УДК 519.65

АППРОКСИМАЦИЯ ДАННЫХ ГИСТОГРАММЫ МЕТОДОМ УСЛОВНОЙ МИНИМИЗАЦИИ ДЛИНЫ КУБИЧЕСКОГО СПЛАЙНА КЛАССА C^1, ОБЛАДАЮЩЕГО СВОЙСТВАМИ НЕОТРИЦАТЕЛЬНОСТИ И ЛОКАЛЬНОЙ МОНОТОННОСТИ. ЧАСТЬ 1

С. В. Мжачих, Н. В. Колобянина, Ю. Н. Лапшина
(ФГУП "РФЯЦ-ВНИИЭФ", г. Саров Нижегородской области)

Проблема, рассматриваемая в настоящей работе, относится к обработке данных, которые доступны в виде функции одной переменной, заданной ступенчатой гистограммой. Построив для таких данных аппроксимацию в виде кубического сплайна класса C^1 с желаемыми свойствами, среди которых могут быть неотрицательность и локальная монотонность, исследователь сможет оценить участки монотонности и значения в заданных точках, причем как для самой функции, так и для ее производной.

В статье представлены основные положения методики расчета аппроксимации. Задача решается с помощью метода минимизации целевой функции, представляющей собой зависимость длины кривой сплайна от векторной величины. Для искомого вектора формируется область, которая задает сплайну нужные свойства. Наиболее точно проблема решается с помощью ограничений в виде нелинейных неравенств, но в некоторых задачах допустимыми могут быть и линейные неравенства. В последнем случае решение находится приближенно, но быстро и практически безаварийно.

Ключевые слова: кубический сплайн, условная минимизация, локально монотонная аппроксимация, неотрицательная аппроксимация, модифицированный метод Лагранжа, метод градиентного спуска, метод линеаризации Ньютона, внутригрупповой спектр.

Введение

Задача, которая рассматривается в настоящей работе, относится к обработке данных, заданных скалярной (одномерной) функцией одного аргумента в виде ступенчатой гистограммы. Такие данные типичны, например, для разностных задач, когда одномерному или многомерному множеству (пространственной ячейке, энергетической группе, шагу по времени и т. п.) ставится в соответствие одно числовое значение, которое является некоторым усредненным значением непрерывной функции на рассматриваемом множестве. Аналитическое представление функции не известно, но иногда для анализа нужна оценка ее вида. Поэтому делается попытка конструирования непрерывной аппроксимации. В некоторых исследованиях интересен вид производной функции. Для таких случаев необходимо строить аппроксимацию класса C^1.

К аппроксимирующей функции могут предъявляться разные требования. Так, в некоторых задачах физическая величина, для которой рассчитывается аппроксимация, не может быть отрицательной по определению (концентрация частиц, энергия и т. п.). Часто требуется наличие у аппроксимирующей функции свойства локальной монотонности (ЛМ), т. е. когда на отрезках данных пользователя функция либо не возрастает, либо не убывает.

Решать задачу расчета аппроксимации данных гистограммы можно разными путями (см., например, работы [1, 2]). В работе [2] схематично описан метод восстановления непрерывных внутригрупповых нейтронных спектров. Рассчитывались энергетические зависимости, используемые

для описания концентрации частиц, при этом сеточная функция усреднялась по различным пространственным областям (зонам). В качестве независимой использовалась переменная летаргии ($\ln(10/E)$, E — энергия, выраженная в МэВ). Метод из работы [2] прошел испытание временем и хорошо себя зарекомендовал. Его идея заключалась в аппроксимации ступенчатой гистограммы непрерывной функцией, заданной параболой в пределах каждой энергетической группы. Для расчета использовался метод условной минимизации. В результате получалась непрерывная, но не дифференцируемая (на границах групп) неотрицательная зависимость со свойством ЛМ.

В настоящей работе достаточно подробно описан метод, который является дальнейшим развитием метода из работы [2]. Несмотря на наличие многих общих черт, новый алгоритм существенно отличается от предыдущего.

Во-первых, теперь в качестве искомой функции используется кубический сплайн класса C^1, который намного точнее аппроксимирует ступенчатые данные с большими перепадами значений.

Во-вторых, используется другая целевая функция (ЦФ), точка минимума которой определяет решение задачи. В новом методе ищется минимум длины кривой кубического сплайна класса C^1. Ранее была использована ЦФ, которая имеет минимальное нулевое значение, если искомая кусочно-параболическая функция относится к классу C^1.

В-третьих, используются другие неравенства, которые задают ограничения на искомый вектор решения. Новые идеи появились после того, как авторы больше узнали о локально комонотонных сплайнах класса C^1 (см., например, работы [3, 4]).

В-четвертых, используется другой решатель системы линейных уравнений. Вместо метода исключения Гаусса применяется метод квадратного корня, разработанный специально для систем с симметричной матрицей (см., например, книгу [5]).

Знания о методе условной минимизации, использованном в настоящей работе, были получены, главным образом, из книги [6].

1. Аппроксимирующая функция в виде кубического сплайна класса C^1

Рассмотрим множество из отрезков $[t_{n-1}, t_n]$, где $n = \overline{1, N}$, $n \in \mathbb{Z}$ ($\mathbb{Z}$ — множество целых чисел). Для каждого n-го отрезка известно положительное значение $\bar{f}_n$. Величины $\bar{f}_n$ образуют вектор[*] $\bar{\mathbf{f}} \equiv \left(\bar{f}_1, \bar{f}_2, \ldots, \bar{f}_N\right)^{\mathrm{T}} \equiv \left(\bar{f}_n\right)_{n=\overline{1,N}}$.

В предлагаемом методе используются абсолютные значения рассчитываемых величин, при этом полагается, что некоторое среднее значение элементов вектора $\bar{\mathbf{f}}$ равно единице. С помощью множителя проведем нормирование исходного вектора $\bar{\mathbf{f}}$ так, чтобы истинным стало равенство

$$\sum_{n=1}^{N} \bar{f}_n \Delta_n = \sum_{n=1}^{N} \Delta_n = t_N - t_0, \tag{1}$$

где $\Delta_n \equiv \Delta t_n \equiv t_n - t_{n-1}$.

Будем также полагать, что для всех значений индекса n истинны неравенства

$$\bar{f}_n > \varphi_0, \tag{2}$$

где φ_0 — неотрицательный параметр. Если для данных, которые требуют обработки, некоторые неравенства (2) оказываются ложными, то задачу надо переформулировать, удалив из рассмотрения "плохие" отрезки. Возможно, что при этом задачу придется разбить на несколько подзадач, для которых равенства (1) и неравенства (2) истинны. Действия подсказаны практикой расчетов. Дело в том, что для "плохих" отрезков не всегда удается получать решение с приемлемой точностью, при этом сам процесс расчета может быть слишком дорогостоящим.

Ступенчатую гистограмму $\bar{\mathbf{f}}$ будем аппроксимировать сплайном $F(t)$ класса $C^1[t_0, t_N]$, заданным на отрезке $[t_{n-1}, t_n]$ в виде полинома третьей степени

$$F_n(t) \equiv b_n (t - t_{n-1})^3 + c_n (t - t_{n-1})^2 + d_{n-1} (t - t_{n-1}) + f_{n-1}, \quad t \in [t_{n-1}, t_n], \tag{3}$$

[*] Символ тождества иногда будет использоваться для задания обозначения (записи) или равенства по определению.

откуда $F(t_i) = f_i$, $F'(t_i) = d_i$, где $i = \overline{0, N-1}$. Полагаем $F(t_N) = f_N$, $F'(t_N) = d_N$.

Из условий непрерывности функций $F(t)$ и $F'(t)$ получаем коэффициенты (см. [3])

$$c_n = \frac{3}{\Delta_n}\left(\delta_n - \frac{2d_{n-1} + d_n}{3}\right); \quad b_n = \frac{2}{\Delta_n^2}\left(\frac{d_n + d_{n-1}}{2} - \delta_n\right), \tag{4}$$

где

$$\delta_n \equiv \frac{\Delta f_n}{\Delta_n}, \quad \Delta f_n \equiv f_n - f_{n-1}, \quad n = \overline{1, N}.$$

Рассмотрим некоторые свойства кубического полинома $F_n(t)$.

