ОБЪЕДИНЕННЫЙ ИНСТИТУТ ЯДЕРНЫХ ИССЛЕДОВАНИЙ

Лаборатория теоретической физики

С.Н. Соколов, И.Н. Силин

Д-810

НАХОЖДЕНИЕ МИНИМУМОВ ФУНКЦИОНАЛОВ МЕТОДОМ ЛИНЕАРИЗАЦИИ

Дубна 1961 год

С.Н. Соколов, И.Н. Силин

Д-810

НАХОЖДЕНИЕ МИНИМУМОВ ФУНКЦИОНАЛОВ МЕТОДОМ ЛИНЕАРИЗАЦИИ

Аннотация

Предлагается метод минимизации функционалов, зависящих от искомых параметров только через свой функциональный аргумент, дающий значительно лучшую сходимость, чем применяемые общие методы. Рассмотрена устойчивость находимых решений по отношению к внешним, не входящим в функциональный аргумент, параметрам. Метод показал высокую эффективность при решении ряда задач.

Сейчас для минимизации любых выражений, зависящих от параметров, применяется почти исключительно метод градиента (скорейшего спуска)[1,2]. Естественным следствием универсальности этого метода является, прежде всего, неудовлетворительная скорость сходимости, причем необходимое для достижения минимума число шагов быстро растет с числом варьируемых параметров. Этот же недостаток присущ и другим общим методам, например, релаксационному[1].

Ввиду того, что подавляющее большинство выражений, с минимизацией которых приходится иметь дело на практике, довольно шаблонны по своей структуре, представляется целесообразным выделить из них два-три основных типа и разработать соответствующие специализированные методы поиска минимумов, более эффективные, чем метод градиента. Такие специализированные методы могут быть полезны в чистом виде или в комбинации с некоторой поощрительной кибернетической процедурой, так же, как, например, метод градиента используется в методе оврагов[3].

В настоящей работе в качестве одного из таких специализированых методов предлагается метод нахождения минимумов функционалов $M\{y(a,x)\}$, зависящих от искомых параметров $a=\{a_1,\dots,a_m\}$ только через свой функциональный аргумент $y(a,x)$, т.е.

$$\frac{\partial M}{\partial a_k}=\int\frac{\delta M}{\delta y(a,x)}\,\frac{\partial y(a,x)}{\partial a_k}\,dx \qquad (1)$$

(под вариационной производной $\frac{\delta M}{\delta y(x)}$ мы понимаем ядро производной Фреше[12] $M'(y,f)\equiv\int\frac{\delta M}{\delta y_x}f\,dx$). Например, $M=\int f[y(a,x)]\,dx$. Переменная x может быть дискретной[x)] и непрерывной, одномерной и многомерной. Функционал M мы будем считать дважды непрерывно-дифференцируемым по своему функциональному аргументу во всех областях, где это может понадобиться в дальнейшем.

С формальной точки зрения предлагаемый метод состоит в замене точных уравнений экстремума некоторой системой линейных уравнений, в связи с чем он назван методом линеаризации.

---

x) Здесь и в дальнейшем подразумевается, что если функционал задан на дискретном множестве точек x_ξ, $\xi=1,\dots,n$, то вариационные производные заменяются частными, а интегрирование – суммированием, например: $\frac{\partial M}{\partial a_k}=\sum_{\xi=1}^{n}\frac{\partial M}{\partial y(a,x_\xi)}\,\frac{\partial y(a,x_\xi)}{\partial a_k}$. Число точек x_ξ должно быть не меньше числа параметров a.

Из достоинств метода линеаризации следует упомянуть, что, как будет видно ниже, скорость достижения минимума почти не падает с ростом числа параметров a .

Метод линеаризации применялся в ряде случаев [4,5,6]. Минимум находился за 5 - 10 приближений при числе параметров от 2-х до 16-ти. В Объединенном институте ядерных исследований имеется стандартная программа метода линеаризации в применении к методу наименьших квадратов.

Функционал M может иметь много минимумов разных типов, причем далеко не все из них обязательно соответствуют решениям поставленной задачи. Для отбрасывания ложных решений необходимо исследовать устойчивость положений минимумов по отношению к смещению внешних (не входящих в функциональный аргумент $y(a,x)$) параметров. Отметим, что такое исследование не сводится к изучению формы найденных ямок. Для метода линеаризации оказывается легко установить соответствие между типом минимума и тем, какой характер имеет процесс поиска в его окрестности.

§ 1. Формула шага

В методе линеаризации функционал $M\{y\}$ аппроксимируется квадратичным

$$M\{y\} \simeq \tfrac{1}{2} \int \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} [y(a,z) - y(a^\circ, z)][y(a,x) - y(a^\circ,x)]\, dz dx + \int \frac{\delta M}{\delta y(x)} [y(a,x) - y(a^\circ,x)]dx + \text{const}, \tag{1.1}$$

а зависимость $y(a)$ линейной [x]

$$y(a) \simeq y(a^\circ, x) + \sum_{k=1}^{m} \frac{\partial y}{\partial a_k} \Delta a_k . \tag{1.2}$$

Функциональная окрестность начального приближения, в которой справедлива аппроксимация (1.1), будет предполагаться настолько большой, что минимум M находится в этой окрестности [xx]. Относительно (1.2) такого предположения

---

x) В применении к методу наименьших квадратов на целесообразность аппроксимации (1.2) указывалось различными авторами, например, [7,8].

xx) Под функциональной окрестностью минимума M понимается область, в которой $\int f(z) \frac{\delta^2 M}{\delta y(z)\, \delta y(x)} f(x)\, dz dx > 0$ для произвольной ненулевой функции f .

