从瞎子爬山到最优化方法

袁亚湘（中国科学院数学与系统科学研究院）

看到标题，读者一定会问：瞎子爬山和最优化方法有什么关系？事实上，爬山的目标是登上山顶，也就是要找海拔最高的点；而最优化是在一定约束条件下寻求某个目标函数的最大值或最小值。所以爬山本身就是一个优化问题。给定一个点，计算机可以计算目标函数在该点的信息（如函数值，梯度值），但不知道其他点的信息。这正如一个瞎子在山坡上能感觉到脚下的坡度（这是海拔函数在当前点的梯度值），但不知道山上的其他点的任何情况。可见计算机的能力和瞎子是差不多的。正因为如此，我们说，用计算机求解最优化问题和瞎子爬山有惊人的相似之处。

黄山　天都峰

把计算机的能力和瞎子对比可能已经出人意料了，但我想问一个更让大家吃惊的问题：计算机和瞎子谁更聪明？我国已故著名数学家华罗庚先生曾把一个简单的优化方法称之为“瞎子爬山法”，该方法就是相当于瞎子在爬山时用明杖前后左右轮流试，能往上走

就迈一步直到四面都不高了就是山顶。这个方法本质上就是坐标轮换搜索法。现实生活中，瞎子肯定不会这样爬山的，可见瞎子就比采用坐标轮换法的计算机聪明。我更偏向于把最速下降法称为“瞎子爬山法”，理由是瞎子能知道山的坡度。

华罗庚 (1910-1985)

最速下降法是利用最速下降方向求函数极小的方法，这相当于在爬山中沿着山坡最陡的方向往前爬。在数学上，就是求解极小化问题

$$\min_{x\in\Re^n} f(x) \tag{1}$$

的迭代法：

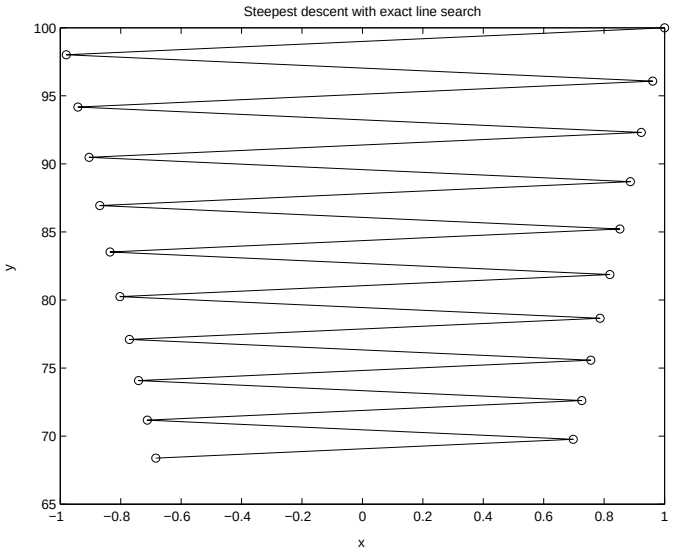
$$x_{k+1} = x_k + \alpha_k(-\nabla f(x_k))\,,$$

其中 $\alpha_k > 0$ 是步长。 α_k 的一个直观的选取是使得目标函数 $f(x)$ 尽可能的小，也就是让 $\alpha_k = \alpha^*$ 满足精确搜索条件：

$$f(x_k - \alpha^*\nabla f(x_k)) = \min_{\alpha>0} f(x_k - \alpha\nabla f(x_k))\,.$$

