного объекта при изучении метода Ньютона-Канторовича, см. [I - 3]. Оно имеет вещественные корни при $\lambda \le 0,5$. Изучение этого уравнения (17) оправдано тем, что задача об оптимальном чередовании здесь может быть эффективно решена не только для малых значений λ, как в случае (16), но и, скажем, для $\lambda = 0,5$, то есть на границе области существования решения.

Отметим следующий элементарный факт. Пусть $y=Q(x)$ – квадратичная функция. Замена переменных (x, y) на (ζ, η) по формулам

$$x = x_0 - \frac{Q(x_0)}{Q'(x_0)}\zeta , \qquad y = Q(x_0)\eta \tag{18}$$

преобразует квадратичную функцию $y=Q(x)$ к виду (сравн. (17))

$$\eta \equiv \bar{Q}(\zeta) = 1 - \zeta + \frac{\bar{\lambda}}{2}\zeta^2 , \quad \text{где} \quad \bar{\lambda} = Q(x_0)Q''(x_0)/Q'^2(x_0) . \tag{19}$$

При этом смешанному приближенному процессу $x_n(s,Q)$ (для уравнения $Q(x)=0$), исходящему из начальной точки $x_0(s,Q)=x_0$, отвечает (в смысле формул замены (18)) процесс $\zeta_n(s,\bar{Q})$ для уравнения $\bar{Q}(\zeta)=0$ с начальной точкой $\zeta_0(s,\bar{Q})=0$. Сравнивая формулы (17) и (19) и обозначая приближение смешанного процесса для (17) при $x_0=0$ через $x_n(s,\lambda)$, приходим к следующему утверждению.

ЛЕММА 2. Для уравнения (17)

$$x_{n+\ell}(\beta_1,\beta_2,\lambda) = x_n(\beta_1,\lambda) - \frac{f(x_n(\beta_1,\lambda))}{f'(x_n(\beta_1,\lambda))}\, x_\ell(\beta_2,\bar\lambda),$$

$$\bar\lambda = \lambda f(x_n(\beta,\lambda))\left[f'(x_n(\beta,\lambda))\right]^{-2}; \qquad n,\ell = 0,1,2,\ldots \tag{20}$$

СЛЕДСТВИЕ 1. Пусть $\mu_n \equiv \mu_n(\lambda)$ - приближения модифицированного процесса Ньютона для уравнения (17) при $\mu_0 = 0$, то есть

$$\mu_{n+1} = 1 + 0{,}5\,\lambda\mu_n^2 \quad . \tag{21}$$

Тогда (обозначение, - см. выше формулу (4)):

$$x_{n+\ell}(\beta,\lambda) = \mu_n + (\mu_{n+1} - \mu_n)(1-\lambda\mu_n)^{-1}\,\mu_\ell(\bar\lambda_n) \qquad \beta = (0,n)\ ,$$

$$\bar\lambda = \lambda(\mu_{n+1} - \mu_n)(1-\lambda\mu_n)^{-2}; \qquad n,\ell = 0,1,2,3,\ldots \tag{22}$$

СЛЕДСТВИЕ 2. Пусть $x^* \equiv x^*(\lambda)$ - решение (2). Тогда

$$x^*(\lambda) = \mu - \frac{f(\mu)}{f'(\mu)}\, x^*(\bar\lambda(\mu)), \qquad \bar\lambda(\mu) = \lambda\frac{f(\mu)}{f'^2(\mu)}, (0<\mu<x^*). \tag{23}$$

СЛЕДСТВИЕ 3. В случае $\lambda = 0{,}5$ (на границе области существования решения уравнения (17)) формулы (20) принимают вид (здесь $x_n(\beta) \equiv x_n(\beta, 0{,}5)$):

$\bar\lambda = \lambda$,

$$\bar x_{n+\ell}(\beta_1,\beta_2) = x_{n+\ell}(\beta_2,\beta_1) = x_n(\beta_1) + x_\ell(\beta_2) - \frac{1}{2}\, x_n(\beta_1)\, x_\ell(\beta_2)\,. \tag{24}$$