не делается. Для оценки направления и расстояния до минимума M получается система линейных уравнений

$$\frac{\partial M}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \int \frac{\partial y(z)}{\partial a_l} \frac{\delta^2 M}{\delta y(z)\ \delta y(x)} \frac{\partial y(x)}{\partial a_k} dz\, dx = 0 . \qquad (1.3)$$

Аппроксимации (1.1), (1.2) приводят к отбрасыванию в выражении для шага не только членов высшего порядка малости по Δa, но и части членов первого порядка малости, как это будет видно ниже. Поэтому такая процедура нуждается в некоторых пояснениях.

Найдем вектор $\Delta a = \{\Delta a_1, \ldots, \Delta a_m\}$, который в бесконечно малой окрестности минимума M указывал бы точно в минимум. Разлагая производные $\frac{\partial M}{\partial a_k}$ в ряд по степеням вектора Δa

$$\frac{\partial M(a + \Delta a)}{\partial a_k} = \frac{\partial M(a)}{\partial a_k} + \sum_{l=1}^{m} \frac{\partial^2 M(a)}{\partial a_l\ \partial a_k} \Delta a_l + \ldots \qquad (1.4)$$

и используя условие экстремума

$$\frac{\partial M(a + \Delta a)}{\partial a_k} = 0, \qquad (1.5)$$

получим, пренебрегая высшим степеням Δa_i, линейную систему уравнений

$$\frac{\partial M(a)}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \frac{\partial^2 M(a)}{\partial a_l\ \partial a_k} = 0, \qquad k = 1, \ldots, m, \qquad (1.6)$$

или, вычисляя явно вторые производные,

$$\frac{\partial M}{\partial a_k} + \sum_{l=1}^{m} \Delta a_l \int \frac{\partial y(z)}{\partial a_l} \frac{\delta^2 M}{\delta y(z)\ \delta y(x)} \frac{\partial y(x)}{\partial a_k} dz\, dx +$$

$$+ \sum_{l=1}^{m} \Delta a_l \int \frac{\delta M}{\delta y(x)} \frac{\partial^2 y(x)}{\partial a_l\ \partial a_k} dx = 0 . \qquad (1.7)$$

Вектор Δa , компоненты которого удовлетворяют системе (1.7), в малой окрестности экстремума дает точное направление и расстояние до экстремума.

Вдали от минимума вектор Δa указывает грубо направление на ближайший экстремум без различия его типа, так что, если начальное приближение a^0 оказалось, например, вблизи седловой точки, то движение вдоль Δa приведет в седловую точку. Это делает затруднительным использование уравнения (1.7) для поисков минимумов, так как минимизируемые функции вблизи начального приближения могут иметь много разных конкурирующих экстремумов.

Сравнивая (1.3) и (1.7), видим, что уравнения (1.3) получены из (1.7) при помощи своеобразной линеаризации - отбрасывания члена

$$\sum_{l=1}^{m} \Delta a_l \int \frac{\delta M}{\delta y(x)} \quad \frac{\partial^2 y(x)}{\partial a_l \; \partial a_k} \quad dx = \sum_{l=1}^{m} \Delta a_l \quad Q_{lk} , \tag{1.8}$$

который учитывает нелинейность $y(a)$ и не является, вообще говоря, малым по сравнению с оставленным в (1.3) членом

$$\sum_{l=1}^{m} \Delta a_l \int \frac{\partial y(z)}{\partial a_l} \quad \frac{\delta^2 M}{\delta y(z) \; \delta y(x)} \quad \frac{\partial y(x)}{\partial a_k} \; dz \; dx = \sum_{l=1}^{m} \Delta a_l \quad G_{lk} . \tag{1.9}$$

Выбрасывание Q_{ik} приводит к нескольким преимуществам системы (1.3) перед системой (1.7). В частности, вектор Δa , определяемый системой (1.3), всегда указывает в сторону уменьшения M и прекращается конкуренция экстремумов разных типов.

Практически используемая в методе линеаризации система уравнений отличается от (1.3) тем, что вместо полного шага Δa берется некоторая его доля $\overline{\Delta a} = \lambda \Delta a$, $0 < \lambda \leq 1$, которая определяется из условия оптимальной сходимости процесса минимизации (см. § 3). Качественно λ можно оценить как максимальное $\lambda \leq 1$, при котором линейная аппроксимация

$$y(a + \lambda \Delta a) - y(a) \simeq \overline{\Delta y} = \sum_{k=1}^{m} \lambda \Delta a_k \frac{\partial y}{\partial a_k} \tag{1.10}$$

является еще грубо справедливой.

