# 梯度下降方法

最速下降法为最早最简单，也是最常用的优化方法，该方法于1847年由法国数学家Cauchy提出。

1. **引言**

考虑无约束优化模型



其中为连续可微函数。记为最优解，为最优解。

在处理这类问题时，我们总希望从某一点出发，选择一个目标函数值下降最快的方向，以利于尽快达到极小点。正是基于这样一种愿望，最速下降法被提出，最速下降法也称梯度下降法。

1. **最速下降法的思想：在当前位置沿负梯度方向进行搜索。**

因为目标函数的负梯度方向为当前位置的最快下降方向，因此该算法被称为是**“最速下降法”**。

最速下降法的迭代公式为：



其中分别表示当前位置和下一位置；为迭代步长；为当前位置的梯度。

1. **算法描述**

Step 0. 选取初始点，允许误差，令k=1；

Step 1.计算.若,算法停止，输出作为近似最优解.否则，转Step 2.

Step 2. 取搜索方向，利用线搜索技术确定步长.

Step 3. 令，，转step 1.

1. **几何解释：**

算法旨在追求每步求解较易问题，考虑函数在的近似二次表达



其中是在处的一阶线性近似，而是加权邻近项。

显然是二次强凸函数，易求有唯一的最优解且满足

即



解得，这正是梯度下降算法的迭代公式。

1. **收敛性分析**

梯度下降方法的快慢取决于步长，而步长有多种取法，既可以使用**固定步长，也可以使用精确线搜索步长，还可以使用非精确线搜索步长，在理论上都能够保证算法的全局收敛性**。

1）**固定步长**（Fixed Step Size）

**定理：假设f为可微凸函数，且梯度Lipschitz连续，即**

**其中，为Lipschitz常数。对于固定步长的梯度下降法，产生的点列满足**.

**我们称梯度下降方法的收敛速率为，即要使，我们需要进行步迭代。**

证明：根据假设是梯度Lipschitz连续，即

从而，由泰勒展开式近似可得

利用梯度下降迭代公式，取代入上式

取步长，可知

同时，由f的凸性可得

结合上述条件，可得

即

注意到单调不增数列，所以

从而，定理得证。

**2）精确线推导（Exact line search）**

一般地，上述优化问题不易精确求解。近似求解的效果经常不如回溯法有效。特别地，精确线搜索的梯度下降法具有“zigzag”（锯齿）现象。现在我们考虑二次凸规则问题，取目标函数



其中，是对称正定矩阵。显然，梯度和最优解是唯一的且为线性方程组的解。

**定理2（二次函数规划收敛性）：**对于强凸二次规划问题,其中是对称正定矩阵。最优步长下的最速下降法线性收敛，且产生的点列收敛到唯一的最优解，且满足

 (\*)

等价于

 (\*\*)

其中为Q的特征值，定义为



**证明：**

首先易知，对于二次函数



有梯度

最优解满足，即

从而

这证明（\*）式与（\*\*）式等价

下面根据迭代方法求最优步长

计算得满足

即

从而

下面开始证明（\*\*）式，

=

=

=

=

=

=

=

其中

根据*Kantorovich*不等式，对于任意，成立

其中和分别为正定矩阵的最小和最大特征值。

从而，根据前面推导有

显然，从而收敛至，即收敛到唯一最优解。

上述收敛性定理证明**以线性速率收敛到**。此推导最早由Luenberger[195-NO]证明，作为特殊情况，如果的所有特征值都相等，算法可以一步收敛到最优解。一般来说，矩阵的条件数越大，算法收敛速度越慢，出现所谓zigzag现象。