ДОКАЗАТЕЛЬСТВО. Первая из формул (20) выражает полугрупповое свойство динамического процесса итераций. Вторая формула (для λ) следует из (19). Формула (22) следует из (20) и (17). Формулу (23) можно вывести из (22), устремив $\ell \to \infty$. Тождество (23) легко проверяется также и непосредственно. Соотношение (24), из которого следует перестановочность процессов β_1 и β_2 при $\lambda=0{,}5$ выводится из (20) и (I9), поскольку при этом значении λ имеем

$$f(x)=[f'(x)]^2 .$$

Переходим к определению главных членов при малых λ ($\lambda\to 0$) функций $x^*(\lambda)-x_m(\beta,\lambda)$, дающих погрешность приближений смешанного процесса. Пусть заданы p целых чисел $K_1,K_2,\dots,K_p$, удовлетворяющих соотношению (3). Мы будем обозначать символом $x_{m(i)}(\beta_i,\lambda)(m(i)=\sum_{\nu=1}^{i}K_\nu)$ - приближение к решению x^* уравнения (I7), (вместо $x(\beta_1,\lambda)$ пишем $x(K_1,\lambda)$), полученные смешанным процессом с перевычислением производной на итерациях с номерами $m(i)$ $(m(0)=0,\quad i=0,\dots,n-1)$.

ЛЕММА 3.

$$x^*-x_{m(i)}(\beta_i,\lambda)=2^{-\mathfrak{T}_i(z)}\,\lambda^{d_i(z)-1}(1+o(1)),\quad z_\nu=K_\nu+1,$$

$$d_i(z)=z_1 z_2\dots z_i\ (i=\overline{1,n}),\quad \mathfrak{T}_i(z)=1+\sum_{\nu=2}^{n} z_n z_{n-1}\dots z_\nu \tag{25}$$

ДОКАЗАТЕЛЬСТВО. Из формул (23), (20) следует, что

$$x^*(\lambda)-x(\beta_{i+1},\lambda)=\frac{f(x(\beta_i))}{1-\lambda x(\beta_i)}\left[x^*(\lambda(\beta_i))-x(K_{i+1},\lambda(\beta_i))\right]. \tag{26}$$

По индукции нетрудно доказать, что

$$x^*-x(K,\lambda)=2^{-1}\lambda^K(1+o(1));\qquad f'(x(\beta_i))=-1+o(1),$$

$$f(x(\beta_i))=[x^*-x(\beta_i,\lambda)][1+o(1)]=\lambda^{-1}\ \lambda(\beta_i)\ . \tag{27}$$

Учитывая это и предполагая, что соотношение (25) выполнено для некоторого $i>1$ (для $i=1$ оно выполнено в силу (27)), мы получаем из (26):

$$x^*-x(\beta_{i+1},\lambda)=2^{d_i}\,\lambda^{\mathfrak{T}_i(x)-1}\,2^{-1}[\lambda(\beta_i)]^{x_{i+1}-1}(1+o(1))=$$
$$=2^{-(1+d_i x_i+1)}\ \lambda^{\mathfrak{T}_i(x)x_{i+1}-1}\ .$$

откуда можно получить формулы (26) для $i+1$. Лемма доказана.

Из этой леммы вытекает, что оптимальное чередование основного и модифицированного шагов для решения уравнения (I7) при малых λ такое же, как это описано в теореме I. (сравн. ф-лы (II,I2) и (25)).

ТЕОРЕМА 2. Пусть задано общее число итераций m и число p, $p<m$, итераций основного метода. Тогда существует такое число δ_m, зависящее от m, что для малых λ, $|\lambda|<\delta_m$ оптимальное чередование основного и модифицированного процесса в задаче А для уравнения (I7) даёт процесс типа $(K_1, K_2, \ldots, K_p)$, где $K_i=[(m+i-1)/p]$.

ДОКАЗАТЕЛЬСТВО. Теорема следует из лемм I и 3.