Пусть $b_n \neq 0$. Полином может иметь два действительных локальных экстремума:

$$F_n'\left(t_n^{\pm}\right) = 0, \quad t_n^{\pm} = \widetilde{t}_n \pm \frac{\sqrt{p_n}}{3b_n}, \quad \widetilde{t}_n = t_{n-1} - \frac{c_n}{3b_n} = \frac{1}{2}\left(t_n^{+} + t_n^{-}\right), \quad p_n = c_n^2 - 3b_n d_{n-1}. \tag{5}$$

Если $p_n < 0$, то уравнение $F_n'(t) = 0$ не имеет действительных корней и функция $F_n(t)$ на всей действительной оси (на множестве $\mathbb{R}$) ведет себя строго монотонно. Если $p_n > 0$, то $F_n''\left(t_n^{\pm}\right) = \pm 2\sqrt{p_n}$. Следовательно, точка t_n^{+} всегда является точкой минимума, а точка t_n^{-} — точкой максимума. Если $p_n = 0$, то точки экстремумов совпадают, функция при этом монотонна. В случае $b_n \neq 0$ такая точка называется точкой перегиба.

Пусть $b_n = 0$. Имеем $p_n = c_n^2 \geq 0$. Если $c_n < 0$, то парабола имеет точку максимума, если $c_n > 0$ — точку минимума. В случае $c_n = 0 \quad (p_n = 0)$ функция $F_n(t)$ линейна.

Итак, для точки перегиба и точки локального минимума, включая случай $b_n = 0 \ \left(\sqrt{p_n} = c_n\right)$, справедливы эквивалентные формулы

$$t_n^{+} = t_{n-1} + \frac{-c_n + \sqrt{p_n}}{3b_n} = t_{n-1} + \frac{-d_{n-1}}{c_n + \sqrt{p_n}}.$$

Для расчета аппроксимирующей функции $F(t)$ привлечем N равенств

$$\int_{t_{n-1}}^{t_n} F_n(t)\, dt = \bar{f}_n \Delta_n, \qquad n = \overline{1, N}. \tag{6}$$

Учитывая (4), из равенств (6) можно получить рекуррентное соотношение

$$d_n - d_{n-1} - \frac{6}{\Delta_n}\left(f_n + f_{n-1} - 2\bar{f}_n\right) = 0, \qquad n = \overline{1, N}. \tag{7}$$

Зададим для сплайна в крайней правой точке граничное условие (ГУ) в виде $F''(t_N) = 0$, которое принято называть естественным. Другие ГУ возможны, но в этой работе не рассматриваются. Поскольку $F''(t_N) = 6b_N\Delta_N + 2c_N$, то, учитывая (4), получаем соотношение

$$d_N = \frac{1}{2}\left(-d_{N-1} + 3\delta_N\right). \tag{8}$$

Из соотношения (8) и соотношения (7), записанного для $n = N$, следует формула

$$d_N = \frac{1}{\Delta_N}\left(3f_N + f_{N-1} - 4\bar{f}_N\right). \tag{9}$$

Аналогичное задание ГУ в крайней левой точке приводит к формуле

$$d_0 = \frac{1}{\Delta_1}\left(-3f_0 - f_1 + 4\bar{f}_1\right). \tag{10}$$

Из формул (9) и (10) следует выбрать одну. Если вектор $\mathbf{f} \equiv (f_n)_{n=\overline{0,N}}$ считать искомым, то с помощью соотношений (7) можно рассчитать вектор $\mathbf{d} \equiv (d_n)_{n=\overline{0,N}}$, а затем по формулам (4) определить векторы $\mathbf{c} \equiv (c_n)_{n=\overline{1,N}}$ и $\mathbf{b} \equiv (b_n)_{n=\overline{1,N}}$. Переменные b_n, c_n, d_n, p_n задают значения соответствующих функций вектора $\mathbf{f}$: $b_n = B_n(\mathbf{f})$, $c_n = C_n(\mathbf{f})$, $d_n = D_n(\mathbf{f})$, $p_n = P_n(\mathbf{f})$.

Далее будем использовать обозначения $\nabla Y \equiv \left(\dfrac{\partial Y}{\partial f_i}\right)_{i=\overline{0,N}}$, $\nabla^2 Y \equiv \left(\dfrac{\partial^2 Y}{\partial f_j \partial f_i}\right)_{i,j=\overline{0,N}}$ соответственно для вектора градиента и симметричной матрицы Гессе для произвольной функции $Y(\mathbf{f})$ класса C^1. Заметим, что величины ∇D_n, ∇B_n, ∇C_n не зависят от вектора $\mathbf{f}$, поэтому они определяются на стадии расчета начальных данных (РНД) задачи.

2. Постановка задачи расчета точки условного минимума для функции нескольких переменных

Будем решать задачу конструирования кубического сплайна $F(t)$ (см. (3)), который должен:
1) принадлежать классу $\mathrm{C}^1[t_0, t_N]$;
2) сохранять известные интегралы гистограммы (см. (6));
3) обладать свойством ЛМ;
4) быть неотрицательным.

Проблема построения аппроксимирующей функции с заданными свойствами относится к задаче восстановления данных. Для такой задачи, как правило, единственного решения не существует, а заключение о пригодности решения, полученного на основании выбранной математической модели расчета, для конкретного приложения во многом является субъективным.

Для решения проблемы воспользуемся методом условной минимизации. Для этого, во-первых, нужно определить зависящую от вектора $\mathbf{f}$ скалярную ЦФ, для которой требуется найти минимальное значение. И, во-вторых, нужно определить семейство равенств и неравенств, задающих область, в пределах которой вектор решения $\mathbf{f}$ формирует сплайн с требуемыми свойствами.

За свойство 1 (из указанных четырех) отвечают формулы (4), за свойство 2 — рекуррентное соотношение (7). Метод представляет собой итерационную процедуру. Равенства-ограничения (7) учтем так, чтобы для каждой итерации решения $\mathbf{f}_k$ они выполнялись точно. Для этого навяжем сплайну ГУ в одной из граничных точек отрезка $[t_0, t_N]$. Можно задавать любое ГУ, которое сводится к известной функции $D_0(\mathbf{f})$ или $D_N(\mathbf{f})$. В работе рассматривается условие $F''(t_N) = 0$ или $F''(t_0) = 0$ (т. е. используется (9) или (10)).

Наличие у аппроксимирующей функции свойства 2 для некоторых приложений может быть критически важным. Так, например, если искомая функция задает энергетическую зависимость концентрации частиц (спектр), то свойство 2 позволяет аппроксимации точно воспроизводить известные числа частиц в энергетических группах.

Для задания свойств 3 и 4 привлекаются неравенства. Под точным будем понимать решение сформулированной выше задачи с заданным ГУ при использовании ЦФ и нелинейных неравенств-ограничений, приводимых далее. В такой постановке минимум ЦФ ищется на самом широком множестве значений $\mathbf{f}$, для которого свойства 3 и 4 гарантируются. Для решения проблемы, обозначенной в начале данного раздела, можно использовать другие варианты метода, которые получаются при задании ограничений с помощью линейных неравенств. Возможно, пользователь предпочтет один из них, так как у них есть свои преимущества. Следует заметить, что искомый сплайн $F(t)$ зависит от некоторых параметров метода решения.