Подставляя $\overline{\Delta a} = \lambda \Delta a$ и (1.9) в (1.3), получаем

$$\lambda \frac{\partial M}{\partial a_k} + \sum_{l=1}^{m} \overline{\overline{A}}\, \overline{\overline{a_l}}\, G_{lk} = 0, \qquad (1.11)$$

откуда находим шаг в пространстве параметров

$$\overline{\overline{\Delta a_l}} = -\lambda \sum_{k=1}^{m} G^{-1}_{lk} \frac{\partial M}{\partial a_k} \qquad (1.12)$$

и функциональный шаг

$$\overline{\overline{\Delta y(z)}} = -\lambda \sum_{i,k=1}^{m} \frac{\partial y(z)}{\partial a_i} G^{-1}_{ik} \frac{\partial M}{\partial a_k} . \qquad (1.13)$$

§ 2. Простой пример

Пусть $y(A)$ – аналитическая функция параметра $A = a_1 + ia_2$ и $M\{y\} = |y|^2$. Тогда минимумы M равны нулю и соответствуют корням уравнения $y(a) = 0$. Подставляя $M = |y|^2$ в (1.9), (1.12) и полагая $\lambda = 1$, получаем

$$G = 2|y'_A| \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \Delta A = A^{(n+1)} - A^{(n)} = -y/y'_A . \qquad (2.1)$$

Очевидно, мы пришли к методу Ньютона для отыскания комплексных корней уравнения[x] $y(A) = 0$. Следует подчеркнуть, что метод линеаризации, вообще говоря, не эквивалентен методу Ньютона. В частности, если искать решение уравнений экстремума $\frac{\partial M}{\partial a_k} = 0$ методом Ньютона, мы получим не систему (1.3), а систему (1.7), недостатки которой мы уже обсуждали.

---

x) Относительно формулы (2.1) см. также работу [9]. Заметим, что в случае неаналитических функций $y(A)$ матрица G перестает быть диагональной и формулы (2.1) усложняются.

§ 3. Некоторые свойства формулы шага и выбор λ

Функциональный шаг $\overline{\Delta y(z)}$ инвариантен к любой (в том числе и нелинейной) замене параметров

$$a_k \to b_k(a_1, \dots, a_m). \tag{3.1}$$

Выражение (1.13) для $\overline{\Delta y(z)}$ можно интерпретировать как некоторый градиент, а именно, как градиент M в пространстве $\frac{\partial y}{\partial a}$ с метрикой $\frac{\delta^2 M}{\delta y(z)\delta y(x)}$ (т.е. со скалярным произведением $(f_1, f_2) = \int f_1(z)\ \frac{\delta^2 M}{\delta y(z)\ \delta y(x)}\ f_2(x)\ dz\ dx$). Действительно, для любого $dy = \sum_{p=1}^{m} \frac{\partial y}{\partial a_p}\ da_p$ из (1.13) при $\lambda = -1$ имеем

$$(dy, \overline{\Delta y}|_{\lambda=-1}) = \sum_{p,k,l=1}^{m} da_p G_{pk} G^{-1}_{kl} \frac{\partial M}{\partial a_l} = dM, \tag{3.2}$$

т.е. выполнено условие, которое является определением градиента. Аналогично, выражение (1.12) для $\overline{\Delta a}$ при $\lambda = -1$ является градиентом M в пространстве a с метрикой[x] G_{ik}.

Доказательство того, что при использовании (1.12), (1.13) мы будем идти к минимуму M, не отличается по существу от такого же доказательства для обычного метода градиента. Действительно, при малых $\lambda > 0$ приращение M в результате шага $\overline{\Delta y}$ будет обязательно отрицательным, если только метрика G является положительно определенной. Последнее, очевидно, всегда имеет место, если начальное приближение $y(a^0, x)$ расположено в функциональной окрестности минимума M.

Перейдем к практическому использованию метода. Если $y(a^0, x)$ окажется вне функциональной окрестности минимума, метрику G_{ik} можно временно заменить на метрику

$$\overline{G}_{ik} = \int \left| \frac{\delta^2 M}{\delta y(z)\ \delta y(x)} \right| dz\ \frac{\partial y(x)}{\partial a_i}\ \frac{\partial y(x)}{\partial a_k}\ dx, \tag{3.3}$$

которая всюду положительно определена.

[x] Метрика G_{ik} преобразуется при замене (3.1) таким образом, что градиент (1.12) в отличие от обычного градиента $\nabla_i = \frac{\partial}{\partial a_i}$ преобразуется при линейной замене параметров, как радиус вектор a.

При малых λ приращение функционала $\Delta M = M\{y(a+\lambda\Delta a)\} - M\{y(a)\}$ растет с ростом λ, но при дальнейшем увеличении λ влияние отброшенных при аппроксимации (1.10) членов может стать настолько большим, что ΔM изменит знак. Максимальный шаг l_i по каждому параметру a_i, при котором нелинейная часть приращения y не может повлиять на знак ΔM, зависит от величины и строения матриц G и Q. Так как при значительном числе параметров a подсчитать матрицу Q сложнее, чем сделать несколько лишних шагов, то оптимальный шаг выгоднее находить подбором. Величины l_i от шага к шагу меняются медленнее, чем λ, и их пересмотр можно делать реже. Поэтому удобнее подбирать l_i, а λ вычислять по формуле

$$\lambda = \frac{1}{\max\left\{1, \dfrac{\Delta a_i}{l_i}\right\}} . \qquad (3.4)$$

В дальнейшем числа l_i будут предполагаться выбранными следующим образом. Первоначальные l_i берутся несколько завышенными так, чтобы нелинейная часть приращения $y(a_i \pm l_i) - y(a_i)$ в среднем равнялась линейной части. Если в результате шага Δa значение функционала M возросло, все l_i уменьшаются в 2 раза, и шаг повторяется. Если в результате двух последних шагов функционал M уменьшался, то перед следующим шагом те l_i, для которых $\Delta a_i > l_i$, удваиваются.