Аналогичная теорема может быть доказана и для более общего уравнения (I6) и притом без предположения об аналитичности функции φ. При этом можно пользоваться методами, предложенными в книге [5]. Не останавливаясь на этом подробнее, приведём лишь формулы, полученные с помощью ряда Лагранжа, из которых следует отмеченный факт, относящийся к уравнению (I6) при $m=3$ и $n=4$.

$$x^*-x_3(1,2)=0{,}5\varphi^3\varphi''^2\lambda^5+o(\lambda^5)$$

$$x^*-x_3(2,1)=0{,}5\varphi^2\varphi'^2\varphi''\lambda^5+o(\lambda^5)$$

$$x^*-x_4(1,3)=0{,}5\varphi^4\varphi''^3\lambda^7+o(\lambda^7)$$

$$x^*-x_4(2,2)=0{,}5\varphi^3\varphi'^3\varphi''^2\lambda^8+o(\lambda^8)$$

Перейдём к уравнению (I7) при $\lambda=0{,}5$

ТЕОРЕМА 3. Решение задачи А, для уравнения (I7) при $\lambda=0{,}5$ даёт любой процесс типа $\beta_\pi=(K_{\pi(i)})$ $i=\overline{1,n}$, где π - какая-нибудь перестановка множества $\{1,\ldots,n\}$, числа $\{K_i\}$ - те же, что и в теореме I.

ДОКАЗАТЕЛЬСТВО. Используя следствие 3 леммы 2, сравним $x_{p+q}(p,q)$ и $x_{p+q}(p+1,\ q-1)$:

$$
\begin{aligned}
&x_{p+q}(p,q)=\mu_p+\mu_q-\tfrac{1}{2}\mu_p\mu_q, \\
&x_{p+q}(p+1,\ q-1)=\mu_{p+1}+\mu_{q-1}-\tfrac{1}{2}\mu_{p+1}\mu_{q-1}.
\end{aligned}
\tag{28}
$$

Последовательность $\{\mu_p\}$ определена формулами (21); заметим, что при $\lambda = 0{,}5$ имеем

$$\Delta_p\equiv\mu_{p+1}-\mu_p\equiv 1+\tfrac{1}{4}\mu_p^2-\mu_p=\left(1-\tfrac{1}{2}\mu_p\right)^2.$$

Теперь из формул (28) следует, что

$$
\begin{aligned}
&x_{p+q}(p+1,\ q-1)-x_{p+q}(p,q)=\Delta_p-\Delta_{q-1}- \\
&-\tfrac{1}{2}\left[\mu_{p+1}\mu_{q-1}-\mu_p\mu_{q-1}+\mu_p\mu_{q-1}-\mu_p\mu_q\right]= \\
&=\left(1-\tfrac{1}{2}\mu_{q-1}\right)\Delta_p-\left(1-\tfrac{1}{2}\mu_p\right)\Delta_{q-1}= \\
&=\left(1-\tfrac{1}{2}\mu_{q-1}\right)\left(1-\tfrac{1}{2}\mu_p\right)^2-\left(1-\tfrac{1}{2}\mu_p\right)\left(1-\tfrac{1}{2}\mu_{q-1}\right)^2= \\
&=\left(1-\tfrac{1}{2}\mu_{q-1}\right)\left(1-\tfrac{1}{2}\mu_p\right)\left(\mu_{q-1}-\mu_p\right).
\end{aligned}
\tag{29}
$$

Вспомним, что последовательность $\{\mu_p\}$, определенная формулой (21) - возрастающая и ограниченная сверху числом 2. Поэтому из соотношения (29) следует, что

$$x_{p+q}(p+1,\ q-1)>x_{p+q}(p,q) \quad \text{если} \quad q-1>p \tag{30}$$

(и, кроме того, справедливы еще два соотношения типа (30), полученные заменой знака $>$ на знаки $=$, либо $<$). Итак, процесс ($p+1$, $q-1$) даёт лучшее приближение, чем процесс (p , q), если $p<q-1$. Отсюда с учетом перестановочности, выраженной следствием 3 леммы 2, получаем, что в оптимальном процессе (κ_1^* , κ_2^* ,..., κ_n^*) любые два числа κ_i^* и κ_j^* могут отличаться самое большое на единицу:

$$|\kappa_i^*-\kappa_j^*|\leqslant 1, \qquad i,j=1,2,\dots,n.$$

Сравнивая это неравенство с неравенством (15), которое играло основную роль в доказательстве леммы I, мы можем завершить рассуждение подобно тому, как это сделано в случае леммы I. На этот раз, конечно, нет условия (13), которое обеспечивало единственность решения.