2.1. Целевая функция. Будем искать аппроксимирующий сплайн $F(t)$, соответствующий минимуму длины его кривой, т. е. будем решать задачу вида

$$\sum_{n=1}^{N} \int_{t_{n-1}}^{t_n} \sqrt{1 + S_n^2(t)}\,dt \to \min, \qquad \text{где} \quad S_n(t) \equiv F_n'(t)\,. \tag{11}$$

Интеграл, используемый в (11), не может быть выражен через элементарные функции, поэтому длину кривой будем рассчитывать приближенно. Применим метод численного интегрирования, например, метод Гаусса. Пусть M_G — число квадратур на каждом отрезке $[t_{n-1}, t_n]$. Зададим ЦФ в виде

$$\Gamma(\mathbf{f}) \equiv \sum_{n=1}^{N} \Gamma_n(\mathbf{f}), \quad \Gamma_n(\mathbf{f}) \equiv \sum_{m=1}^{M_G} \Delta_{n,m} \sqrt{E_{n,m}(\mathbf{f})} \approx \int_{t_{n-1}}^{t_n} \sqrt{1 + S_n^2(t)}\, dt, \quad E_{n,m}(\mathbf{f}) \equiv 1 + S_{n,m}^2(\mathbf{f}).$$

Здесь

$$\Delta_{n,m} = \frac{1}{2}\omega_m \Delta_n; \quad S_{n,m}(\mathbf{f}) \equiv F_n'(t_{n,m}) = \beta_{2,n,m} b_n + \beta_{1,n,m} c_n + d_{n-1};$$

$$t_{n,m} = t_{n-1} + \tau_{n,m}, \quad \tau_{n,m} = \frac{1}{2}(\mu_m + 1)\Delta_n; \quad \beta_{k,n,m} = (k+1)\tau_{n,m}^k, \quad k = \overline{1,2};$$

$$\omega_m,\ \mu_m \ - \text{веса и узлы квадратуры}, \quad \sum_{m=1}^{M_G} \omega_m = 2;$$

$$\frac{\partial \Gamma}{\partial f_i} = \sum_{n=1}^{N} \sum_{m=1}^{M_G} \frac{\Delta_{n,m} S_{n,m}(\mathbf{f})}{E_{n,m}^{1/2}} \frac{\partial S_{n,m}(\mathbf{f})}{\partial f_i}; \qquad \frac{\partial^2 \Gamma}{\partial f_j \partial f_i} = \sum_{n=1}^{N} \sum_{m=1}^{M_G} \frac{\Delta_{n,m}}{E_{n,m}^{3/2}} \frac{\partial S_{n,m}(\mathbf{f})}{\partial f_i} \frac{\partial S_{n,m}(\mathbf{f})}{\partial f_j}; \tag{12}$$

$$\nabla S_{n,m} = \beta_{2,n,m} \nabla B_n + \beta_{1,n,m} \nabla C_n + \nabla D_{n-1}.$$

Значения $\beta_{1,n,m}$, $\beta_{2,n,m}$ и $\nabla S_{n,m}$ определяются на стадии РНД задачи. Матрица $\nabla^2 \Gamma$ полностью заполнена, ее элементы на главной диагонали положительны.

Если для рекуррентного соотношения используется значение d_0, то $D_0(\mathbf{f}) \equiv D_0(f_0, f_1)$, а для $n \in [1, N]$ зависимость $D_n(\mathbf{f})$ превращается в зависимость вида $D_n(f_0, f_1, \ldots, f_n)$. Следовательно, если $i \notin [0, n]$, то $\dfrac{\partial D_n}{\partial f_i} = \dfrac{\partial B_n}{\partial f_i} = \dfrac{\partial C_n}{\partial f_i} = 0$. Дальнейший анализ приводит к выводу, что в первой формуле (12) суммирование ведется от $n = \max(1, i)$, а во второй — от $n = \max(1, i, j)$.

Если для рекуррентного соотношения используется значение d_N, то можно показать, что $\dfrac{\partial D_{n-1}}{\partial f_i} = \dfrac{\partial B_n}{\partial f_i} = \dfrac{\partial C_n}{\partial f_i} = 0$, если $i \notin [n-1, N]$. Поэтому в первой формуле (12) задаем суммирование до $n = \min(N, i+1)$, а во второй — до $n = \min(N, i+1, j+1)$.

На практике было исследовано несколько вариантов задания ЦФ. Выбор авторов обусловлен полезным свойством данной ЦФ, которое было предсказано, а затем подтвердилось в процессе тестирования методики. Было замечено, что расчеты с рассматриваемой ЦФ без использования ограничений, относящихся к ЛМ, приводили к наиболее сглаженным аппроксимациям, при этом эффективно подавлялись численные (физически не обоснованные) осцилляции. Это свойство продемонстрировано на расчетах задач безусловной минимизации. Отметим, что выбранная ЦФ не определяет свойство ЛМ сплайна.

2.2. Ограничения, относящиеся к ЛМ сплайна, в виде нелинейных неравенств. Пусть значения гистограммы на отрезках $[t_{n-2}, t_{n-1}]$, $[t_{n-1}, t_n]$ и $[t_n, t_{n+1}]$ монотонно возрастают, при этом выполняется двойное неравенство

$$\rho \bar{f}_{n-1} < \bar{f}_n < \frac{1}{\rho} \bar{f}_{n+1}, \tag{13}$$

где $\rho = \dfrac{2+\varsigma}{2-\varsigma}$ — параметр, зависящий от неотрицательного параметра ς. Заметим, что $\rho \geq 1$.

Условие (13) формируется из неравенств

$$\bar{f}_n - \bar{f}_{n-1} > \frac{\varsigma}{2}\left(\bar{f}_{n-1} + \bar{f}_n\right) \geq 0; \quad \bar{f}_{n+1} - \bar{f}_n > \frac{\varsigma}{2}\left(\bar{f}_n + \bar{f}_{n+1}\right) \geq 0.$$

Если условие (13) выполнено, то потребуем от функции $F_n(t)$ неубывания на отрезке $[t_{n-1}, t_n]$, т. е. $F_n'(t) \geq 0$, $\forall t \in [t_{n-1}, t_n]$.

Условие монотонности сплайна $F(t)$ на отрезке $[t_{n-1}, t_n]$ в случае $\delta_n \neq 0$ задается системой трех неравенств (см. [3])

$$x_n + y_n - 3 - \sqrt{x_n y_n} \leq 0; \tag{14}$$

$$x_n \geq 0; \quad y_n \geq 0, \tag{15}$$

где

$$x_n \equiv \frac{d_{n-1}}{\delta_n}; \quad y_n \equiv \frac{d_n}{\delta_n}.$$

Вместо неравенства (14), функция которого не дифференцируется при $x_n = 0$ и $y_n = 0$, лучше использовать эквивалентное (для области (15)) неравенство

$$\Psi(x_n, y_n) \leq 0, \tag{16}$$

где

$$\Psi(x,y) \equiv \begin{cases} -xy, & \text{если} \quad x + y - 3 \leq 0; \\ (x+y-3)^2 - xy & \text{в других случаях,} \end{cases} \qquad \Psi \in \mathrm{C}^1(\mathbb{R} \times \mathbb{R}).$$

Запишем неравенство (16) в виде

$$r_n = R_n(\mathbf{f}) \leq 0, \tag{17}$$

При условии (13) полагаем $R_n(\mathbf{f}) \equiv R_n^-(\mathbf{f})$, где

$$R_n^{\pm}(\mathbf{f}) \equiv \begin{cases} 0, & \text{если } \pm\delta_n > 0 \text{ или } (\pm d_{n-1} > 0 \text{ и } \pm d_n > 0); \\ -\Delta_n^2 d_{n-1} d_n, & \text{если } \pm(d_{n-1} + d_n - 3\delta_n) \geq 0; \\ \Delta_n^2 \left[(d_{n-1} + d_n - 3\delta_n)^2 - d_{n-1} d_n \right] & \text{в других случаях.} \end{cases}$$

Пусть значения гистограммы на трех соседних отрезках монотонно убывают, при этом выполняется двойное неравенство

$$\rho \bar{f}_{n+1} < \bar{f}_n < \frac{1}{\rho} \bar{f}_{n-1}, \tag{18}$$

которое формируется из неравенств

$$\bar{f}_{n-1} - \bar{f}_n > \frac{\varsigma}{2}\left(\bar{f}_{n-1} + \bar{f}_n\right) \geq 0; \quad \bar{f}_n - \bar{f}_{n+1} > \frac{\varsigma}{2}\left(\bar{f}_n + \bar{f}_{n+1}\right) \geq 0.$$

Если условие (18) выполнено, то потребуем от функции $F_n(t)$ невозрастания на отрезке $[t_{n-1}, t_n]$, т. е. $F_n'(t) \leq 0$, $\forall t \in [t_{n-1}, t_n]$.

В этом случае используем неравенство (17), полагая $R_n(\mathbf{f}) \equiv R_n^+(\mathbf{f})$.

Сформулируем в общем виде ограничения, которые определяют свойство ЛМ. Для отрезка $[t_{n-1}, t_n]$ при истинности одного из условий (13) или (18), которые не могут быть истинными одновременно, зададим ограничение в виде нелинейного неравенства (17), а также линейных неравенств в узлах сплайна

$$g_j = G_j(\mathbf{f}) \leq 0, \tag{19}$$

где $j = \overline{n-1,\, n}$;

$$G_j(\mathbf{f}) \equiv G_j^{\pm}(\mathbf{f}); \quad R_n(\mathbf{f}) \equiv R_n^{\pm}(\mathbf{f}); \tag{20}$$

$$G_j^{\pm}(\mathbf{f}) \equiv \pm\frac{1}{2}(\Delta_j + \Delta_{j+1}) D_j(\mathbf{f}).$$

В выражении (20), а также в подобных выражениях далее используется знак "минус", если истинно неравенство (13), и знак "плюс" при условии (18).

