УДК 519.853.4

РЕШИТЕЛЬНЫЙ ПРОГРЕСС В ПРОБЛЕМЕ МИНИМИЗАЦИИ С ОГРАНИЧЕНИЯМИ

И.Н.Силин

Исследование неприятностей в сложном случае применения FUMIVI привело к серьезному прояснению проблемы минимизации регулярных функций и функционалов с регулярными ограничениями. Под регулярностью у нас понимается непрерывность вторых производных (по крайней мере, в окрестности минимума). Проблема почти закрыта. Выяснилось также, что нужно для окончательного ее закрытия.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Resolute Progress in Constrained Minimization Problem

I.N.Silin

Investigation of troubles in a difficult case of FUMIVI application has resulted in serious clearing of a problem of minimization of regular functions and functionals with regular constraints. Under regularity the second derivatives continuity (at least in a vicinity of a minimum) is understood here. The problem is almost closed. What is necessary for its final closing has been also found out.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

В [1] мы, подобно многим авторам, поддались искушению прямого учета геометрии разрешенной для поиска минимума области (исходной и/или вспомогательной). После ряда успешных применений программы FUMIVI [2] мы неожиданно столкнулись с серьезными проблемами при массовой обработке модельных данных для конкурирующих процессов при поиске $K^- \to \mu^-\nu\gamma\pi^0$ распада на установке ИСТРА-М [2]. Неприятным сюрпризом оказались сложности геометрии границ, не связанные непосредственно с минимумом, но затрудняющие движение к нему.

Какого рода трудности могут возникать, покажем на простых примерах. Рассмотрим поиск минимума $F(x, y)$ с границами

$$y \geq x \quad \text{и} \quad y \geq x \cdot (1 + \varepsilon),$$

причем минимум $F(x, y)$ без ограничений, расположен при $x < 0$, $y << x$. Если начальное приближение расположено при $x > 0$ и $y > x \cdot (1 + \varepsilon)$, то при движении к минимуму мы придем сначала на границу $y \geq x \cdot (1 + \varepsilon)$ и затем вдоль нее в вершину $x = 0$, $y = 0$.

При малом ε мы вынуждены будем для анализа вершины разлагать градиент минимизируемой функции по почти коллинеарным градиентам ограничений — типичный случай плохо обусловленной задачи. В то же время с точки зрения здравого смысла — это всего лишь мелкий (например, в пределах машинной точности) излом границы, не влияющий на решение, расположенное при $x < 0$, $y = x$. В случае двух переменных мы легко, на глаз, разобрались в ситуации. Но глазомер не поможет в m-мерном подпространстве n-мерного пространства.

Ситуация еще хуже при избыточности связей. Усложним пример, добавив неравенство $y \geq x \cdot (1 + \varepsilon/2)$. Вообще говоря, необходимо перебрать все варианты исключения лишних связей, чтобы найти выход к меньшим значениям функции $F(\mathbf{x})$. При большой же размерности и большой переопределенности таких вариантов очень много.

Проблема может показаться слегка надуманной. Разумный потребитель старается избегать подобных проблем ещё при постановке задачи. Однако каждый может споткнуться. Например, в методе исследования долин (программа FUMIVI [1]) для отслеживания искривленных долин и границ нами вводятся временные границы изменения переменных, которые в сложных ситуациях неизбежно приводят к переопределенности. Отказ же от временных границ разрушает сходимость метода.

Это заставило снова вспомнить о, казалось бы, примитивных методах штрафных и барьерных функций. Например, для учета каждого равенства $f_i(\mathbf{x}) = 0$ можно добавлять к минимизируемой функции член $w\,(f_i(\mathbf{x})/sf_i)^2$, а для учета каждого неравенства $f_i(\mathbf{x}) \geq 0$ — добавлять $\ln\,(sf_i/f_i(\mathbf{x}))/w$. $1/w \to 0$. Здесь sf_i — масштаб $f_i(\mathbf{x})$, зависящий от масштабов x_j. В задачах обработки экспериментальной информации наиболее приемлемы в качестве sf_i оценки среднеквадратичных погрешностей $f_i(\mathbf{x}_e)$, как функции оценок $\mathbf{x}_e$ искомых величин x_j и оценок их погрешностей sx_j. К сожалению, часто такие оценки становятся известными только в результате минимизации. Так что, на первой стадии приходится довольствоваться какими-то оценками этих оценок.

