从瞎子爬山到最优化方法

袁亚湘

黄山 天都峰

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、梯度值等），但不知道其它点的信息。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。从这个角度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。

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。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

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

的迭代法：

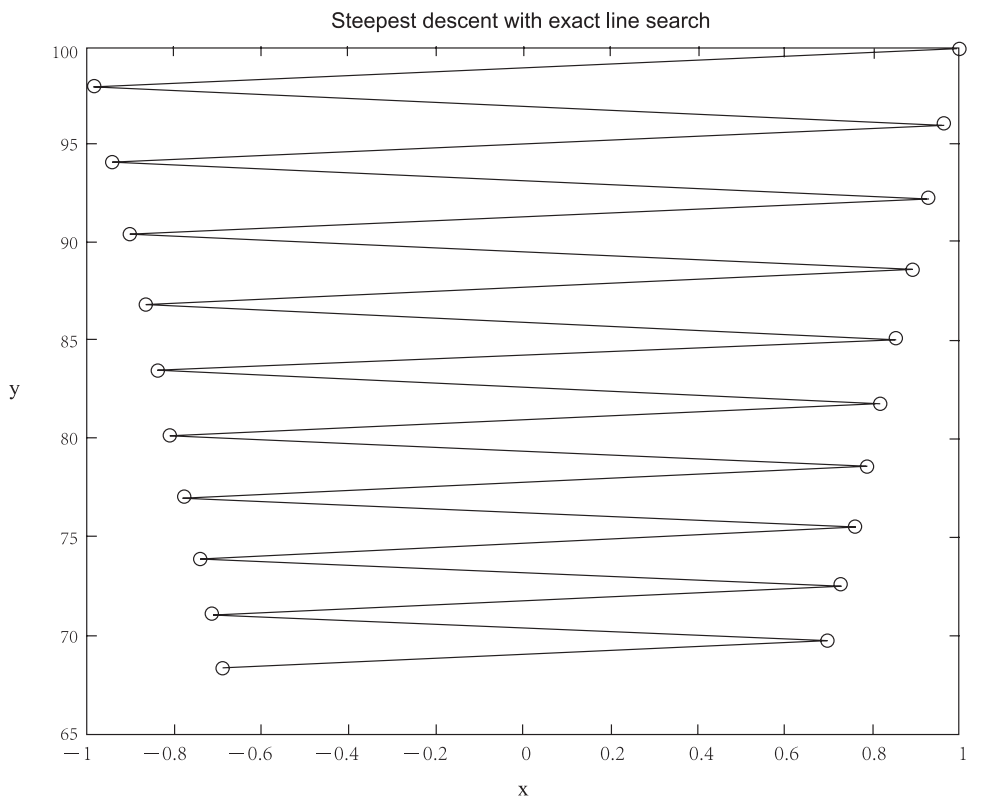
$$x_{k+1} = x_k + \alpha_k(-\nabla f(x_k)),$$

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：

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

这就是精确搜索下的梯度法，通常称为最速下降法。

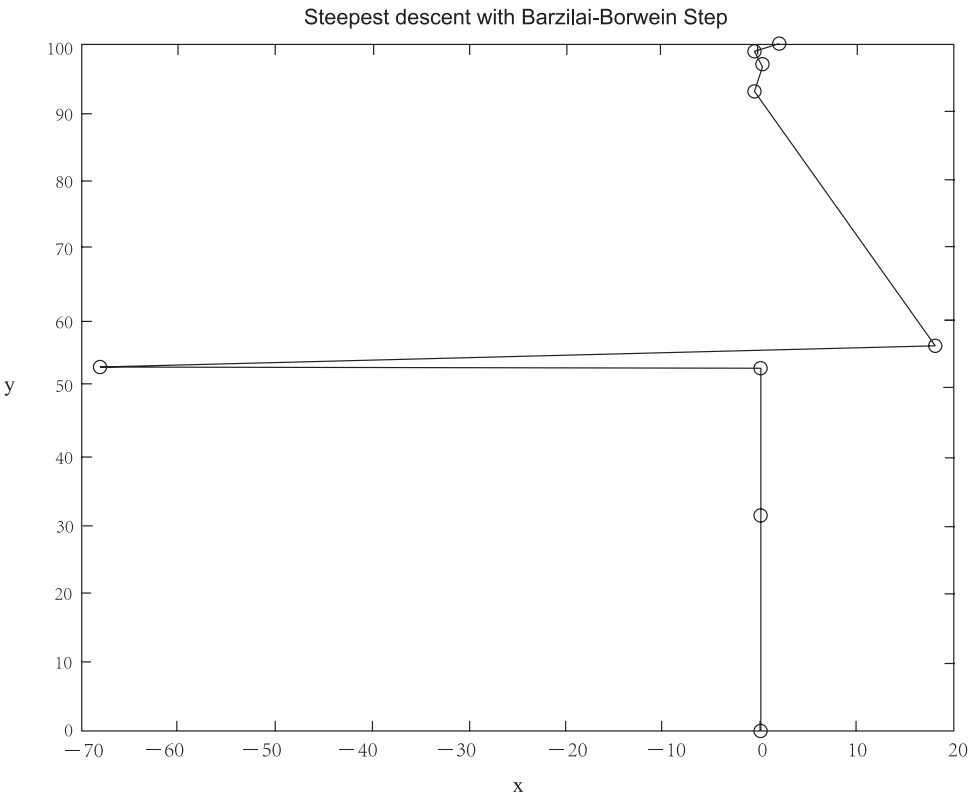
表面上看来，最速下降法是个完美的方法。该方法所用的方向是最好的（使函数降得最快），步长也是最好的（让函数在搜索方向上最小）。但是，最速下降法不仅不是一个最好的方法，反倒是一个很差的方法。下图是用最速下降法求解 $\min f(x, y) = 100x^2 + y^2$，从初始点 (1, 100) 开始迭代的前二十个迭代点：



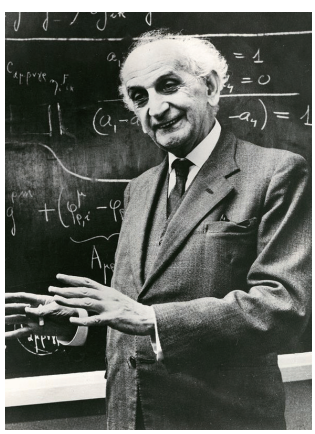
从上图可以看出，最速下降法收敛非常慢。也就是说，“最好”＋“最好”≠“最好”。我在中科院研究生院上课常常跟同学们开玩笑说，班上最好的男生娶班上最好的女生，结果往往不是最好的。

1988 年加拿大数学会前会长、加拿大皇家科学院院士 Jonathan Borwein 教授和合作者 Jonathan 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，几乎可以达到与后面将提到的共轭梯度法差不多的效果。下图是用 Barzilai-Borwein 方法（简称 BB 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。



Cronelius Lanczos（1893-1974）

Magnus Hestenes（1906-1991）

Eduard Stiefel（1909-1978）

优化方法中另外一个应用广泛的方法是共轭梯度法。该方法是用来求解线性方程组的，由著名数学家 Cornelius Lanczos，Magnus Hestenes 和 Eduard Stiefel 等在 1950 年代提出。

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：

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

显然，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 Michael 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。试探点往往要求是最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非线性最小二乘问题 $\min\|F(x)\|_2^2$ 的 Levenberg-Marquardt 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。Gauss-Newton 法是一个很“值钱”的方法，因为 Carl Friedrich Gauss 和 Issac Newton 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