最优步长规则下的最速下降法出现zigzag现象，这与相邻两点的搜索方向正交有关，即。

证明：

易知

即

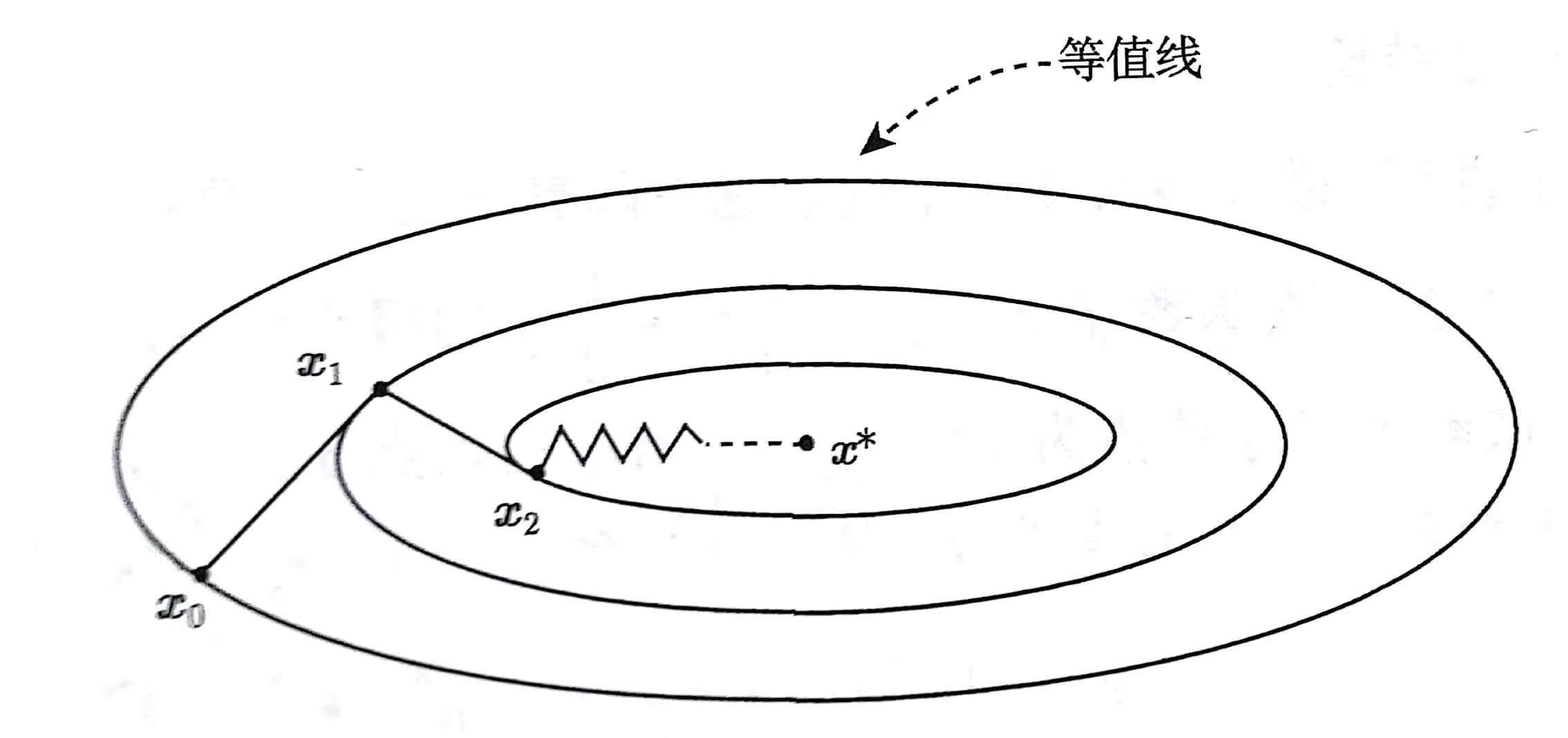


图 最速下降法的迭代过程

**3）回溯法（Backtracking）**

首先取定参数和，在每次迭代中，取，取最小的正整数，故得

在实际中这种选取步长的方法简单而较有效，一般地可简单取。

在和前面同样的假设条件下，即为连续可微的函数，且梯度Lipschitz连续，有以下收敛结果。

**定理（收敛速率）回溯法选择步长的梯度下降方法产生的数列满足**

其中。

注意，如果不太小，这种方法和固定步长类似。

证明：方法和固定步长类似。注意到此时。

**4）强凸情形的收敛性分析**

假设函数是强凸，即存在使得

易知，有下列不等式成立

此不等式可用来给出的界。

1. 事实上，固定，不难得出

最小化左边的二次函数，从而

由于对于所有成立，从而

上述不等式表明如果在一点的梯度很小，则此值接近最优解。此式还可用来作为次优的条件

另外，我们可以选定的界，其中为最优解。

事实上，由强凸定义，有

由于，所以

从而

这也证明了最优解是唯一的。

1. 由强凸定义，可以证明函数在给定点的次水平集是有界的，即

因此，的最大特征值在上是有上界的。

即存在常数使得

从而

两边关于求最小，得到

**在和前面相同条件下，强凸性可得下列收敛性条件（更快的收敛速率）**

**定理：梯度下降方法，取精确步长，或固定步长，或回溯法搜索确定步长，产生的点列满足**

（\*）

其中.

这里，收敛速率为（指数级），即要达到需要步迭代。

证明：

事实上，根据强凸性，只需证明

即可证明（\*）式。

**(i)精确线搜索确定步长的情形**

设，则

=

两边关于求最小，得

两边减去，利用迭代公式得

由于强凸，

代入上式

从而由递推关系可得

其中. 这说明.