Почему следует рекомендовать логарифмический барьер, а не, например, барьер $1/(w \cdot f(\mathbf{x}))$, как часто советуют? Довольно просто убедиться, что при введении квадратичного штрафа погрешность решения в нормальных ситуациях стремится к нулю как $O(1/w)$. Такую же сходимость дает введение логарифмического барьера, в то время как барьер $1/(w \cdot f(\mathbf{x}))$ дает сходимость как $O(1/\sqrt{w})$. К тому же логарифмический барьер был хорошо испытан при минимизации отрицательного логарифма функции правдоподобия. Такая возможность была заложена автором в программу FUMILI [3,4,5] при реализации метода линеаризации функционального аргумента [6] (не путать с методом линеаризации минимизируемой функции, каковой тоже существует). Была предусмотрена и борьба со случайным перепрыгиванием барьера. Правда, у логарифма функции правдоподобия барьер естественный, связанный с требованием положительности вероятностей, и не требует его последовательного подавления. Тогда же автором была надежно решена проблема простых ограничений вида $\min x_i \leq x_i \leq \max x_i$. Разра-

ботанный для FUMILI алгоритм гарантирует нахождение минимума с простыми ограничениями квадратичной строго выпуклой функции за конечное число шагов. Доказательство этого факта получено автором только сейчас с использованием современной техники подобных доказательств. Алгоритм неплохо работает и в неквадратичном случае, если в окрестности минимума функция строго выпукла. Таким образом, метод штрафных и барьерных функций пользователь может реализовать даже с помощью FUMILI, программируя в собственной версии подпрограммы SGZ добавление штрафных и барьерных членов к минимизируемой функции, ее градиенту и приближенной матрице вторых производных. Более того, при обнаружении плохой сходимости пользователь может начать вычислять точные вторые производные. Если это будет вблизи минимума, то FUMILI сможет даже преодолеть временное появление неположительно определенных матриц.

Но, конечно, предпочтительнее использовать FUMIVI (в режиме только простых ограничений). Она специально рассчитана на работу со сложным рельефом, возникающим при больших w (именно такие функции использовались для жесткого тестирования её свойств). Кроме того, FUMIVI рассчитана на функции и функционалы произвольного вида, и к тому же выполняет вспомогательные шаги для учета простых границ, используя квадратичное приближение минимизируемой функции без её перевычисления. Алгоритм FUMIVI гарантирует нахождение за конечное число вспомогательных шагов одного из минимумов с ограничениями для произвольной квадратичной функции (в том числе и линейной).

В [1] не отмечено, что для хорошей сходимости может потребоваться (а для квадратичной сходимости обязателен) учет вторых производных уравнений связи. В методе штрафов и барьеров он произойдет автоматически в случае полного дифференцирования вспомогательной (со штрафами и барьерами) функции.

Мы не зря рассматривали скорость сходимости к решению при стремлении w к бесконечности. Дело в том, что когда она известна, возможна экстраполяция по Ричардсону. Автор не слышал, чтобы кто-то обращал на это внимание в применении к методам штрафов и барьеров. А между тем, применение экстраполяции позволяет избегать слишком больших w. Кроме того, при последовательном уточнении решения с возрастающими w можно начинать минимизацию с хорошего (экстраполированного) начального приближения. Это, в свою очередь, позволяет применять и простые методы минимизации, не рассчитанные на большие перемещения при сложном рельефе.

Все было бы хорошо, но барьерная функция создает долины с особо неприятным несимметричным профилем. К тому же сохраняется проблема попадания в разрешенную область. Можно пробовать несимметричные штрафы, но неприятности будут похожими.