Неравенства (19) относятся к границам отрезков (узлам сплайна), поэтому следует исключить двойное задание одного неравенства для соседних отрезков.

Укажем на некоторые особенности автоматического (программного) задания ограничений, связанных с ЛМ сплайна. Как показала практика расчетов, после того как будет определено множество *монотонных* отрезков по условиям (13) и (18), для некоторых из них ограничения лучше отменить. Прежде всего это касается отрезков, непосредственно соседствующих с отрезками, для которых условия монотонности не выполняются. Замечено, что уменьшение числа монотонных отрезков повышает эффективность метода.

Можно показать, что множество комонотонности (см. (14), (15)) выпукло и находится внутри множества, образуемого касательными линиями к его внешней границе. Так, для решения задачи истинны неравенства

$$x_n + y_n - 6 \le 0; \quad x_n - 4 \le 0; \quad y_n - 4 \le 0,$$

которые представим в виде

$$\widetilde{r}_n = \widetilde{R}_n(\mathbf{f}) \le 0; \quad \overleftarrow{r}_n = \overleftarrow{R}_n(\mathbf{f}) \le 0; \quad \vec{r}_n = \vec{R}_n(\mathbf{f}) \le 0, \tag{21}$$

где

$$\widetilde{R}_n(\mathbf{f}) \equiv \widetilde{R}_n^{\pm}(\mathbf{f}); \quad \overleftarrow{R}_n(\mathbf{f}) \equiv \overleftarrow{R}_n^{\pm}(\mathbf{f}); \quad \vec{R}_n(\mathbf{f}) \equiv \vec{R}_n^{\pm}(\mathbf{f});$$
$$\widetilde{R}_n^{\pm}(\mathbf{f}) \equiv \pm\Delta_n(6\delta_n - d_{n-1} - d_n); \quad \overleftarrow{R}_n^{\pm}(\mathbf{f}) \equiv \pm\Delta_n(4\delta_n - d_{n-1}); \quad \vec{R}_n^{\pm}(\mathbf{f}) \equiv \pm\Delta_n(4\delta_n - d_n).$$

Легко удостовериться, что из неравенств (19) и (21) следует, что $\pm\delta_n \le 0$, т. е. в случае $\delta_n \ne 0$ неравенства (15) истинны. Заметим, что если $\delta_n = 0$, то из тех же неравенств получаем $d_{n-1} = d_n = 0$, т. е. функция $F_n(t)$ на основании (4) монотонна: $F_n(t) \equiv f_{n-1} = f_n$.

Решаем задачу, привлекая только линейные неравенства (19) и (21), затем проверяем нелинейные неравенства (17) и, если они окажутся истинными, считаем задачу решенной. В противном случае продолжаем вычисления, добавив ограничения (17).

2.3. Ограничения, относящиеся к ЛМ сплайна, в виде линейных неравенств. Использование линейных ограничений характеризуется высокой скоростью расчета задачи, так как все производные от функций, задающих неравенства, рассчитываются на стадии РНД. Более того, практика показала, что замена нелинейных ограничений на линейные повышает надежность метода решения задачи условной минимизации.

Считаем истинным неравенство либо (13), либо (18). Предложим способ задания ограничений, относящихся к монотонному поведению функции $F_n(t)$ на множестве $[t_{n-1}, t_n]$, но уже с помощью линейных неравенств. Используем квадратное подмножество комонотонности, впервые представленное в работе [4]. Для того чтобы сплайн $F(t)$ на множестве $[t_{n-1}, t_n]$ был монотонен, достаточно истинности неравенств (15), а также неравенств

$$\Psi_1(x_n, y_n) \equiv y_n - 3 \le 0; \qquad \Psi_2(x_n, y_n) \equiv x_n - 3 \le 0. \tag{22}$$

Переходя к функциям от вектора $\mathbf{f}$, получим систему, включающую неравенства (19) для $j = n-1$ и $j = n$, а также неравенства

$$r_{n,i} = R_{n,i}(\mathbf{f}) \le 0, \quad i = \overline{1, 2I}, \tag{23}$$

где $I = 1$;

$$R_{n,i}(\mathbf{f}) \equiv R_{n,i}^{\pm}(\mathbf{f}); \quad R_{n,i}^{\pm}(\mathbf{f}) \equiv \mp\Delta_n\delta_n\Psi_i(x_n, y_n).$$

Подмножество комонотонности можно расширить, если воспользоваться линейной системой, включающей неравенства (15) и неравенства вида

$$\Psi_i(x_n, y_n) \le 0, \qquad i = \overline{1, 2I}, \tag{24}$$

где $I \geq 1$; $\Psi_{I+j}(x,y) \equiv \Psi_j(y,x)$, $j = \overline{1,I}$.

Случай $I = 1$ представлен формулами (22). Для случая $I \geq 2$ линейные функции $\Psi_i(x,y)$ определим так, чтобы площадь подмножества комонотонности в виде N-угольника, где $N = 2(I+1)$, была максимальной. Вершины многоугольника, симметричного относительно прямой $y = x$, принадлежат границе области комонотонности. В множество вершин всегда входят вершины квадрата Фрича—Карлсона (случай $I = 1$). Функция внешней границы области комонотонности для $x \in [0,3]$ имеет вид

$$\Phi(x) \equiv \frac{1}{2}\left[6 - x + \sqrt{3x(4-x)}\right].$$

Для $I = 2$ задача сводится к поиску точки (x^*, y^*) на кривой $y = \Phi(x)$ в области $x \in (0,3)$, для которой площадь треугольника, образуемого точками (0,3), (x^*, y^*) и (3,3), максимальна. Задача легко решается, получаем точку (1,4). Для $I = 3$ задача сводится к поиску двух вершин на кривой $y = \Phi(x)$ в области $x \in (0,3)$, для которых максимальна площадь четырехугольника, образуемого точками (0,3), (x_1^*, y_1^*), (x_2^*, y_2^*) и (3,3). Решение находится из кубического уравнения $y_1^* = y_2^*$ при условии $x_2^* = x_1^*(4 - x_1^*)$. Получаем $x_1^* = 4\sin^2\dfrac{\pi}{9}$. Итак, площадь подмножества максимальна при задании

$$I = 2 \quad \text{и} \quad \Psi_1(x,y) \equiv y - x - 3, \quad \Psi_2(x,y) \equiv 2y + x - 9;$$
$$I = 3 \quad \text{и} \quad \Psi_1(x,y) \equiv y - \frac{y_1^* - 3}{x_1^*}x - 3, \quad \Psi_2(x,y) \equiv y - y_1^*, \quad \Psi_3(x,y) \equiv y - \frac{3 - y_1^*}{3 - x_2^*}(x - x_2^*) - y_1^*.$$

Площадь подмножества в случае $I = 1$ составляет примерно 68 % от площади всего множества комонотонности. Значение возрастает до 91 %, если $I = 2$, и до 96 %, если $I = 3$.

Множество линейных ограничений является подмножеством области нелинейных ограничений, поэтому линейные ограничения решают проблему ЛМ сплайна точно. Но не следует забывать, что задача минимизации длины кривой сплайна при этом решается приближенно, так как при использовании нелинейных ограничений можно получить сплайн с ЛМ с меньшим значением длины кривой. Однако, как правило, с точки зрения пользователя методики, это не является принципиальным (см. второй абзац разд. 2).