Теорема 3 доказана. Вместе с предыдущей теоремой, она показывает, что в крайних случаях – при малых λ, и при $\lambda = 0,5$ (на границе области существования решения уравнения (17)) – оптимальное поведение в задаче A_1 определяется одинаково, см. формулы (7), причём при $\lambda = 0,5$ решение не единственно. Возникает вопрос о свойствах решения при промежуточных значениях λ. Повидимому, оптимальное поведение не изменяется и при всех λ, $0 < \lambda \leq 0,5$. Не обсуждая здесь подробнее этот вопрос в общем случае, приведем результат сравнения процессов типа (1,2) и (2,1) для уравнения (17):

$$x_3(1,2) - x_3(2,1) = \frac{\lambda^5 (1-2\lambda)}{16(1-\lambda)^3 (1-\lambda-\frac{\lambda^2}{2})} .$$

Очевидно, процесс (1,2) всегда – кроме $\lambda = 0$ и $\lambda = 0,5$ – даёт лучший результат, чем процесс (2,1), (в том числе и при малых отрицательных λ; заметим, что при $\lambda \leq -2$ последний процесс нецелесообразен, так как при этом второе приближение хуже исходного, нулевого).

5. Свойства задачи A_1, выраженные в теоремах 1 – 3, могут быть следующим образом использованы для решения задач A_2 и $Б$. Прежде всего очевидно, что оптимальное решение этих задач следует искать среди величин, удовлетворяющих условиям (7). Далее, мы получаем некоторое достаточное условие того, чтобы система чисел $\{p, m, k_i, (i = \overline{1,p})\}$ была решением задачи A_2. Это условие мы выразим в терминах показателя степени при γ в формуле (11) (при $n = p$), зависящего от m и p, причем числа $k_i = z_i - 1$ определяются через m и p по формулам (7). Введём следующие величины

$$\begin{aligned}
&g(m,p) = \prod_{i=1}^{p} \bar{z}_i , \quad \bar{z}_i - 1 = \bar{k}_i = [m + i - 1/p] ; \\
&\Delta_1(m,p) = m - p - [m_1 - (p+1)] = m - m_1 + 1, \\
&m_1 = min\{m / g(m, p+1) \geq g(m,p)\} ; \\
&\Delta_2(m,p) = [m_2 - (p-1)] - (m-p) = m_2 - m + 1 , \\
&m_2 = min\{m / g(m, p-1) \geq g(m,p)\} .
\end{aligned} \qquad (31)$$

Очевидно, что $\Delta_1 \geqslant 0$, $\Delta_2 > 0$. Первое из этих чисел отражает экономию в числе модифицированных шагов, вызванную введением одного дополнительного шага основного процесса; второе число Δ_2 показывает, напротив, сколько дополнительных модифицированных шагов надо сделать для компенсации уменьшения числа основных шагов на единицу. Все эти величины g, Δ_1, Δ_2 могут быть найдены в приводимой таблице I для $3 \leqslant p \leqslant 10$, причём диапазон величины m выбирается так, чтобы при заданном $\bar{p}$ достичь показателя $g(m,\bar{p}) \geqslant 2^{10} = 1024$. При $p = 1$ и $p = 2$ расчёты удобно проводить по следующим простым формулам:

$$g(m,1)=m+1, \quad g(2n,2)=(n+1)^2, \quad g(2n-1,2)=n(n+1),$$

$$m(h)=\min\{m \mid g(m,2) \geqslant h\}=[\sqrt{h}]+\left[\frac{h}{2\sqrt{h}}\right]-2$$

$$(h=4,5,6,\ldots).$$

ЛЕММА 4. Для того, чтобы числа $(m, \bar{p}, \bar{K}_i)$ решали задачу A_2, необходимо выполнение следующих условий:

$$\Delta_1(\bar{m},\bar{p}) \leqslant c/c' \leqslant \Delta_2(\bar{m},\bar{p}) \qquad (\bar{m} > \bar{p} > 1) \; . \tag{32}$$

причём $\bar{K}_i$, $i=\overline{1,\bar{p}}$, определяются формулой (31). При $\bar{m}=\bar{p}$ остаётся лишь последнее, а при $\bar{p}=1$ – первое неравенство.