这就是精确搜索下的梯度法，通常称为最速下降法。

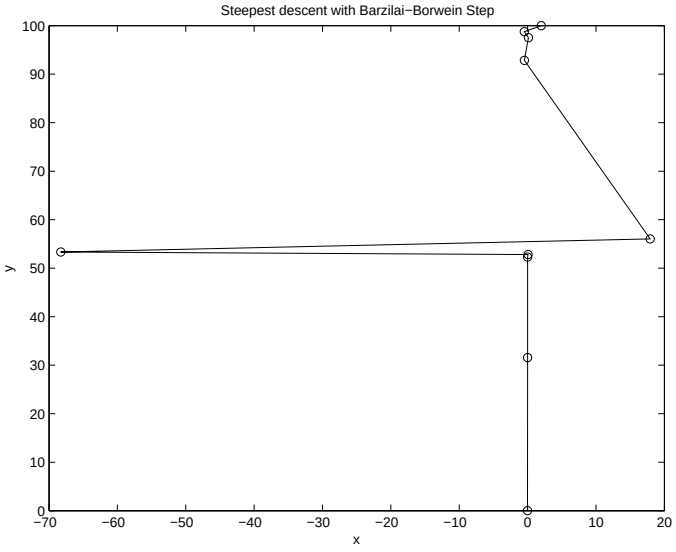
表面上看来，最速下降法是个完美的方法。该方法所用的方向是最好的（使函数降得最快），步长也是最好的（让函数在搜索方向上最小）。但是，最速下降法不仅不是一个最好的方法，反倒是一个很差的方法。下图是用最速下降法求解 $\min f(x,y) = 100x^2 + y^2$ 从初始点 $(1, 100)$ 开始迭代的前二十个迭代点：



从上图可以看出，最速下降法收敛非常慢。也就是说，“最好” + “最好” $\neq$ “最好”。我在中科院研究生院上课时常常跟同学们开玩笑说，班上最好的男生娶班上最好的女生，结果往往不是最好的。

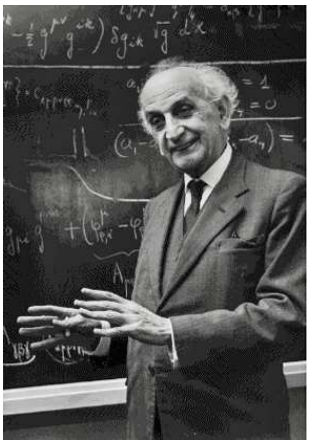
1988 年加拿大数学会前会长、加拿大皇家科学院院士 Borwein 教授和合作者 Barzilai 提出了一个巧妙的办法来改进最速下降法。他们把上一次迭代的最好步长留着下一次迭代用。这一小小的改动，导致新算法效率惊人地提高，几乎可以达到和共轭梯度法差不多的

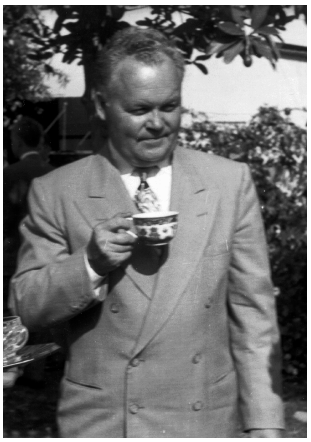
效果。下图是用 Barzilai-Borwein 方法求解 $\min f(x,y) = 100x^2 + y^2$ 从初始点 $(1, 100)$ 开始迭代的表现：



由此图可知， BB 方法只需九次迭代就得到一个非常高精度的解。 BB 方法的提出使得优化专家们对梯度法不得不重新认识，并引发了大量的后续研究，英国皇家学会会员、优化最高奖 Dantzig 奖获得者 Roger Fletcher 等著名学者也对这个问题作了深入研究。但是，如此重要的 BB 方法本质上却如此简单，就是把最好的步长延迟一步用。继续上面提到的玩笑就是，班上最好的男生应该找低年级最好的女生。

优化方法中另外一个应用广泛的方法是共轭梯度法。该方法最早是用来求解线性方程组的，由著名数学家 Cornelius Lanczos(1893-1974), Magnus Hestenes(1906-1991) 和 Eduard Stiefel(1909-1978) 等提出。


Cronelius Lanczos


Magnus Hestenes


Eduard Stiefel

共轭梯度法的基本思想是把一个 N 维问题转化为 N 个一维问题。方法的关键是构造一组两两共轭的方向。巧妙的是，共轭方向可以由上次搜索方向和当前点的梯度方向之组合来逐步产生：

$$d_{k+1} = -\nabla f(x_{k+1}) + \beta_k d_k.$$

不同的 β_k 导致不同的非线性共轭梯度法，著名的方法有：Hestenes-Stiefel 方法、 Fletcher-Reeves 方法、 Polak-Ribiére-Polyak 方法和 Dai-Yuan 方法，其对应的 β_k 的选取分别为：