Если вблизи минимума

$$|Q_{ik}| << \sqrt{G_{ii}\, G_{kk}} \qquad (3.5)$$

для всех i,k, то $\lambda \to 1$, процесс итераций становится близким к ньютоновскому, и сходимость в конце процесса будет быстрой[x]. Уточнения положения минимума в этом случае будут происходить фактически по одномерной формуле Ньютона (2.1), где роль параметра A будет играть тот из параметров a_j, который окажется самым нелинейным. Случай нарушения условия (3.5) будет обсуждаться в следующем параграфе.

Упомянем две основные причины, способствующие выполнению условия (3.5)

---

x) Относительно сходимости приближенно-ньютоновских итераций см. /2/, гл. XУIII § 2.

на практике. Во-первых, функции $\frac{\partial^2 y(x)}{\partial a_i \partial a_k}$ обычно хорошо аппроксимируются линейной комбинацией производных $\frac{\partial y}{\partial a_i}$, а производные $\frac{\partial y}{\partial a_i}$ дают (вблизи минимума) при интегрировании с $\frac{\delta M}{\delta y(x)}$ нуль в силу уравнений экстремума $\frac{\partial M}{\partial a_i}=0$. Во-вторых, функция $\frac{\delta M}{\delta y(x)}$ вблизи минимума нередко близка к нулю из-за того, что абсолютный минимум функционала M близок к минимуму на семействе функций $y(a,x)$. Последнее имеет место, например, в методе наименьших квадратов, когда $M = \sum_{\xi} [y(a, x_\xi) - t_\xi]^2 w_\xi$.

Отметим, что в обоих упомянутых случаях условие (3.5) имеет тенденцию выполняться все лучше по мере того, как растет число параметров a и семейство $y(a,x)$ становится богаче. Этим, в частности, объясняется тот факт, что в методе линеаризации число шагов, необходимое для достижения минимума, практически не растет с увеличением числа параметров a.

§ 4. Устойчивость минимумов функционалов

Пусть функционал M, кроме параметров a, по которым производится минимизация, зависит также от некоторых параметров $t = (t_1, \dots, t_\mu, \dots)$. Мы сразу ограничимся случаем, когда параметры t не входят совсем в функциональный аргумент $y(a,x)$

$$\frac{\partial y(a,x)}{\partial t_\mu} \equiv 0, \tag{4.1}$$

и функционал M зависит от них только явно: $M = M\{t; y(a,x)\}$.
Нас будет интересовать зависимость положения минимума M от смещений параметров t.

Физических параметров, когда возникает подобного рода вопрос, можно привести очень много. Параметры t могут быть экспериментальными величинами, известными с ограниченной точностью. Насколько сместится минимум M, если параметры t сдвинутся на свои ошибки? Другим примером могут служить разного рода поправки и высшие члены разложения в ряды, которыми вследствие их малости при составлении функционала M мы пренебрегли, заменив нулями. Чувствительны ли результаты к этим поправкам?

Возьмем линейный член в разложении M в окрестности минимума в ряд Тейлора по приращениям параметра a_k и t_μ

$$M \simeq M_{min} + \sum_\mu \frac{\partial M}{\partial t_\mu} \Delta t_\mu + \sum_{k=1}^{m} \frac{\partial M}{\partial a_k} \Delta a_k \qquad (4.2)$$

и подставим это разложение в уравнения экстремума $\frac{\partial M}{\partial a_i} = 0$. Мы получим систему уравнений

$$\sum_\mu \frac{\partial^2 M}{\partial t_\mu \partial a_i} \Delta t_\mu + \sum_{k=1}^{m} \frac{\partial^2 M}{\partial a_k \partial a_i} \Delta a_k = 0, \quad i=1, \dots, m, \qquad (4.3)$$

дающую связь между смещением положения минимума Δa и приращением параметров t .

Качественные свойства системы (4.3) следующие. Если система (4.3) разрешима относительно Δa_k и определитель матрицы

$$\tilde{G}_{ki} = \frac{\partial^2 M}{\partial a_k \partial a_i} \qquad (4.4)$$

не является малым, то небольшие смещения параметров t будут вызывать незначительные перемещения минимума Δa , т.е. минимум будет устойчивым.

Когда в окрестности минимума $\det \tilde{G}$ мал, то элементы обратной матрицы $\tilde{G}^{-1}$ оказываются большими, и даже небольшие смещения параметров t могут вызвать значительные перемещения положения минимума. Мы назовем такие минимумы относительно неустойчивыми; подробнее этот случай рассматривается в § 5.

Особый случай возникает, если некоторые производные $\frac{\partial y(a,x)}{\partial a_k}$ (или их линейные комбинации) обращаются в минимуме в нуль. Такие минимумы мы будем называть вырожденными.