2.4. Ограничения, относящиеся к неотрицательности сплайна, в виде нелинейных неравенств. Прежде всего зададим простейшие константные ограничения в узлах сплайна

$$f_n \geq 0, \quad n = \overline{0,N}. \tag{25}$$

Потребуем от функции $F_n(t)$ неотрицательности в точке локального минимума t_n^+ (см. (5)), если $t_n^+ \in (t_{n-1}, t_n)$. Вместо переменной t_n^+ будем использовать переменную t_n^*, рассчитываемую по формуле $t_n^* = t_{n-1} + \tau_n^*$, где

$$\tau_n^* = T_n(\mathbf{f}) \equiv \begin{cases} \dfrac{-d_{n-1}}{z_n^+}, & \text{если} \quad c_n > 0, \quad d_{n-1} < 0, \quad z_n^+ \Delta_n > -d_{n-1}; \\ \dfrac{z_n^-}{3b_n}, & \text{если} \quad c_n \leq 0, \quad b_n > 0, \quad z_n^- < 3b_n\Delta_n; \end{cases} \tag{26}$$
$$z_n^\pm \equiv \Lambda(p_n) \pm c_n,$$
$$\Lambda(p) \equiv \begin{cases} p^{1/2}, & \text{если} \quad p \geq \varphi_1; \\ 0, & \text{если} \quad p \leq 0; \\ \dfrac{1}{2}\varphi_1^{-3/2} p(3\varphi_1 - p), & \text{если} \quad p \in (0, \varphi_1), \end{cases} \qquad \Lambda \in \mathrm{C}^1(0, +\infty), \quad \Lambda \in \mathrm{C}(\mathbb{R}).$$

Здесь φ_1 — малый положительный параметр. Значение t_n^* определяется только в случае $\tau_n^* \in (0, \Delta_n)$: условия расчета приведены в (26). Если $p_n \geq \varphi_1$ или $p_n = 0$, то справедливо точное равенство $t_n^* = t_n^+$. Если $p_n \leq 0$, что возможно только в случае $b_n \neq 0$, то $t_n^* = \widetilde{t}_n$.

Для формулирования ограничения, относящегося к неотрицательности функции $F_n(t)$ внутри интервала (t_{n-1}, t_n), используем переменную $f_n^* \equiv F_n(t_n^*)$ (см. (3)).

Предлагается еще один вариант расчета значения f_n^*. Из разложения Тейлора для полинома третьей степени $F_n(t)$ в окрестности точки локального минимума $t_n^* = t_n^+$ имеем

$$F_n(t) = f_n^* + (t - t_n^*)^2 \left[c_n + (t + 2t_n^* - 3t_{n-1})\, b_n\right],$$

следовательно,

$$f_n = f_n^* + (\Delta_n - \tau_n^*)^2 \left[c_n + (2\tau_n^* + \Delta_n)\, b_n\right], \quad f_{n-1} = f_n^* + (\tau_n^*)^2 \left(c_n + 2\tau_n^* b_n\right); \tag{27}$$

$$\int_{t_{n-1}}^{t_n} F_n(t)\, dt = \bar{f}_n \Delta_n = f_n^* \Delta_n + \frac{c_n}{3}\left[(\Delta_n - \tau_n^*)^3 + (\tau_n^*)^3\right] + \frac{\Delta_n b_n}{4}\left[\Delta_n^3 - 6\Delta_n (\tau_n^*)^2 + 8(\tau_n^*)^3\right]. \tag{28}$$

Используя (27), исключаем в (28) слагаемое со множителем c_n и получаем соотношение

$$f_n^* = \frac{3}{2}\bar{f}_n - \frac{1}{2}(1 - \gamma_n^*) f_n - \frac{1}{2}\gamma_n^* f_{n-1} + \Delta_n^3 U(\gamma_n^*)\, b_n,$$

где

$$\gamma_n^* = \frac{\tau_n^*}{\Delta_n}, \quad \gamma_n^* \in [0,1]; \quad U(\gamma) \equiv \frac{1}{2}\left(\frac{1}{2} - \gamma\right)\left[\left(\frac{1}{2} - \gamma\right)^2 + \frac{1}{4}\right].$$

На принадлежность точки t_n^* интервалу (t_{n-1}, t_n) укажет область положительных значений финитной функции класса $\mathrm{C}^1(\mathbb{R})$

$$W_n(t) \equiv \begin{cases} 1, & \text{если} \quad t \in [t_{n-1} + \theta_n, t_n - \theta_n]; \\ x_{1,n}^2 (3 - 2x_{1,n}), & \text{если} \quad t \in (t_{n-1}, t_{n-1} + \theta_n); \\ x_{2,n}^2 (3 - 2x_{2,n}), & \text{если} \quad t \in (t_n - \theta_n, t_n); \\ 0, & \text{в других случаях} \ \ (t \notin (t_{n-1}, t_n)), \end{cases}$$

где

$$x_{1,n}(t) = \frac{t - t_{n-1}}{\theta_n}, \qquad x_{2,n}(t) = \frac{t_n - t}{\theta_n}, \qquad x_{1,n}, x_{2,n} \in (0,1);$$
$$\theta_n = \lambda \Delta_n, \qquad \lambda \ - \text{параметр}: \ \lambda \in \left(0, \frac{1}{2}\right].$$

Для тех отрезков $[t_{n-1}, t_n]$, которые ранее не были задействованы в формировании ограничений, относящихся к ЛМ сплайна, зададим ограничения для искомого вектора $\mathbf{f}$ в виде неравенств

$$q_n = Q_n(\mathbf{f}) \equiv -f_n^* w_n^* \le 0, \quad n \in [1, N], \tag{29}$$

где $w_n^* = W_n(t_n^*)$. Если величина t_n^* на интервале (t_{n-1}, t_n) не рассчитывалась, то значения функции $Q_n(\mathbf{f})$ и ее производных равны нулю.

Проведем анализ неравенства (29).

Если $t_n^* \notin (t_{n-1}, t_n)$, то $w_n^* = 0$ и неравенство (29) является истинным. В рассматриваемом случае сплайн не имеет точки локального минимума на интервале и его неотрицательность на интервале гарантируется неравенствами (25).

Если $p_n < 0$, то искомая функция $F_n(t)$ строго монотонна на множестве $\mathbb{R}$. В случае $p_n = 0$ функция $F_n(t)$ либо имеет точку перегиба ($b_n \ne 0$), либо она линейна ($b_n = c_n = 0$). Следовательно, в случае $p_n \le 0$ функция монотонна и неравенства (25) гарантируют неотрицательность сплайна на отрезке $[t_{n-1}, t_n]$. При этом неравенство (29) справедливо, даже если $w_n^* > 0$. Случай $p_n \in (0, \varphi_1)$ на практике не отличим от случая $p_n \le 0$.

Пусть t_n^* является точкой локального минимума функции $F_n(t)$ и $t_n^* \in (t_{n-1}, t_n)$, т. е. $p_n \geq \varphi_1$ и $w_n^* > 0$. Следовательно, неравенство (29) является истинным при условии $f_n^* \geq 0$. Если при этом точка локального максимума t_n^- функции $F_n(t)$ не принадлежит множеству (t_{n-1}, t_n), то неотрицательность функции на отрезке обеспечивается только неравенством (29). Если же $t_n^- \in (t_{n-1}, t_n)$, то минимальное значение функции на отрезке определяется как минимум значений f_n, f_{n-1} и f_n^*. Неотрицательность значений функции в граничных точках отрезка обеспечивается введением ограничений вида (25).

2.5. Ограничения, относящиеся к неотрицательности сплайна, в виде линейных неравенств. В некоторых задачах требование неотрицательности сплайна $F_n(t)$ на отрезке $[t_{n-1}, t_n]$ можно ослабить, разрешив полиному третьей степени $F_n(t)$ иметь отрицательные значения на некотором открытом подмножестве отрезка, мера которого задается коэффициентом величины Δ_n. В таких случаях разумно использовать ограничения в виде линейных неравенств. О достоинствах постановки задачи с линейными ограничениями было сказано ранее.

Потребуем для решения выполнения неравенств (25).

На отрезке $[t_{n-1}, t_n]$ определим сетку из узлов $t_{n,m}$:

$$t_{n,m} = t_{n-1} + \frac{\Delta_n}{M} m, \quad m = \overline{1, M-1}, \quad M \geq 2.$$

Используем ограничения в виде системы линейных неравенств

$$q_{n,m} = Q_{n,m}(\mathbf{f}) \equiv -F_n(t_{n,m}) \leq 0, \quad n \in [1, N]. \tag{30}$$

Условия (30) являются необходимыми условиями неотрицательности функции $F_n(t)$ на интервале (t_{n-1}, t_n). Однако эти условия не являются достаточными. Заметим, что искомый полином $F_n(t)$ может быть отрицателен только на открытом подмножестве отрезка $[t_{n-1}, t_n]$, содержащем единственную точку локального минимума t_n^*. Размер такого интервала не превышает значения Δ_n / M. С увеличением числа M дефект решения ослабляется.