$$\begin{aligned}
\beta_k^{HS} &= (g_{k+1} - g_k)^{\mathrm{T}} g_{k+1} / d_k^{\mathrm{T}} (g_{k+1} - g_k) \ , \\
\beta_k^{FR} &= \|g_{k+1}\|_2^2 / \|g_k\|_2^2 \ , \\
\beta_k^{PRP} &= (g_{k+1} - g_k)^{\mathrm{T}} g_{k+1} / \|g_k\|_2^2 \ , \\
\beta_k^{DY} &= \|g_{k+1}\|_2^2 / d_k^{\mathrm{T}} (g_{k+1} - g_k).
\end{aligned}$$

显然可以看出，这四个不同的 β_k 可通过两个分子和两个分母的组合来得到。这给我们的一个启迪是：完备性和对称性能引导我们发现新的方法。

信赖域方法是英国皇家学会会员、美国科学院外籍院士、首届 Dantzig 奖获得者、英国剑桥大学教授 Powell 最先提出的。在过去的

三十年中人们对信赖域方法的研究取得了巨大的进展，并使得信赖域方法一直是非线性优化研究的中心和热点。这样一个对学科发展起了巨大推动作用的方法其基本思想却非常简单。它不像线搜索方法那样先求搜索方向然后求步长，而是每次迭代在一个区域内试图找到一个好的点. 该区域称为信赖域，通常是以当前迭代点为中心的一个小邻域. 试探点往往要求是原优化问题的某个近似问题在信赖域的解. 试探点求出后利用某一评价函数来判断它是否可以被接受为下一个迭代点. 试探点的好坏还被用来决定如何调节信赖域. 粗略地说，如果试探点较好，则信赖域保持不变或扩大；否则将缩小.

正式的教科书追溯信赖域历史往往会提到求解求解非线性最小二乘问题 $\min ||F(x)||_2^2$ 的 Levenberg-Marquadt 方法。因为 Levenberg-Marquardt 步

$$d_k = -(J(x_k)J(x_k)^T + \lambda_k I)^{-1}F(x_k)$$

是线性化最小二乘问题

$$\min ||F(x_k) + J(x_k)d||_2^2$$

在某一个信赖域上的解，其中 $J(x_k) = \nabla F(x_k)$ 。如果没有信赖域约束，该问题的解就是 Gauss-Newton 步。又可以开个玩笑： Gauss-Newton 法是一个很“值钱”的方法, 因为 Carl Friedrich Gauss(1777-1855) 和 Issac Newton(1642-1727) 都上过各自所在国的货币。

德国马克上的高斯

英国英镑上的牛顿

Newton 与优化的联系是相当多的。事实上,求函数极小 $\min_{x\in\Re^n} f(x)$ 的一个基本方法就是 Newton 法，它的搜索方向是：

$$d_k^N = -(\nabla^2 f(x_k))^{-1}\nabla f(x_k).$$

牛顿法是一个几乎完美的方法，它不仅简单，而且收敛到 $f(x)$ 的极小值点 x^* 的速度非常快。在二阶充分条件下，可证明它具有 Q 二次收敛性：

$$||x_k + d_k^N - x^*|| = O(||x_k - x^*||^2).$$

但是,美好的东西往往是可望不可及的。在实际应用中,特别是对于大规模问题,牛顿法是没法用的,这是因为二阶偏导数矩阵 $\nabla^2 f(x_k)$ 的计算量太大，甚至根本无法计算。

1959 年诞生的拟牛顿 (quasi-Newton) 方法将牛顿法中的二阶偏导数矩阵用一个拟牛顿矩阵来代替,避免了计算二阶偏导数,而且通过逐步修正拟牛顿阵,也能使方法达到超线性收敛。英国皇家学会会员、牛津大学的 Trefethen 教授将拟牛顿法与有限元、快速傅立叶变换及小波等并列为二十世纪最重要的计算方法之一。欧美优化界的好几位院士都在拟牛顿法方面有深入的研究。 Fletcher 和 Powell 关于拟牛顿法的第一篇文章的 SCI 引用已超过 3100 次。