Обилие вырожденных минимумов характерно для функционалов (1), которые не зависят от a явно. Действительно, пусть

$$\frac{\partial y(a,x)}{\partial a_j} = 0, \qquad \frac{\partial^2 M}{\partial a_j^2} > 0 \qquad \text{при некотором} \quad a_j = \bar{a}_j, \qquad (4.5)$$

где равенство нулю производной выполняется тождественно по x и $\bar{a}_j$ может быть функцией остальных параметров $a_{k\neq j}$. Положим $a_j = \bar{a}_j$. Согласно (1), это влечет за собой выполнение уравнения экстремума $\frac{\partial M}{\partial a_j}=0$. Минимизируя M по параметрам $a_{k\neq j}$, мы можем добиться выполнения остальных уравнений экстремума $\frac{\partial M}{\partial a_{k\neq j}}=0$. Таким образом, каждое обращение в нуль одной из производных $\frac{\partial y}{\partial a_k}$ может привести к появлению вырожденного минимума функционала M.

С точки зрения устойчивости вырожденный минимум отличается тем, что по некоторым направлениям в пространстве параметров a он не может быть смещен никаким изменением параметров t ("сверхустойчив" в этих направлениях[x]). Например, в случае (4.5) минимум может перемещаться только по поверхности, заданной уравнением $a_j = \bar{a}_j$. Заметим, что высказанные только что утверждения существенно опираются на ограничение (4.1).

Набор минимумов, который имеет функционал, связан с тем, как параметризован его функциональный аргумент. Пусть функционал $M\{y(a,x)\}$ имеет некоторый спектр минимумов $M^{(i)}_{min} = M\{y^{(i)}(x)\}$. Если семейство $y(a,x)$ параметризовано так, что функциональный аргумент может принимать одно и то же значение $y^{(i)}(x)$ при нескольких значениях параметров a, то M как функция параметров a будет иметь несколько идентичных экземпляров этого минимума с точно совпадающей глубиной, и соответствующее спектральное значение $M^{(i)}_{min}$ будет кратным.

Введем новую параметризацию $b_k = b_k(a_1, \ldots, a_m)$ того же семейства. Очевидно, любой минимум $M\{y^{(i)}(x)\}$ функционала $M\{y(a,x)\}$, в котором якобиан $\frac{\partial(b_1,\ldots,b_m)}{\partial(a_1,\ldots,a_m)}$ отличен от нуля, должен присутствовать (по крайней мере в одном экземпляре) и в спектре $M\{y(b,x)\}$, так что изменения спектра минимумов могут быть связаны только с нулями якобианов $\frac{\partial(b)}{\partial(a)}$ и $\frac{\partial(a)}{\partial(b)}$. Так как невырожденный минимум при изменении параметров t перемещается в m-мерной области, а якобианы $\frac{\partial(b)}{\partial(a)}$, $\frac{\partial(a)}{\partial(b)}$ могут обращаться в нуль только

---

x) Устойчивость по остальным направлениям можно исследовать обычным образом, если вычеркнуть из системы (4.3) уравнения, которые не содержат приращений Δt_μ.

в подобласти меньшей размерности, то невырожденный минимум не может быть создан или уничтожен заменой параметров, хотя при такой замене число его экземпляров может измениться. Отсюда, в частности, следует, что если существует такая параметризация семейства $y(a,x)$, при которой M имеет единственный невырожденный минимум $M^{(0)}_{min}$, то спектр минимумов $M\{y(a,x)\}$ начинается с $M^{(0)}_{min}$, и все $M_{min} > M^{(0)}_{min}$ являются вырожденными.

В тех задачах, где функционал M является вспомогательной величиной и прямой интерес представляют лишь те значения параметров a, при которых он минимален, вырожденные минимумы соответствуют ложным решениям задачи. В самом деле, вырожденный минимум легко создать искусственно при любом значении $\tilde{a}_k$ любого параметра a_k, не изменяя существенно самого функционала M, а изменив только формально способ параметризации его функционального аргумента $y(x)$. Заменим, например, параметры a на b по формуле $b_k^2 + b_k^3 = (a_k - \tilde{a}_k)\operatorname{sign}\frac{\partial M}{\partial a_k}$. Теперь $\frac{\partial y}{\partial b_k} = (2b + 3b^2)\frac{\partial y}{\partial a_k} = 0$, $\frac{\partial^2 M}{\partial b_k^2} > 0$ при $b_k = 0$, и функционал M имеет m-кратно вырожденный минимум в выбранной нами точке $a_k = \tilde{a}_k$. Проделывая подобную процедуру в обратном порядке, формальной заменой параметров a_k можно избавить функционал M от любого из обнаруженных вырожденных минимумов (при этом в других местах могут возникнуть новые вырожденные минимумы).

В вырожденных минимумах процесс итераций сходится за счет уменьшения чисел l_i при каждом шаге. При такой "вынужденной" сходимости в конце процесса $\lambda \to 0$, что позволяет легко отличить вырожденные минимумы от обычных. Заметим, что могут существовать минимумы, в которых λ продолжает колебаться до самого конца процесса. Например, это минимумы, которые близки к вырожденным и переходят в последние при небольшом изменении параметров t (в таких минимумах условие (3.5) нарушено).