2.6. Ограничения, относящиеся к крайним точкам и отрезкам сплайна, в виде неравенств. В некоторых задачах требуется, чтобы сплайн не убывал (не возрастал) в точке t_0. Аналогичное условие может потребоваться и для точки t_N. При необходимости вводим неравенство $g_0 \leq 0$ и/или неравенство $g_N \leq 0$. Используем выражение (20), в котором знаки диктуются задачей. Определим $\Delta_0 = \Delta_1$ и $\Delta_{N+1} = \Delta_N$.

В подразд. 2.2 были использованы неравенства для задания ЛМ сплайна на отрезках $[t_{n-1}, t_n]$ для $n \in [2, N-1]$. В некоторых задачах от сплайна требуется монотонность в крайних интервалах. Для этого зададим фиктивные значения $\bar{f}_0 = \bar{f}_{N+1} = 0$ и используем их при формировании неравенств для первого ($n = 1$) и последнего ($n = N$) отрезков $[t_{n-1}, t_n]$.

3. Модифицированный метод Лагранжа расчета точки минимума функции нескольких переменных при наличии ограничений в виде неравенств

Задачу поиска точки минимума ЦФ $\Gamma(\mathbf{f})$ при наличии S ограничений-неравенств запишем в виде

$$\Gamma(\mathbf{f}) \to \min; \quad \mathbf{\Omega}(\mathbf{f}) \leq 0, \tag{31}$$

где $\mathbf{\Omega}(\mathbf{f}) \equiv \Big(\Omega_s(\mathbf{f})\Big)_{s=\overline{1,S}}$, $\mathbf{\Omega} \in \mathbb{R}^S$ ($\Omega_s \in \mathbb{R}$, $\forall s$).

Для решения задачи условной минимизации (31) будем использовать модифицированный метод Лагранжа [6]. Задача сводится к последовательности решения подзадач нахождения итераций $\mathbf{f}_k$ решения $\mathbf{f}$.

3.1. Модифицированная функция Лагранжа. Определим для итерационного индекса k ($k = \overline{1,K}$) модифицированную функцию Лагранжа

$$\mathcal{L}_k(\mathbf{f}) \equiv \Gamma(\mathbf{f}) + \Big(\boldsymbol{\eta}_k, \boldsymbol{\Omega}^+(\mathbf{f}, \boldsymbol{\eta}_k, \chi_k)\Big) + \frac{\chi_k}{2}\left|\boldsymbol{\Omega}^+(\mathbf{f}, \boldsymbol{\eta}_k, \chi_k)\right|^2.$$

Здесь используются следующие переменные и обозначения:

$\boldsymbol{\eta}_k$ — вектор множителей Лагранжа: $\boldsymbol{\eta}_k \equiv (\eta_{k,s})_{s=\overline{1,S}}$, $\eta_{k,s} \in \mathbb{R}$, $\eta_{k,s} \geq 0$;

χ_k — параметр квадратичного штрафа: $\chi_k \in \mathbb{R}$, $\chi_k > 0$;

$$\boldsymbol{\Omega}^+(\mathbf{f}, \boldsymbol{\eta}_k, \chi_k) \equiv \Big(\Omega_s^+(\mathbf{f}, \boldsymbol{\eta}_k, \chi_k)\Big)_{s=\overline{1,S}}, \quad \boldsymbol{\Omega}^+ \in \mathbb{R}^S; \qquad \Omega_s^+(\mathbf{f}, \boldsymbol{\eta}_k, \chi_k) \equiv \max\left\{\Omega_s(\mathbf{f}), -\frac{\eta_{k,s}}{\chi_k}\right\};$$

$$(\mathbf{x}, \mathbf{y}) = \sum_{s=1}^{S} x_s y_s, \quad \text{если} \quad \mathbf{x} \equiv (x_s)_{s=\overline{1,S}}, \quad \mathbf{y} \equiv (y_s)_{s=\overline{1,S}}; \quad |\mathbf{x}| \equiv \sqrt{(\mathbf{x}, \mathbf{x})}.$$

Определим множество ограничений, выполняющихся на интервалах (t_{n-1}, t_n), $n = \overline{1,N}$:

$$\boldsymbol{\Omega}(\mathbf{f}) \equiv \Big(\boldsymbol{\Omega}_0(\mathbf{f}), \boldsymbol{\Omega}_1(\mathbf{f}), \ldots, \boldsymbol{\Omega}_6(\mathbf{f})\Big)^{\mathrm{T}}, \quad \boldsymbol{\Omega}_i(\mathbf{f}) \equiv \Big(\Omega_{i,n}(\mathbf{f})\Big)_{n\in\boldsymbol{\omega}_i}, \quad i = \overline{0,6};$$

$$\Omega_{0,n}(\mathbf{f}) \equiv -f_n; \quad \boldsymbol{\Omega}_1(\mathbf{f}) \equiv \mathbf{Q}(\mathbf{f}); \quad \boldsymbol{\Omega}_2(\mathbf{f}) \equiv \mathbf{G}(\mathbf{f}); \quad \boldsymbol{\Omega}_3(\mathbf{f}) \equiv \widetilde{\mathbf{R}}(\mathbf{f});$$

$$\boldsymbol{\Omega}_4(\mathbf{f}) \equiv \overleftarrow{\mathbf{R}}; \quad \boldsymbol{\Omega}_5(\mathbf{f}) \equiv \vec{\mathbf{R}}(\mathbf{f}); \quad \boldsymbol{\Omega}_6(\mathbf{f}) \equiv \mathbf{R}(\mathbf{f});$$

$$\boldsymbol{\omega}_i \equiv (\omega_{i,s})_{s=\overline{1,S_i}} \equiv \{n : n \in [0,N], \text{ задано ограничение } \Omega_{i,n}(\mathbf{f}) \leq 0\}.$$

Целочисленные множества-векторы $\boldsymbol{\omega}_i$ формируются на стадии РНД задачи. Заметим, что

$$\boldsymbol{\omega}_0 \equiv \{\overline{0,N}\}; \quad \boldsymbol{\omega}_3 = \boldsymbol{\omega}_4 = \boldsymbol{\omega}_5 = \boldsymbol{\omega}_6; \quad \sum_{i=0}^{6} S_i = S.$$

Функцию Лагранжа можно преобразовать к виду

$$\mathcal{L}_k(\mathbf{f}) = \Gamma(\mathbf{f}) + \frac{1}{2\chi_k}\sum_i \sum_{n\in\boldsymbol{\omega}_i} \big(\max^2\{0, \eta_{i,k,n} + \chi_k \Omega_{i,n}(\mathbf{f})\} - \eta_{i,k,n}^2\big),$$

где

$$\boldsymbol{\eta}_k \equiv \big(\boldsymbol{\eta}_{0,k}, \boldsymbol{\eta}_{1,k}, \ldots, \boldsymbol{\eta}_{6,k}\big)^{\mathrm{T}}; \qquad \boldsymbol{\eta}_{i,k} \equiv (\eta_{i,k,n})_{n\in\boldsymbol{\omega}_i}.$$

Если использованы линейные ограничения вида (30), то вместо слагаемого для $i = 1$ либо дополнительно следует задать слагаемое

$$\frac{1}{2\chi_k}\sum_{n\in\boldsymbol{\omega}_7}\sum_{m=1}^{M-1} \big(\max^2\{0, \eta_{7,k,n,m} + \chi_k Q_{n,m}(\mathbf{f})\} - \eta_{7,k,n,m}^2\big),$$

где $\boldsymbol{\omega}_7 = \boldsymbol{\omega}_1$.

Если использованы линейные ограничения вида (23), то вместо слагаемых для $i = \overline{3,6}$ следует задать слагаемое

$$\frac{1}{2\chi_k}\sum_{n\in\boldsymbol{\omega}_8}\sum_{m=1}^{2I} \big(\max^2\{0, \eta_{8,k,n,m} + \chi_k R_{n,m}(\mathbf{f})\} - \eta_{8,k,n,m}^2\big),$$

где $\boldsymbol{\omega}_8 = \boldsymbol{\omega}_3$.