拟牛顿法的核心就是将 Newton 法的 $\nabla^2 f(x_k)$ 用一个拟牛顿矩阵 B_k 代替。拟牛顿矩阵满足拟牛顿公式：

$$B_{k+1}s_k = y_k$$

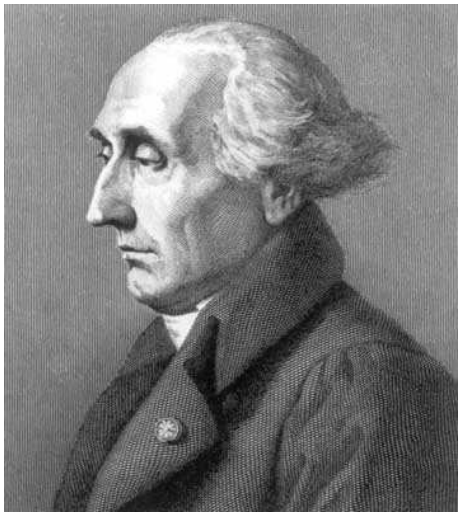
其中 $s_k = x_{k+1} - x_k$, $y_k = \nabla f(x_{k+1}) - \nabla f(x_k)$. 第一个出现的拟牛顿法是 Davidon-Fletcher-Powell 方法，其拟牛顿矩阵修正公式为

$$B_{k+1}^{DFP} = B_k - \frac{B_k s_k y_k^T + y_k s_k^T B_k}{s_k^T y_k} + \left(1 + \frac{s_k^T B_k s_k}{s_k^T y_k}\right)\frac{y_k y_k^T}{s_k^T y_k}.$$

而目前公认最好的拟牛顿法是 Broyden-Fletcher-Goldfarb-Shanno 方法：

$$B_{k+1}^{BFGS} = B_k - \frac{B_k s_k s_k^T B_k}{s_k^T B_k s_k} + \frac{y_k y_k^T}{s_k^T y_k}.$$

美国西北大学 Nocedal 教授（ 1998 年国际数学家大会 45 分钟报告者） 1992 年在剑桥大学出版的综述论文集《 Acta Numerica 》中提出两个公开问题：“ DFP 方法对强凸函数是否收敛？”和“ BFGS 方法对于非凸函数是否收敛？”第二个公开问题被戴虹利用 Powell 关于 PRP 方法的例子巧妙解决。而第一个公开问题虽有一些进展但至今还未彻底解决。拟牛顿法给我们的启迪是：近似和逼近是构造优化方法有力武器。



Joseph Louis Lagrange (1736-1813)

另一个著名数学家， Lagrange 和优化也有紧密的联系。事实上寻找有约束条件的优化问题

$$\min_{x \in \Re^n} \quad f(x) \tag{2}$$

$$\text{s. t.} \quad c_i(x) = 0, \quad i = 1, 2, .., m_e; \tag{3}$$

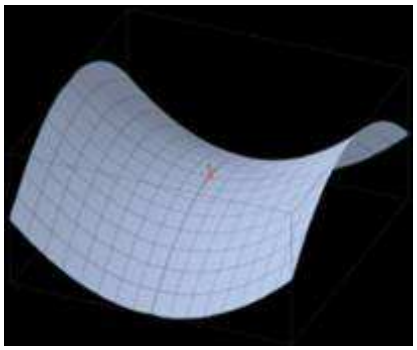
$$c_i(x) \geq 0, \quad i = m_e + 1, ..., m. \tag{4}$$

最优解等价于计算 Lagrange 函数

$$L(x, \lambda) = f(x) - \sum_{i=1}^{m} \lambda_i c(x)$$

的鞍点。

中国古代马鞍



matlab 图形 $z = x^2 - y^2$

λ_i 称为 Lagrange 乘子，对应于不等式约束的 Lagrange 乘子应该是非负的。 Lagrange 函数对优化的重要性不仅体现在刻画最优性条件上，同时它在优化计算方法的构造上起了巨大的作用。例如，著名的逐步二次规划方法 (SQP) 就是基于 Lagrange-Newton 法发展起来的。