ДОКАЗАТЕЛЬСТВО. Если оказывается, что $R=c-c'\Delta_1(m,p)<0$, то, как видно из соотношений (31), мы имеем $g(m,p+1) \geqslant g(m,p)$ Поэтому, взяв общее число итераций m_1 вместо m, причём итераций по основному методу будет $p+1$ вместо p, мы получим экономию в затратах в количестве $|R|$ единиц, обеспечивая при этом по крайней мере такую же точность, как и для пары (m, p). Итак, в оптимальном решении $R \geqslant 0$, что равносильно одному из неравенств (32). Аналогично поступаем и со вторым неравенством.

В таблице 2 отражены некоторые результаты решения задачи A_2 при $Q = 25$, $c' = 1$, и для c, меняющегося от 2,5 до

12,5 ; ($\ell = m - p$), в выделенных клетках стоят оптимальные величины.

Таблица 3 показывает, как решается задача Б в следующей видоизмененной постановке: задан – вместо числа Δ в исходной постановке – показатель H, и требуется с минимальными затратами добиться того, чтобы показатель $g(m, p)$, определенный соотношениями (3I), был бы не меньше H. Пользуясь таблицей 3, нетрудно найти, что, например, при $H = 700$, $c' = 1$, $c = 3$, решение видоизмененной задачи Б дает величины

$$\bar{m} = 14, \quad \bar{p} = 5, \quad \bar{\ell} = \bar{m} - \bar{p} = 9, \quad c\bar{p} + c'\bar{\ell} = 29.$$

В таблице 3 $S = g(m,p) - H$ – избыток достигаемого показателя g над заданным H.

6. В общем многомерном случае справедливо следующее утверждение (сравн. с теоремой I).

ТЕОРЕМА 4. Пусть задано функциональное уравнение (I), имеющее решение $x^* \in X$. Пусть оператор P – дважды дифференцируемый в окрестности решения x^*; пусть существует $\Gamma = [P'(x^*)]^{-1}$ и выполнено следующее условие для билинейного оператора $\gamma \equiv \Gamma P'': X \times X \to X$: существует константа $\mu > 0$ такая, что

$$\|\gamma(X_1, X_2)\| \geq \mu \|X_1\| \|X_2\|. \qquad (33)$$

Тогда найдётся такая окрестность $S = S(x^*)$ решения x^*, что для любого начального приближения $x_0 \in S$ задача А имеет решение того же вида, какой указан в теореме 3, с тем отличием, что на этот раз π – не произвольная перестановка индексов: $\pi \in K$, где K – некоторый не пустой класс перестановок (возможно, состоящий из одного элемента), зависящий, вообще говоря, от P и от x^*.

ДОКАЗАТЕЛЬСТВО в существенном аналогично обоснованию теоремы I. Мы используем разложения для $P(x_0)$ и $P'(x_0)$, подобные имеющемуся в формуле (8), а также просто проверяемое тождество

$$[P'(x^*)+K]^{-1}(g+h)=\Gamma g+[P'(x^*)+K]^{-1}(h-K\Gamma g)$$

(аналог правила деления многочлена на многочлен; здесь $K: X\to Y$ – малый по норме линейный оператор). Существование обратного оператора в последней формуле очевидно. Методом математической индукции доказываем, что для приближений $\{x_n\}$ модифицированного процесса Ньютона-Канторовича справедливы разложения:

$$\Delta_n = x_n - x^* = R_n\, \gamma(\Delta_0, \Delta_{n-1}) + o(\|\Delta_0\|^{n+1})$$

$$(\; n = 1,2,\ldots, \qquad R_1 = 0{,}5; \qquad R_n = 1 \text{ для } \quad n>1 \;).$$

(сравн. с формулой (10)). Условие (33) гарантирует выделение главных членов в этих формулах и применимость леммы I. Однако, рассуждения, связанные с условием (13) и с показателем $\mathcal{T}_n$ в формуле (11) здесь, в отличие от доказательства теоремы I, не обоснованы, так как нет перестановочности операторных сомножителей, очевидной для одномерного случая. Поэтому оптимальный процесс может отличаться от указанного в теореме I, основанного на условии (13) процесса, получаясь из него с помощью некоторой перестановки, как в теореме 3.