Для каждого индекса k будем решать задачу безусловной минимизации модифицированной функции Лагранжа. Запишем задачу в виде

$$\mathcal{L}_k(\mathbf{f}) \to \min, \quad \mathbf{f} \in \mathbb{R}^{N+1}. \tag{32}$$

Задача (32) сводится к нахождению действительного вектора $\mathbf{f}_k$, который является решением системы нелинейных уравнений

$$\nabla\mathcal{L}_k\left(\mathbf{f}_k\right)=0. \tag{33}$$

Задача (33) решается при фиксированных значениях $\boldsymbol{\eta}_k$ и χ_k, которые пересчитываются в конце каждого итерационного шага (см. далее подразд. 3.6).

3.2. Метод Ньютона. Эффективным методом решения задачи (33) является метод линеаризации Ньютона, который порождает итерационный цикл по индексу l ($l=\overline{1,L}$). Пусть значение индекса $l=L+1$ отвечает условию нормального завершения полного цикла. В цикле решаются системы линейных уравнений

$$H_{k,l}\,\delta\mathbf{f}_{k,l}=-\nabla\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right), \tag{34}$$

где $\delta\mathbf{f}_{k,l}$ — шаг смещения вектора $\mathbf{f}_{k,l-1}$; $\mathbf{f}_{k,0}=\mathbf{f}_{k-1}$ (см. (33)). Новая итерация решения рассчитывается согласно формуле

$$\mathbf{f}_{k,l}=\mathbf{f}_{k,l-1}+\alpha_{k,l}\,\delta\mathbf{f}_{k,l}. \tag{35}$$

Для расчета значения $\alpha_{k,l}$, где $\alpha_{k,l}\in(0,1]$, применяется специальный алгоритм.

Используем в качестве матрицы $H_{k,l}$ матрицу Гессе, т. е. $H_{k,l}=\nabla^2\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right)$. Ниже приводятся формулы расчета вектора $\nabla\mathcal{L}_k$ и матрицы $\nabla^2\mathcal{L}_k$ в точке $\mathbf{f}_{k,l-1}$:

$$\frac{\partial\mathcal{L}_k}{\partial f_i}=\frac{\partial\Gamma}{\partial f_i}+\sum_{s=1}^{S}\begin{cases}0, & \text{если } \eta_{k,s}+\chi_k\Omega_s\le 0;\\ \left(\eta_{k,s}+\chi_k\Omega_s\right)\dfrac{\partial\Omega_s}{\partial f_i} & \text{в других случаях;}\end{cases}$$

$$\frac{\partial^2\mathcal{L}_k}{\partial f_j\partial f_i}=\frac{\partial^2\Gamma}{\partial f_j\partial f_i}+\sum_{s=1}^{S}\begin{cases}0, & \text{если } \eta_{k,s}+\chi_k\Omega_s\le 0;\\ \chi_k\dfrac{\partial\Omega_s}{\partial f_j}\dfrac{\partial\Omega_s}{\partial f_i}+\left(\eta_{k,s}+\chi_k\Omega_s\right)\dfrac{\partial^2\Omega_s}{\partial f_j\partial f_i} & \text{в других случаях.}\end{cases}$$

Как видим, матрица $\nabla^2\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right)$ полностью заполнена. Для решения систем уравнений с симметричной матрицей используется метод квадратного корня [5].

3.3. Критерий завершения расчета задачи безусловной минимизации. Полагаем, что задача (32) для текущего номера итерации k успешно решена, если для текущего индекса l, где $l\in[1,L]$, выполнены все следующие условия:

1) последняя система уравнений (34) была решена для $H_{k,l}=\nabla^2\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right)$ (в подразд. 3.4 будет показано, что в некоторых случаях для решения может привлекаться метод градиентного спуска (МГС) с диагональной матрицей $H_{k,l}$);
2) истинны неравенства

$$\begin{aligned}&\left(\delta\mathbf{f}_{k,l},\nabla\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right)\right)\le 0;\\ &\mathcal{L}_k\left(\mathbf{f}_{k,l}\right)\le\left(1+\varepsilon_3\right)\mathcal{L}_k\left(\mathbf{f}_{k,l-1}\right);\\ &\left|\Gamma_n\left(\mathbf{f}_{k,l-1}\right)-\Gamma_n\left(\mathbf{f}_{k,l}\right)\right|\le\begin{cases}\varepsilon_4\Gamma_n\left(\mathbf{f}_{k,l}\right), & \text{если } \bar{f}_n>\varphi_5;\\ \varepsilon_5\Gamma_n\left(\mathbf{f}_{k,l}\right) & \text{в противном случае;}\end{cases}\qquad n=\overline{1,N},\end{aligned} \tag{36}$$

где ε_3, ε_4, ε_5, φ_5 — неотрицательные параметры. Здесь вектор $\mathbf{f}_{k,l}$ рассчитан по формуле (35) при $\alpha_{k,l}=1$.

Если задача (32) для текущего номера итерации k была решена, то следует принудительно выйти из цикла по индексу l.

Из неравенства (36) следует, что существует окрестность точки $\mathbf{f}_{k,l-1}$, в которой функция $\mathcal{L}_k\left(\mathbf{f}\right)$ не возрастает в направлении вектора $\delta\mathbf{f}_{k,l}$. Несложно показать, что для положительно определенной матрицы $H_{k,l}^{-1}$ неравенство (36) всегда истинно.

Если неравенство (36) ложно, то пересчитаем шаг $\delta\mathbf{f}_{k,l}$, используя МГС (см. подразд. 3.4). В противном случае определим значение $\alpha_{k,l}$ с помощью специального итерационного алгоритма, который предполагается описать в части 2 данной статьи.

Использование шагового множителя $\alpha_{k,l}$ обеспечивает понижение значения функции $\mathcal{L}_k(\mathbf{f})$ на каждом l-м цикле. Вычислительный опыт показывает, что, начиная с некоторого значения l, в качестве множителя $\alpha_{k,l}$ принимается его стандартное единичное значение. Однако для первых итераций иногда требуется значительно уменьшить шаг $\delta\mathbf{f}_{k,l}$.

3.4. Метод градиентного спуска. МГС не относится к эффективным алгоритмам, поэтому задействуем его только в следующих случаях:

– метод Ньютона не может должным образом понизить функцию Лагранжа, т. е. либо неравенство (36) является ложным, либо не удается определить значение $\alpha_{k,l}$;
– при решении системы (34) с матрицей Гессе произошел отказ метода квадратного корня (на главной диагонали был получен близкий к нулю элемент).

Стандартно используем МГС с масштабированием. Для этого модифицируем матрицу Гессе, обнулив все ее элементы вне главной диагонали.

Матрица, обратная к диагональной матрице $H_{k,l}$, является диагональной. При этом если у исходной матрицы все элементы положительны, то и у обратной матрицы диагональ также положительна. Очевидно, такая матрица $H_{k,l}^{-1}$ является положительно определенной.

Уточним: МГС с масштабированием будем задействовать только в случае положительности всех диагональных элементов матрицы $H_{k,l}$. В противном случае привлечем к расчету МГС в варианте метода наискорейшего спуска. Для этого используем диагональную единичную матрицу, т. е. полагаем $H_{k,l} = H_{k,l}^{-1} = E$.

С учетом положительной определенности матрицы $H_{k,l}^{-1}$ для МГС истинность неравенства (36) следует априори. После расчета вектора $\delta\mathbf{f}_{k,l}$ из системы (34) запускаем алгоритм расчета коэффициента $\alpha_{k,l}$, затем используем формулу (35). Если произошел отказ алгоритма расчета коэффициента $\alpha_{k,l}$, то, скорее всего, задача не решается с требуемой точностью.

По окончании расчета с применением МГС следует вернуться к расчету по методу Ньютона на следующем шаге цикла по l.