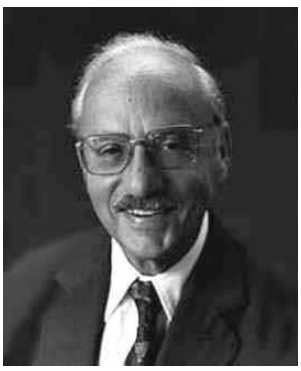
上个世纪优化的另一个重大突破是内点法提出和兴起。线性规划是最简单的约束优化问题，它的标准形式如下：

$$\min_{x \in \Re^n} \quad c^T x \tag{5}$$

$$\text{subject to} \quad Ax = b \tag{6}$$

$$x \geq 0 \tag{7}$$

其中 $c \in \Re^n$, $A \in \Re^{m\times n}$, $b \in \Re^m$. 线性规划在经济，国防等许多重要领域有着广泛的应用。线性规划的奠基者有优化先驱 George Dantzig(1914-2005)，诺贝尔奖获得者 Leonid Kantorovich (1912-1986)，和著名数学家 John von Neumann(1903-1957)。


George Dantzig

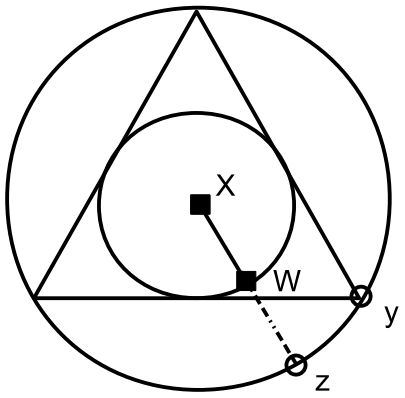

Leonid Kantorovich


John von Neumann

在几何上，线性规划可理解为求凸多面体最低的点。Dantzig 提出的求解线性规划的单纯形法 (Simplex Method) 本质上就是每次从凸多面体的一个顶点走到相邻的一个更低的顶点而逐步找到最低点的方法。单纯形法具有简单、直观等优点，同时它用来求解大多数线性规划问题也是非常快的。但是，可以构造例子使得单纯形法走遍凸多面体的每个顶点，于是可知单纯形法的复杂度是指数的。

1984 年美国贝尔实验室的印度数学家 Karmarkar 提出了一个具有多项式复杂度的方法。Karmarkar 方法的基本思想是从凸多面体的内部而不像单纯形那样在边界上去逐步靠近最优解。Karmarkar 方法可看成是利用 Newton 法求解 log 罚函数的方法。事实上，Karmarkar 方法有一个很简单的同时也是很巧妙的几何解释。举一个简单的例子来说明这一点。假定我们需要寻找三角形的最低顶点 y. 设当前点是三角形的中心 x，重力方向是 $\vec{xz} = -c$. 作三角形的内切圆和外接圆，从中心点 x 出发沿重力方向交内切圆和外接圆分

别为 w 和 z. 由于外接圆包含三角形而三角形又包含内切圆，我们就知道高度函数 $f(x) = c^T x$ 满足 $f(z) \le f(y) \le f(w)$.



f(x)-f(w)=0.5(f(x)-f(z))

由于内切圆半径是外接圆半径的一半，而且 $f(x)$ 是线性函数，所以我们有

$$f(x) - f(w) = \frac{1}{2}[f(x) - f(z)] \ge \frac{1}{2}[f(x) - f(y)]\,.$$

也就是说，通过一次迭代（从 x 到 w ）我们就可以让目标函数到最优函数值的距离缩小一半，这样就很容易得到多项式复杂度（与精度有关）的算法。当然，真正的 Karmarkar 算法没有那么简单，我们只是用这个例子来说明它的基本思想。这给我们另一个启迪是：了解问题的几何本质对构造高效的计算方法是非常有帮助的。

内点法在过去的二十多年一直是十分热门的研究方向。许多国际著名学者，如美国科学院院士、美国工程院院士、美国纽约大学 Courant 研究所计算科学系系主任 Wright 教授；Dangtzig 奖获得者、美国 Cornell 大学的 Todd 教授等都在内点法方面有深入的研究。可喜的是，许多华人学者，如美国斯坦福大学的叶荫宇教授、美国 Rice

大学的张寅教授、美国 Minnesota 大学的罗智泉教授、香港中文大学的张树中教授等在这个国际热门研究领域中也作出了突出的贡献。近年来新兴的优化方向如半定规划、锥优化等的主要求解方法都是内点法。