В связи с вычислительными процессами, изученными в данной статье, отметим, что в книге [4] рассмотрен с другой точки зрения, чем здесь, чередующийся способ типа $\beta = (2,\, 2,\, 2,\ldots)$ и показано, что для приближений $y_n = x_{2n}$ это способ третьего порядка. Подобный факт для функциональных уравнений получен в статье [12а], а в недавней статье тех же авторов [12б] показано (по существу, в других обозначениях), что более общий процесс $(\, p\,,\, p\,,\, p\,,\ldots)$ имеем асимптотический порядок сходимости $1+p$; в этих статьях, однако, даны слишком стеснительные достаточные условия сходимости, сравн. с теоремой I работы [13]. Отметим, что результаты работ [12 - 13] допускают обобщения для класса уравнений, изученных в работе [15], то есть на случай, когда P' удовлетворяет условию Гёльдера. Любопытно, что некоторые результаты, полученные в [15] для этого класса уравнений, вновь получены (отчасти – в более слабой форме) в недавней статье [17], автор которой, видимо, не был знаком с работой [15], а также с работами [16,5].

В перечисленных статьях не рассматривалась задача об оптимальном чередовании процессов.

Добавим, что развитые приемы позволяют решать вопрос и об оптимальном чередовании других методов. Так, для уравнения (I7) последовательное применение сначала одной итерации по методу Ньютона, а затем - Чебышева, приводит к приближению $x_3(1,2)$, а две итерации тех же процессов, но в обратном порядке очередности, к $x_3(2,1)$. Применяя теорему 2, находим, что в первом случае получается лучший результат.

7. Выводы. I. В работе показаны возможности решения проблемы выбора вычислительного алгоритма чередования основного и модифицированного способа Ньютона-Канторовича [I-3], оптимального с точки зрения затрат ресурсов, необходимых для счёта. При этом сравнение основано не на изучении асимптотики вычислительного процесса, как это предложено в книге [4] и более подробно проведено в недавней работе [I0], а на оценке эффекта от конечного числа начальных итераций, что и является особенно интересным с точки зрения вычислительной практики.

2. Результаты работы позволяют уточнить одно замечание А.М. Островского (см. известную и чрезвычайно содержательную книгу [4]) о модифицированном процессе, который в этой книге поставлен под сомнение из-за сравнительно медленной скорости сходимости: дело в том, что замедление сходимости может компенсироваться уменьшением затрат на один шаг процесса. Количественная сторона этого явления и отражена в задачах А и Б.

3. Характеристики оптимальных решений задач А и Б устойчивы в том смысле, что для всех начальных приближений, достаточно близких к искомому корню, оптимальным оказывается один и тот же процесс.

Литература

1. Л.В.Канторович, Функциональный анализ и вычислительная математика, УМН, т.3, вып.6 (1948), стр. 89-185.
2. Л.В.Канторович, Некоторые дальнейшие применения метода Ньютона для функциональных уравнений, Вестник ЛГУ (матем.) № 7, в.2 (1957).
3. Л.В.Канторович, Г.П.Акилов, Функциональный анализ в нормированных пространствах, М., 1959.
4. А.М.Островский, Решение уравнений и систем, ИЛ, М., 1963.

5. М.А.Красносельский, Г.М.Вайникко, П.П.Забрейко, Я.Б.Рутицкий, В.Я.Стеценко, Приближенное решение операторных уравнений. "Наука", М., 1969.

6. Л.Коллатц, Функциональный анализ и вычислительная математика, Изд., "Мир", М., 1969.

7. Р.Беллман, С.Дрейфус, Прикладные задачи динамического программирования, "Наука", М., 1965.

8. H.Ehrmann, Jterations ver fahren... Arch. Rat. Mech. Anal. 4 (1959), 45-88.

9. R.G.Bartle, Newtons Method in Banach Spaces, Proc. Am.Math. Soc, 6, No.5, (1955), 827-831.

10. A.Feldstein, B.Firestone, A study of Ostrowski efficiency for composite iterat. algorithms, Proc. ACM 24-th Nat. Conf., 1969, 147-155.

11. W.C.Rheinboldt, A.Unif. Convergence Theory for a class of Jterdt. Process., SIAM J. Numer. Anal, 5 N 1, 1968, 42-64.

12a. W. Bosarge, Jr, P Falb., A Multipoint Methog, JOTA 4 N 3, 1969, 156-166.

12b. Numer. Math. 14, N 3 1970, 264-287.

13. J.E.Dennis, Jr. On the Kantorovich Hypothesis for Newton's Method, SIAM J. Numer. Anal. 6 N 3, 1969, 493-508.

14. Э.Гурса, Курс математического анализа, т.I.