特别的，为使，需要至多迭代步数为

关于迭代步数的这个界可以给出梯度方法的如下解释：分子可以解释为关于初始次优和最终次优的对数比率。这一项表明迭代次数依赖于初始值的选取和最终精度的要求。

分母是的函数，而是的条件数的一个界，或者在子水平集的条件数的界。对于较大的条件数，有

因此迭代步数的界的增长近似线性依赖.

关系式，表明误差至少以几何级数速率收敛到。在数值计算方法中，我们称之为线性收敛（Linear convergence）。

**(ii)回溯法确定步长的情形**

由回溯法知步长*t*满足，则

事实上，得到

再者，的凸性和梯度Lipschitz性质，对于

（）

由于回溯法线搜索步长选取或的值，这就为目标函数的下降提供了一个下界，当时

而时

合并得

下面过程与精确线搜索(i)相同，即

其中.

1. **总结**

最速下降方向具有计算简单，存储量小的优点。从局部看，最速下降方向确是函数值下降最快的方向，选择这样的方向进行搜索是有利的；但从全局看，在远离极小值点处，每次迭代能够使目标函数有较大的下降，但越接近极小值点，由于锯齿现象的影响，算法收敛速率显著减慢。

# 

# 牛顿算法

1. 模型

考虑无约束优化模型



假设二阶连续光滑且Hessian阵可逆。

1. **思想：**

**牛顿法用一个二次函数去近似目标函数**，**然后精确地求出这个二次函数的极小点.**

1. 算法描述**迭代公式**：



其中**牛顿方向。**

考虑函数在点的二次泰勒近似



由于可逆，则的一阶稳定点满足，即



由上式可得



这正是**牛顿法的迭代公式**。由此说明，**牛顿法在每次迭代中考虑函数在当前的局部二次泰勒展开，其迭代步长为1，迭代方向是梯度方向经过Hessian逆阵的调整**。如果函数为凸函数，此时正定，从而，即牛顿方向是下降方向。

1. **算法步骤**

Step 0. 选取初始点，允许误差，令；

Step 1. 计算，.若,算法停止，输出作为近似最优解；否则，转Step 2。

Step 2. 计算.令，转Step 1.

1. **收敛性分析**

这里讨论牛顿算法的**局部收敛性质**，这是由于Hessian阵可能不总是正定，即牛顿方向可能不总是下降方向。我们假设目标函数为凸函数，二阶连续可微，最优解处正定。此时，在附近的点，其也是正定的。从而牛顿方法是有定义的，并且在迭代步长最终取1的情况下具有二次收敛性。

**定理：假设函数****二阶可微，其Hessian阵****在最优解****的邻域是Lipschitz连续，且在****处满足充分条件**且**正定。考虑迭代公式**，**则下面结论成立：**

**（i）如果初始点充分接近解点****,算法产生的点列收敛到****；**

**（ii）点列收敛速率是二次的；**

**（iii）梯度模长的点列二次收敛到0.**

证明：根据迭代公式和最优性条件,我们有



注意到由泰勒中值定理，



结合上式，我们可以得出



其中L是在点处的Lipschitz常数。由于非奇异，故存在常数，对于任意满足，有。从而结合上式，有



其中。选择初始点使得。我们用此不等式通过递推得到数列收敛到，且为二次收敛。

根据关系式和，我们有



即得****收敛到0且二阶收敛。

1. **用Newton法求解无约束问题会出现以下情形：**

（1）收敛到极小点；

（2）收敛到鞍点；

（3）Hesse矩阵不可逆，无法迭代下去。

1. **评价**

**1）优点:**

①Newton法产生的点列若收敛，则收敛速度快---具有至少二阶收敛速率；

②Newton法具有二次终止性，即对强凸函数Newton法只需迭代一次就能得到最优值。

**证明：**设A为对称正定矩阵，且



令得.从任一点出发，由迭代公式可得



**2）缺点**

①可能会出现在某步迭代时，目标函数值上升；

②当初始点远离极小点时，牛顿法产生的点列可能不收敛，或者收敛到鞍点，或者Hesse矩阵不可逆，无法计算；

③需要计算Hesse矩阵，计算量大.

1. **阻尼牛顿法**

*Newton*法的有效性严重依赖初始点的选择，即初始点需要充分靠近极小值点，否则可能导致算法不收敛。由于实际问题的精确最小值点一般是不知道的，因此初始点的选择给算法的实现带来了很大的挑战。为了解决这一问题，可引入线搜索技术以得到大范围的收敛算法，即阻尼牛顿法。

Step 0. 选取初始点，允许误差，令；

Step 1. 计算，.若,算法停止，输出作为近似最优解；否则，转Step 2.

Step 2. 取搜索方向，利用线搜索技术确定步长

Step 3. 令.令，转Step 1.