§ 5. Проблема неустойчивости

Значительная неустойчивость минимумов вызывается, как правило, некорректной постановкой задачи минимизации и не может быть устранена чисто формальным путем. Независимо от применяемого метода минимизации, неустойчивость неизбежно приводит к замедлению процесса поиска минимума и к потере точности при вычислениях из-за накопления ошибок округления. Механизм возникновения

этих затруднений может быть разным в разных случаях. В частности, в методе линеаризации становится малым λ и растет относительная ошибка элементов матрицы G^{-1} и величин, которые через эту матрицу вычисляются.

Неустойчивость минимума обычно проявляется задолго до того, как минимум найден. Удобным индикатором может служить безразмерная величина

$$\rho = \frac{G_{11} G_{22} \dots G_{mm}}{\det G}, \qquad \rho \geq 1, \tag{5.1}$$

которая близка к единице в районе относительно устойчивого минимума и много больше единицы в районе относительно неустойчивого. Более детальные сведения дают факторы корреляции

$$R_k = G_{kk} \, G^{-1}_{kk}, \qquad R_k \geq 1, \tag{5.2}$$

которые ведут себя так же как ρ и указывают неустойчивость по параметру a_k. Величины ρ и R_k тесно связаны между собой: если фиксировать параметр a_k, величина ρ уменьшится в R_k раз.

Факторы корреляции нужно знать для контроля за точностью вычислений. Можно показать, что относительная точность k-го диагонального элемента матрицы G^{-1} не меньше, чем в R_k раз хуже относительной точности того же элемента матрицы G. Пусть, например, матрица G известна с 9-ю значащими цифрами. Тогда, если хотя бы один из R_k превышает 10^{+8}, то в матрице G^{-1} нельзя будет гарантировать ни один верный знак независимо от способа ее вычисления.

ρ и R_k имеют простой геометрический смысл. Если функции $\frac{\partial y(x)}{\partial a_k}$ считать векторами в пространстве с метрикой $\frac{\delta^2 M}{\delta y(z) \delta y(x)}$ (ср. § 3), то ρ^{-1} равно квадрату объема параллелепипеда с единичными ребрами, построенного на этих векторах, а $R_k^{-1} = \sin^2 \psi$, где ψ – угол, который образует вектор $\frac{\partial y(x)}{\partial a_k}$ с плоскостью, в которой лежат остальные векторы.

Типичной причиной неустойчивости является неправильный выбор семейства функций $y(a,x)$, которое является функциональным аргументом. Этот случай легко узнать по тому признаку, что добавление каждого нового параметра a_k мало снижает минимум M и резко увеличивает ρ. Устойчивость восстанав-

ливается, если для $y(a,x)$ взять семейство, более отвечающее существу задачи. Пусть, например, $y(a,x)$ выбрано в виде $y=\sum_k a_k x^{k-1}$, а низких значений M следует ожидать, если $y(x)$ имеет вид кривой, изображенной на рис. 1.

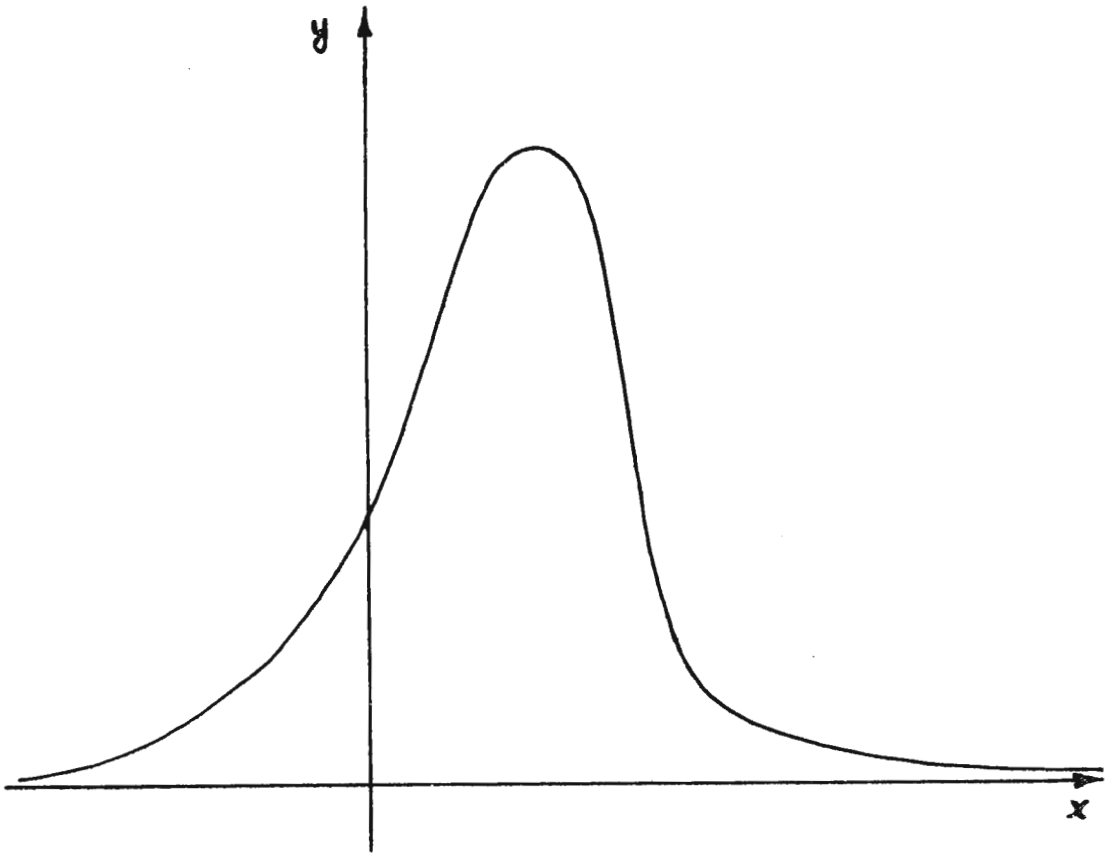

Рис. 1.