15а. Б.А.Вертгейм, Об условиях применения метода Ньютона, ДАН 110 № 5 (1956), 719-722.

15б. Успехи Мат. наук, 12, № I (1957), 166-169.

16. М.А.Красносельский, Я.Б.Рутицкий, О некоторых приближенных методах решения нелинейных операторных уравнений, основанных на линеаризации; ДАН 141, № 4 (1961).

17. Keller Newton's Method under Mild Differentiability Conditions, Journ. Comp. Syst. Sci., 4 N 1 (1970) 15-28.

18. Б.А.Вертгейм. Приближенное решение нелинейных уравнений как управляемый процесс, ДАН 194, № I (1970), 16-20.

Поступила в редакцию
6.V. 1970 г.

Таблица I.

3	3	8	0	2	4	4	16	0	2	6	6	64	0	2
	4	12	1	2		5	24	1	2		7	96	1	2
	5	18	1	3		6	36	1	2		8	144	1	2
	6	27	1	4		7	54	1	3		9	216	1	2
	7	36	2	4		8	81	1	4		10	324	1	2
	8	48	2	5		9	108	2	4		11	486	1	3
	9	64	2	6		10	144	2	4		12	729	1	3
	10	80	3	7	4	11	192	2	5	6	13	972	2	3
	11	100	3	8		12	256	2	5		14	1296	2	4
	12	125	3	10		13	320	3	6	7	7	128	0	2
	13	150	3	11		14	400	3	7		8	192	1	2
3	14	180	4	12		15	500	3	7		9	288	1	2
	15	216	4	14		16	625	3	8		10	432	1	2
	16	252	5	15		17	750	4	9		11	648	1	2
	17	294	5	17	4	18	900	4	9	7	12	972	1	2
	18	343	5	19		19	1080	4	10		13	1458	1	3
	19	392	6	20	5	5	32	0	2	8	8	256	0	2
	20	448	6	22		6	48	1	2		9	384	1	2
	21	512	6	24		7	72	1	2		10	576	1	2
	22	576	7	25		8	108	1	2		11	864	1	2
	23	648	7	27	5	9	162	1	3		12	1296	1	2
	24	729	8	29		10	243	1	3	9	9	512	0	2
	25	810	8	31		11	324	2	4		10	768	1	2
	26	900	9	33		12	432	2	4		11	1152	1	2
3	27	1000	9	36		13	576	2	4	10	10	1024	0	2
	28	1100	9	38	5	14	768	2	5					
						15	1024	2	5					

Таблица 2.

2,5	23	I	24	3	I7	4	750	6	I0	3	80
	20	3	448		I5	5	I024	7	I3	2	56
	I7	5	I600		I3	6	972	8	II	2	42
	I6	6	2304	3,5	I5	4	500	9	9	2	30
	I4	7	2I87	4,5	II	4	I92	I0	7	2	20
	II	9	II52	5	I3	3	I50	II	I5	I	I6
	I0	I0	I024	4	I3	4	320	I2	I4	I	I5

Таблица 3.

I6	I5	I	0	I28	8	6	I6	I024	I023	I	0
	6	2	0	I28	7	7	0		62	2	0
	5	3	2	256	255	I	0		28	3	76
I6	4	4	0		30	2	0		I9	4	56
32	3I	I	0		I7	3	38		I5	5	0
	I0	2	4		I2	4	0		I4	6	272
	7	3	4		II	5	72		I3	7	334
	6	4	4		I0	6	72		I2	8	272
32	5	5	0		9	7	32		II	9	I28
64	63	I	0	256	8	8	0	I024	I0	I0	0
	I4	2	0	5I2	5II	I	0	700	699	I	0
	9	3	0		44	2	I7		5I	2	2
	8	4	I7		2I	3	0		24	3	29
	7	5	8		I6	4	II3		I7	4	50
64	6	6	0		I3	5	64		I4	5	68
I28	I27	I	0		I2	6	2I7		I2	6	29
	2I	2	4		II	7	I36		I2	7	272
	I3	3	22		I0	8	64		II	8	I64
	I0	4	I6	5I2	9	9	0		I0	9	68
I28	9	5	34					700	I0	I0	324