3.5. Критерий завершения расчета задачи условной минимизации. После принудительного выхода из цикла по индексу l ($l \le L$) или по его завершении ($l > L$) определим $\mathbf{f}_k = \mathbf{f}_{k,\min\{l,L\}}$ и $L_k = l$. Задача (31) для текущего номера итерации k, где $k \in [2, K]$, считается успешно решенной, если выполнены все следующие условия:

1) последняя подзадача (32) была успешно решена (см. подразд. 3.3);
2) истинны неравенства

$$|\Gamma_n(\mathbf{f}_{k-1}) - \Gamma_n(\mathbf{f}_k)| \le \varepsilon_5 \Gamma_n(\mathbf{f}_k), \quad n = \overline{1,N},$$

где ε_5 — параметр, введенный в подразд. 3.3;
3) для каждого индекса n из множества $\boldsymbol{\omega}_2$ истинно неравенство

$$g_n \le \Delta_1(\varepsilon_2 |\delta_1| + \varphi_2), \quad \text{если} \quad n = 0; \qquad g_n \le \Delta_N(\varepsilon_2 |\delta_N| + \varphi_2), \quad \text{если} \quad n = N;$$
$$g_n \le \frac{1}{2}(\Delta_n + \Delta_{n+1})(\varepsilon_2 \min\{|\delta_n|, |\delta_{n+1}|\} + \varphi_2), \quad \text{если} \quad n \in [1, N-1],$$

где ε_2, φ_2 — неотрицательные параметры, величины δ_n определены в разд. 1;
4) если от сплайна требуется свойство неотрицательности, то должны быть выполнены все следующие условия:

– истинны неравенства

$$-f_0 \le \varepsilon_1 \bar{f}_1 + \varphi_4; \quad -f_N \le \varepsilon_1 \bar{f}_N + \varphi_4; \quad -f_n \le \varepsilon_1 \min\left\{\bar{f}_n, \bar{f}_{n+1}\right\} + \varphi_4, \quad n = \overline{1, N-1},$$

где ε_1, φ_4 — неотрицательные параметры;

– для каждого индекса n из множества $\boldsymbol{\omega}_1$ либо выполняется неравенство

$$q_n \le \varepsilon_2 \bar{f}_n + \varphi_2, \tag{37}$$

либо истинны неравенства

$$q_{n,m} \le \varepsilon_2 \bar{f}_n + \varphi_2 \quad \forall m. \tag{38}$$

Неравенство (37) проверяется, если проблема неотрицательности решается точно с помощью нелинейных неравенств (29). Если проблема решается приближенно с помощью линейных неравенств (30), то проверяются неравенства (38);

5) если от сплайна требуется свойство ЛМ, то для каждого индекса n из множества $\boldsymbol{\omega}_3$ должно быть выполнено соответствующее задаче условие из следующих:

– если проблема монотонности решается точно с привлечением нелинейного неравенства (17), то истинны неравенства

$$\pm\delta_n \le \varphi_3; \qquad r_n \le \Delta_n r_n^{\max},$$

где $r_n^{\max} = \Delta_n \left(\varepsilon_2 \left|\delta_n\right| + \varphi_2\right)$; φ_3 — неотрицательный параметр;

– если проблема монотонности решается точно с привлечением линейных неравенств (23), то истинны неравенства

$$\pm\delta_n \le \varphi_3; \qquad r_{n,i} \le r_n^{\max}, \quad i = \overline{1, 2I};$$

– если проблема монотонности решается приближенно с привлечением линейных неравенств (21), то истинны неравенства

$$\widetilde{r}_n \le r_n^{\max}, \quad \overleftarrow{r}_n \le r_n^{\max}, \quad \vec{r}_n \le r_n^{\max}.$$

3.6. Расчет векторов Лагранжа и параметра штрафа. Если критерий решения задачи (31) не выполнен и $k < K$, то прежде чем перейти к расчету новой итерации $\mathbf{f}_{k+1}$, рассчитываем вектор Лагранжа $\boldsymbol{\eta}_{k+1}$ и параметр штрафа χ_{k+1}.

Последовательность векторов Лагранжа будем вычислять по рекуррентной формуле

$$\eta_{k+1,s} = \max\left\{0, \eta_{k,s} + \chi_k \Omega_s\left(\mathbf{f}_k\right)\right\}, \quad s = \overline{1, S}, \quad 1 \le k < K; \quad \boldsymbol{\eta}_1 = 0.$$

Положительная последовательность штрафа χ_k может выбираться как заранее, так и в ходе вычислений, исходя из анализа решения. На практике хорошо зарекомендовала себя следующая схема (см. [6]). Начальное значение χ_1 выбирается относительно небольшим, например $\chi_1 = 1/32$. В ходе итерационного процесса обеспечивается монотонное возрастание значений последовательности χ_k, например, соотношением $\chi_{k+1} = \theta_{k+1}\chi_k$, где $\theta_{k+1} \in [2, 8]$. Стандартное значение $\theta_{k+1} = 4$.

Однако следует учитывать, что с ростом параметра штрафа ухудшается обусловленность задачи, и это влечет за собой уменьшение эффективности применяемых методов. В частности, для решения плохо обусловленных задач заведомо не применим МГС. Даже использование метода Ньютона может быть сопряжено с серьезными трудностями, если значение χ_k очень велико, а начальная точка недостаточно близка к решению. По этой причине увеличивать штрафной параметр следует относительно медленно.

В работе [6] предложена еще одна схема расчета параметра штрафа, которая состоит в том, чтобы пересчитывать значение штрафа только в тех случаях, когда не обеспечивается линейная скорость убывания модуля вектора $\boldsymbol{\Omega}_k^+ \equiv \boldsymbol{\Omega}^+\left(\mathbf{f}_k, \boldsymbol{\eta}_k, \chi_k\right)$.

Заключение

В настоящей статье достаточно подробно представлена методика, положения которой тщательно выверялись на практике на большом множестве тестов. Например, варьировались формы записи неравенств, критерии обрыва итерационных циклов, подбирались значения параметров. Однако некоторые важные вопросы остались в статье не освещенными. Не был рассмотрен алгоритм расчета шагового множителя $\alpha_{k,l}$. Также не была исследована проблема расчета начального приближения (вектора $\mathbf{f}_0$). Вопрос значим, так как метод Ньютона при всех его достоинствах имеет существенный недостаток: он относится к методам локальной итерационной сходимости и при неудачном выборе стартовой итерации может отказать.

Упомянутые выше вопросы теории будут рассмотрены в следующей публикации. Там же авторы намерены обсудить результаты численного тестирования методики, а также представить оптимальную тактику счета и выводы о качестве получаемого решения.

Список литературы

1. *Горелов В. П., Фарафонтов Г. Г.* Оценка энергетического спектра нейтронов в слоях цилиндрической многослойной ячейки Вигнера—Зейца // Вопросы атомной науки и техники. Сер. Математическое моделирование физических процессов. 1993. Вып. 1. С. 19—23.
Gorelov V. P., Farafontov G. G. Otsenka energeticheskogo spectra neytronov v sloyakh tsilindricheskoy mnogosloynoy yacheyki Vignera—Zeytsa // Voprosy atomnoy nauki i tekhniki. Ser. Matematicheskoe modlirovanie fizicheskikh protsessov. 1993. Vyp. 1. S. 19—23.

2. *Гребенников А. Н., Фарафонтов Г. Г., Алексеев А. В., Мжачих С. В., Крутько Н. А.* Технология подготовки групповых макроскопических констант и методика их уточнения в процессе расчета задач переноса нейтронов // Там же. 2005. Вып. 4. С. 15—24.
Grebennikov A. N., Farafontov G. G., Alekseev A. V., Mzhachikh S. V., Krutko N. A. Tekhnologiya podgotovki gruppovykh makroskopicheskikh constant i metodika ikh utochneniya v protsesse raschyeta zadach perenosa neytronov // Tam zhe. 2005. Vyp. 4. S. 15—24.

3. *Мжачих С. В., Лапшина Ю. Н.* Об одном локально комонотонном кубическом сплайне класса C^1 // Там же. 2021. Вып. 2. С. 56—69.
Mzhachikh S. V., Lapshina Yu. N. Ob odnom localno komonotonnom kubicheskom splayne klassa C^1 // Tam zhe. 2021. Vyp. 2. S. 56—69.

4. *Fritsch F. N., Carlson R. E.* Monotone piecewise cubic interpolation // SIAM J. Numer. Anal. 1980. Vol. 17, No 2. P. 238—246.

5. *Фадеев Д. К., Фадеева В. Н.* Вычислительные методы линейной алгебры. С.-Пб.: Лань, 2009.
Fadeev D. K., Fadeeva V. N. Vychislitelnye metody lineynoy algebry. S.-Pb.: Lan, 2009.

6. *Бертсекас Д.* Условная оптимизация и методы множителей Лагранжа. М.: Радио и связь, 1987.
Bertsekas D. Uslovnaya optimizatsiya i metody mnozhiteley Lagranzha. M.: Radio i svyaz, 1987.

Статья поступила в редакцию 30.05.22.