Ясно, что полиномом такую кривую описать трудно и коэффициенты a_k будут определяться из условия $M = \min$ очень плохо. В этом случае гораздо лучше, например, положить

$$y = (b_2 + b_4 x + \ldots)(1 + b_1 x + b_3 x^2 + \ldots)^{-1}.$$

Может случиться, что семейство $y(a,x)$ выбрано в соответствии с существом задачи, но ρ тем не менее велико. Это означает, что неудачна параметри-

зация семейства $y(a,x)$ и требуется нелинейная замена параметров[x] с большими R. При такой замене следует стремиться к тому, чтобы каждая из функций $\frac{\partial y(x)}{\partial a_k}$ имела максимум там, где остальные функции малы.

В худшем случае, если введенные параметры a_k являются именно теми величинами, ради которых производится минимизация, при больших ρ приходится отказаться от определения параметров с большими R, фиксировав часть из этих параметров.

§ 6. Некоторые применения

Квадратичный функционал

$$M = \sum_{\xi} [y(a,x_\xi) - t_\xi]^2 \; w_\xi , \tag{6.1}$$

минимизация которого требуется в методе наименьших квадратов[7,10,11], является идеальным объектом для применения метода линеаризации. Действительно, для функционала (6.1) малая функциональная окрестность минимума совпадает со всем функциональным пространством, и матрица Q в минимуме мала. Последнее связано с тем, что в

$$Q_{ik} = 2 \sum_{\xi} \frac{\partial^2 y(x_\xi)}{\partial a_i \, \partial a_k} \; [y(x_\xi) - t_\xi] \; w_\xi \tag{6.2}$$

входит малая в минимуме разность $y(x_\xi) - t_\xi$. Кроме того, если функциональный аргумент $y(a, x_\xi)$ допускает линейную параметризацию, то M имеет единственный невырожденный минимум[xx].

Подсчитаем матрицу ошибок параметров a

$$\sigma_{ik}^2 \equiv \overline{\Delta a_i \, \Delta a_k} \tag{6.3}$$

---

[x] Линейная замена, во-первых, не избавляет от потери точности, во-вторых, чтобы достаточно точно знать ее коэффициенты, надо сначала найти минимум.

[xx] Те же свойства имеют практически все функционалы $M = -2 \ln L$, где L – функция правдоподобия.

(черта сверху означает усреднение), сделав обычное для метода наименьших квадратов допущение, что экспериментальные величины t_ξ независимы

$$\overline{\Delta t_\xi \Delta t_\eta} = 0 \qquad \text{при} \quad \xi \neq \eta \tag{6.4}$$

и имеют дисперсии, равные обратным весам

$$\overline{\Delta t_\xi^2} = w_\xi^{-1} . \tag{6.5}$$

Подставляя (4.3), (4.4) в (6.3) и учитывая (6.4), (6.5), имеем

$$\sigma_{ik}^2 = \overline{\Big(\sum_{j=1}^{m} \sum_{\xi} \tilde{G}_{ij}^{-1} \frac{\partial^2 M}{\partial a_j \partial t_\xi} \Delta t_\xi\Big)\Big(\sum_{l=1}^{m} \sum_{\eta} \tilde{G}_{kl}^{-1} \frac{\partial^2 M}{\partial a_l \partial t_\eta} \Delta t_\eta\Big)} = \tag{6.6}$$

$$= \sum_{j,l=1}^{m} \tilde{G}_{ij}^{-1} \tilde{G}_{kl}^{-1} 2 G_{jl} .$$

Если Q мало, можно положить $\tilde{G}^{-1} \simeq G^{-1}$, тогда[x)]

$$\sigma_{ik}^2 \simeq 2 \sum_{j,l=1}^{m} G_{ij}^{-1} G_{kl}^{-1} G_{jl} = 2 G_{ik}^{-1} . \tag{6.7}$$

В вырожденном минимуме, где $G^{-1} \to \infty$, приближенная формула (6.7) неприменима. Из точной формулы (6.6) видно, что, как этого и следовало ожидать, у параметров a_j, по которым минимум вырожден, дисперсии σ_{jj}^2 равны нулю. С помощью матрицы ошибок σ_{ik}^2 легко вычислить, например, дисперсию функционального аргумента (квадрат коридора ошибок)

$$\overline{[\Delta y(a,x)]^2} \equiv \sigma^2(x) = \sum_{i,k=1}^{m} \frac{\partial y(x)}{\partial a_i} \sigma_{ik}^2 \frac{\partial y(x)}{\partial a_k} . \tag{6.8}$$

Дисперсии параметров a_k просто связаны с факторами корреляции R_k. Согласно (5.2) и (6.7)

$$\overline{\Delta a_i^2} = 2 (G_{ii})^{-1} R_i , \tag{6.9}$$

то есть дисперсию любого параметра можно уменьшить в R_i раз, если фиксировать остальные.