В многочисленных исследованиях по проблеме минимизации с ограничениями высказано множество идей разной степени красоты и полезности (либо вредности). Рекордной по красоте и вредоносности, по-видимому, нужно считать идею множителей Лагранжа. Она породила массу малопродуктивных исследований в попытке их как-то с пользой употребить. Есть еще одна, чем-то похожая, идея — по замене неравенств равенствами. Предлагается заменять неравенство $f(\mathbf{x}) \geq 0$ равенством $f(\mathbf{x}) = t$ и про-

стым неравенством $t \geq 0$. Поскольку с простыми неравенствами мы справляться умеем, то для нас это выход. Равенство можно заменить квадратичным штрафом. Но вполне можно ожидать каких-то неприятностей в духе множителей Лагранжа. Однако оказалось, что в данном случае закон неубывающей хаотичности (или, по-другому, энтропии) не сработал. Парадоксальным образом все настолько хорошо, что введение штрафов в задаче квадратичного программирования с линейными неравенствами не портит квадратичности, и решение по-прежнему находится за конечное число шагов.

Появившиеся при замене переменных дополнительные параметры при минимальном дополнительном усилии мысли могут вычисляться без манипулирования матрицами увеличенной размерности. Шаг по $\mathbf{x}$ для определения направления движения к приближению $k + 1$ вычисляется при значениях $\mathbf{t}$ из приближения k, а шаг по $\mathbf{t}$ — из уравнений $\mathbf{t} + d\mathbf{t} = \mathbf{f}(\mathbf{x} + d\mathbf{x})$. В то же время полностью снимаются проблемы со сложностью геометрии и с попаданием в разрешенную область. Исчезает и проблема обнаружения локальной несовместности уравнений связи. Просто в случае несовместности мы обнаружим, что вклад от штрафных членов с ростом w не убывает, а возрастает, а невязки некоторых из связей стремятся к константам, отличным от нуля.

Этот прием оказывается выгодным даже в задачах линейного программирования с очень сложной геометрией границ.

Тем не менее полностью отказываться от барьерных функций не следует. Бывают случаи, когда минимизируемая функция не определена вне разрешенной области, и минимизацию следует вести, никогда не выходя из нее. Для попадания в разрешенную область можно для начала провести минимизацию со штрафами какой-либо безопасной функции — квадратичной, линейной или даже константы.

После увиденных потрясающих эффектов возникает подозрение, что при попытках прямого учета уравнений связи, в том числе и с множителями Лагранжа, чего-то не досмотрели. Не может быть, что ситуация там принципиально безнадежно хуже, чем в методе штрафов и барьеров.

Для начала переформулируем задачу. Вместо условия выполнения уравнений связи будем требовать минимального нарушения ограничений, конкретно — минимума суммы квадратов невязок в безразмерных масштабах (как и выше) и с заменой не простых неравенств равенствами. Такая постановка задачи законна — некоторая несовместность связей может быть вызвана неточным знанием некоторых констант. В то же время для совместимых связей эта постановка эквивалентна исходной.

Однако если мы согласились с невыполнением условий связи, то должны отказаться от критерия минимальности, связанного со знаками коэффициентов разложения градиента $F(\mathbf{x})$ по градиентам связей $f_i(\mathbf{x})$ и активных простых связей $x_i - c_i$. В этом критерии предполагается, что связи удовлетворены достаточно точно, чтобы градиенты не были искажены нелинейными эффектами.

В работе [1] нами была использована предложенная Курбатовым эффективная техника [7] динамического исключения из квадратичного приближения $F(\mathbf{x})$ части переменных с использованием линейного разложения уравнений связи. Там же было упомянуто, что при возникновении несовместности желательно было бы проводить исключение, пользуясь условиями минимума суммы квадратов нарушений связей. Сейчас

кажется, что это просто необходимо. Может быть применена та же техника, но с использованием линейного разложения уравнений минимума суммы квадратов невязок. При необходимости могут быть учтены и вторые производные уравнений связи. Так как после исключения остались только простые связи, то мы можем ограничиться проверкой только знаков проекций на систему координат **x** (точнее, на активные границы) преобразованного градиента преобразованной квадратичной функции!