关于优化，著名的数学家 Euler 曾说过： "Für, da das Gewebe des Universums am vollkommensten und die Arbeit eines klügsten ist Schöpfers, nichts an findet im Universum statt, in dem irgendeine Richtlinie des Maximums oder des Minimums nicht erscheint" (由于宇宙组成是最完美也是最聪明造物主之产物，宇宙间万物都遵循某种最大或最小准则)。这实际上就是说优化无处不在。

Leonhard Euler(1707-1783)

事实上，在其他科学研究领域中许多问题也归结于优化问题。例如：力学中的最小重量，最大载重，结构最优等；金融中的最大利润，最小风险等；生命科学中的 DNA 序列，蛋白质折叠等；信息科学中模式识别，海量数据处理等；地学中的反演问题；交通中的时刻表安排，最短路程等本质上都是优化问题。

近年来倍受关注的一个问题是压缩感知 (Compressive Sensing) 。菲尔茨奖获得者陶哲轩, 美国科学院院士、斯坦福大学教授 Donoho 等人对该问题有深入的研究。压缩感知问题的实际背景是用尽可能少的存贮记录尽可能清晰的图像。从计算上看, 压缩感知问题是求解线性方程组的最少非零元素解, 它就是如下优化问题:

$$\min_{x\in\Re^n} \quad ||x||_0 \tag{8}$$

$$\text{s. t.} \quad Ax = b \tag{9}$$

其中 $||x||_0$ 是指向量 x 的非零元素的个数, $A \in \Re^{m\times n}$, $b \in \Re^m$, 而且 $m \ll n$. 由于 $||x||_0$ 非凸而且不连续, 问题 (8)-(9) 的求解是 NP 难的。所以人们转而求解 L_1 优化问题:

$$\min_{x\in\Re^n} \quad ||x||_1 \tag{10}$$

$$\text{s. t.} \quad Ax = b, \tag{11}$$

因为 L_1 优化问题 (10)-(11) 是一个凸优化, 有多项式算法。纯数学家关心的是在什么条件下 (8)-(9) 和 (10)-(11) 是等价的。而计算数学家更关注如何快速求解问题 (10)-(11).

生命科学中一个重要的问题是蛋白质折叠问题。由于蛋白质的结构决定它的功能, 因为知道它的空间结构是十分重要的。通常的一种做法是利用核磁共振技术测量出一些原子之间的距离, 然后利用这些距离确定所有原子在空间的位置。这个问题在数学上是一个距离几何问题: 求解 $x_k(k = 1, ..., n) \in \Re^3$ 使得

$$||x_i - x_j||_2 = d_{ij}, \quad (i, j) \in S. \tag{12}$$

这里 d_{ij} $((i, j) \in S)$ 是给定的距离, 而 S 是集合 $\{(i, j)|i = 1, ..., n; j = 1, ..., n\}$ 的一个子集合。距离几何问题可归结为求解非线性最小二乘

$$\min_{x_k\in\Re^3,k=1,...,n} \sum_{(i,j)\in S} \left(||x_i - x_j||_2 - d_{ij}||_2\right)^2$$

或者是非光滑优化

$$\min_{x_k\in\Re^3,k=1,\ldots,n}\sum_{(i,j)\in S}\mid\|x_i-x_j\|_2-d_{ij}\mid$$

的全局最小点。除了在生命科学中的应用，距离几何问题在无线通讯网络定位、图像识别等许多其它科学和工程领域中也有重要的应用。

从学科发展上，优化近年来也越来越受到国际学术界的重视，例如在 2007 年在苏黎世召开的国际工业与应用数学大会上， 27 个大会报告就有 5 个是关于优化的； 2006 年在西班牙召开的国际数学家大会上也有一个优化方面的一小时报告。

历史上，我国广大科技工作者在老一辈科学家华罗庚等的带领下，在优化及其应用方面做出了突出的贡献。当前，国家正处于劳动密集型经济向科技创新型经济转型时期，优化正处大有用武之地，恰逢大有作为之时。我们相信优化将在我国国民经济建设的各个方面发挥更大的作用。