Пусть требуется найти действительные решения системы нелинейных уравнений

---

x) Чтобы избежать в (6.7) коэффициента 2, часто вместо G вводят матрицу, вдвое меньшую.

$$f_k(a_1, \dots, a_m) = 0, \quad k = 1, \dots, m. \tag{6.10}$$

Положим $y_\xi = f_\xi$, $t_\xi = 0$ и введем веса w_ξ произвольным образом, тогда решение системы (6.10) сведется к минимизации функционала (6.1). Аналогично можно искать комплексные решения системы (6.10), положив $y_\xi = \mathrm{Re}\, f_\xi$, $y_{\xi+m} = \mathrm{Im}\, f_\xi$ и взяв в качестве независимых параметров $a_\xi = \mathrm{Re}\, a_\xi$, $a_{\xi+m} = \mathrm{Im}\, a_\xi$ (сравни с (2.1)). Выбором весов w_ξ можно в некоторых пределах влиять на величину факторов корреляции.

В рассматриваемом частном случае, когда число параметров совпадает с числом точек, на котором задан функционал, применение метода линеаризации к минимизации функционала (6.1) эквивалентно применению метода Ньютона непосредственно к системе (6.10), однако метод линеаризации за счет выбора λ обеспечивает сходимость в более широком классе случаев и имеет более удобную систему контроля.

В случае громоздких систем (6.10) минимум M может оказаться относительно неустойчивым к отклонению величин t_ξ от нуля, что приводит к техническим трудностям. Очевидно, неустойчивость такого рода никак не связана с существом задачи и носит формальный характер. Поэтому устойчивость всегда можно восстановить некоторым нелинейным преобразованием системы[x] (6.10).

В вырожденных минимумах функционал $M = \sum_\xi y^2 w_\xi$ не становится равным нулю, так что такие минимумы не являются решениями системы (6.10). Если отыскиваются действительные решения, то некоторые вырожденные минимумы могут указывать на наличие вблизи них пары комплексных решений с малой мнимой частью. В случае же комплексных решений системы (6.10) вырожденные минимумы могут обнаруживаться только в тех точках, где функции f_k не аналитичны по одному или нескольким параметрам a_k. Действительно, если в некоторой точке $\frac{\partial f_k}{\partial a_p} = 0$ и функции f_k аналитичны по a_p, то $|f|$ в одном из направлений обязательно убывает, и процесс поиска минимума $M = \sum_{k=1}^{m} |f_k|^2$ не задержится в такой точке.

---

x) См. примечание на стр. 16.

В заключение авторы считают своим приятным долгом выразить благодарность Н.П. Клепикову, Я.А. Смородинскому, Е.П. Жидкову, Н.Н. Говоруну, Ю.М. Казаринову, Р.М. Джабар-Заде и Г.П. Ососкову за ценные замечания.

Литература

1. Н.С. Березин, Н.П. Жидков. Методы вычислений гл. У1, У11, Физматгиз, Москва, 1960.

2. Л.В. Канторович, Г.П. Акилов. Функциональный анализ в нормированных пространствах. Физматгиз, Москва, 1959.

3. И.М. Гельфанд, М.Л. Цетлин. Принципы нелокального поиска в системах автоматической оптимизации. ДАН, 1961, т.137, № 2.

4. Н.П. Клепиков, В.А. Мещеряков, С.Н. Соколов. Анализ экспериментальных данных по полным сечениям взаимодействия π -мезонов с протонами. Препринт ОИЯИ Д-584, Дубна, 1960.

5. Н.С. Амаглобели, Ю.М. Казаринов, С.Н. Соколов, И.Н. Силин. Определение константы π -мезон-нуклонного взаимодействия по дифференциальным сечениям упругого np -рассеяния. ЖЭТФ, т.39, вып. 4(10), 1960.

6. Лю Юйань, Н.И. Пятов, В.Г. Соловьев, И.Н. Силин, В.И. Фурман. О свойствах ряда сильно-деформированных ядер. ЖЭТФ, 40, 1501 (1961).

7. А. Хальд. Математическая статистика с техническими приложениями. Гл. XX, § 1У, ИЛ, Москва, 1956.

8. И.Н. Бронштейн и К.А. Семендяев. Справочник по математике для инженеров и учащихся втузов. Физматгиз, Москва, 1959, стр. 570.

9. В.В. Воеводин. Применение метода спуска для определения всех корней алгебраического многочлена. ЖВММФ т.1, № 2. 1961 .

10. Т. Крамер. Математические методы статистики. ИИЛ Москва, 1948 .

11. Н.П. Клепиков, С.Н. Соколов. Анализ экспериментальных данных методом максимума правдоподобия. Препринт ОИЯИ Р-235, Дубна, 1958.

12. Л.А. Люстерник, В.И. Соболев. Элементы функционального анализа. § 41 ГИТТЛ, 1951.

Рукопись поступила в издательский отдел
12 октября 1961 года.