По сравнению с [1] алгоритм непосредственного учета связей драматически упростился и преодолены трудности со сложной геометрией. Конечно, кое-какие проблемы могут быть при линейной зависимости уравнений минимального нарушения связей.

Нужно обратить внимание на следующий парадокс. FUMIVI в большинстве случаев работает быстрее даже по числу итераций (при существенно меньшей трудоёмкости итераций) в режиме приближенных вторых производных $F(\mathbf{x})$ (линеаризация функционального аргумента и неучет вторых производных связей). Подобный парадокс существовал в численном многократном интегрировании, принципиально неточный метод Монте-Карло работал лучше, чем рекурсивно повторенные схемы высокого порядка точности (для однократных интегралов). Парадокс был разрешен созданием метода Коробова [8] и ряда других методов с хорошей асимптотической сходимостью. Так что не следует терять надежды на существенное улучшение эффективности алгоритма минимизации функций с неизвестной структурой. У автора есть некоторые смутные идеи в связи с этим. По крайней мере, очевидно, что не имеют перспектив методы переменной метрики в современной форме. Они не способны конкурировать даже с работой FUMIVI без учета функциональной структуры.

Хотелось бы немного коснуться проблемы глобального минимума. Как известно, чисто формально проблема локализации глобального минимума решена средствами интервальных вычислений [9]. При интервальных вычислениях в каждом арифметическом действии и каждой элементарной функции вычисляется интервал, в котором строго расположен результат, если заданы интервалы расположения аргументов. Это позволяет вычислять интервал расположения результата для произвольных функций. Для наших целей интересны большие интервалы переменных минимизируемой функции. Пространство поиска может разбиваться на клетки, и из рассмотрения изымаются большие участки пространства, для которых интервал значений функции лежит выше наименьшего известного значения функции. Учитывая, что вычисление интервала не намного труднее вычисления функции в одной точке, предприятие не кажется совершенно безнадежным. Конечно, есть проблемы с большими интервалами в программах с автоматической проверкой точности результата и, вообще, существенно использующих логические условия. Тем не менее было бы не вредно иметь в трансляторах с распространенных языков программирования режим интервальной арифметики, наподобие комплексной арифметики. Это полезно и для оценки достоверности вычислений. Впрочем, это может быть сделано и на уровне объектно-ориентированного программирования, хотя должно делаться профессионально.

Автор глубоко признателен С.Н.Соколову, В.С.Курбатову и ныне покойному Ю.М.Казаринову, вовлекшим его в данную деятельность и внесшим существенный вклад в решение проблемы.

Литература

1. Kurbatov V.S., Silin I.N. — Nucl. Instr. and Meth., 1994, v.A345, p.346.
2. Артемов В.М. и др. — ЯФ, 1997, т.60, №12, с.2205.
3. Silin I.N. — CERN Program Library, D510, FUMILI.
4. Силин И.Н. — Поиск максимума функции правдоподобия методом линеаризации. Приложение III к русскому переводу книги: Идье В.Т. и др. — Статистические методы в экспериментальной физике. М.: Атомиздат, 1976.
5. Родионов А.И., Силин И.Н. — Алгольный вариант программы FUMILI — минимизация функционалов методом линеаризации. Совместный научный сборник ОИЯИ, Дубна и ЦИФИ, Будапешт «Алгоритмы и программы для решения некоторых задач физики». Будапешт, RFKI-74- 34, 1974, с.113.
6. Соколов С.Н., Силин И.Н. — Препринт ОИЯИ Д-810, Дубна, 1961.
7. Ketikian A.J. et al. — Nucl. Instr. and Meth., 1992, v.A314, p.578.
8. Коробов Н.М. — ДАН СССР, 1959, т.124, №6.
9. Moore R.E. — Interval Arithmetic and Automatic Error Analysis in Digital Computing. Tech. Rep. 25, Stanford Univ. Applied Mathematics and Statistics Laboratories, 1962.

Рукопись поступила 29 мая 1998 года.