# 一阶优化算法：回溯线搜索

北极甜虾 (南半球版)

2026 年 1 月 25 日

Manifold Optimization Notes

回溯线搜索 | 算法的理论保证 | 启发式步长选择


## 1 回溯线搜索和 Armijo 条件

在**黎曼梯度下降 (RGD)** 中，迭代公式为

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

我们需要确定步长 $\alpha_k$. 除固定步长外，确定步长的方法还有:

- **精确线搜索 (Exact Line Search):** 寻找使目标函数沿搜索方向下降最多的步长的计算开销过高，不切实际。
- **非精确线搜索 (Inexact Line Search):** 我们只需要找到一个“足够好”的步长，既能保证目标函数值下降，又能保证步长不会太小导致算法停滞。

我们之前笔记中提到的**回溯线搜索（Backtracking Line Search)** 是一种标准的非精确搜索策略, 同时平衡计算效率与收敛速度，这次来系统地学习一下这个方法。

**定义 1: 回溯线搜索（Backtracking Line Search）**

给定参数 $\bar{\alpha} > 0$（初始步长），$\tau \in (0,1)$（缩减因子）和 $r \in (0,1)$（下降常数）。我们寻找最小的非负整数 $m$，使得 $\alpha_k = \bar{\alpha}\tau^m$ 满足 **Armijo-Goldstein 条件** (aka. **Armijo 条件**):

$$f(R_{x_k}(-\alpha_k \operatorname{grad} f(x_k))) - f(x_k) \leqslant -r\alpha_k \| \operatorname{grad} f(x_k)\|_{x_k}^2$$

其中，右侧的项中代表了我们期望目标函数的最少下降量。

**算法：回溯线搜索 (Backtracking Line Search)**

1. **输入**：目标函数 $f : \mathcal{M} \to \mathbb{R}$ 与收缩映射 $R$，当前迭代点 $x \in \mathcal{M}$，搜索方向 $\eta \in T_x\mathcal{M}$ （通常取 $\eta = -\operatorname{grad} f(x)$）
2. **参数**：初始步长 $\bar{\alpha} > 0$, 缩减因子 $\tau \in (0,1)$ （例如 0.5）, 下降常数 $r \in (0,1)$ （例如 $10^{-4}$）
3. **初始化**：令当前步长 $\alpha = \bar{\alpha}$.
4. **循环**：当 Armijo 条件**不满足**时，即

   $$f(R_x(\alpha\eta)) - f(x) > r\alpha \| \operatorname{grad} f(x)\|_x^2$$

   更新:

   $$\alpha \leftarrow \tau\alpha$$

5. **输出**：满足条件的步长 $\alpha$.

## 2 算法的理论保证

**假设 2: Lipschitz 型正则性条件**

如果对于切丛 $T\mathcal{M}$ 的子集 $S$，存在常数 $L > 0$，使得对于所有 $(x, s) \in S$，满足：

$$f(R_x(s)) \leqslant f(x) + \langle \operatorname{grad} f(x), s\rangle_x + \frac{L}{2}\|s\|_x^2.$$

**定理 3: 线搜索的有限终止**

如果假设 2 对于 $S \supset \{(x, -\alpha \operatorname{grad} f(x)) \in T\mathcal{M} : \alpha \in [0, \bar{\alpha}]\}$ 满足，那么回溯线搜索会在最多

$$n = \max\left\{1, 2 + \log_{\tau^{-1}} \frac{\bar{\alpha} L}{2\tau(1-r)}\right\}$$

次迭代内终止，并且返回的步长满足：

$$\alpha \geqslant \min\left\{\bar{\alpha}, \frac{2\tau(1-r)}{L}\right\} > 0.$$

证明. 根据光滑函数在流形上的泰勒展开，对于曲线 $c(t) = R_{x_k}(-t \operatorname{grad} f(x_k))$，存在常数 $L > 0$，使得对于足够小的 $t \geqslant 0$：

$$f(c(t)) \leqslant f(x_k) - t\|\operatorname{grad} f(x_k)\|_{x_k}^2 + \frac{L}{2}t^2\|\operatorname{grad} f(x_k)\|_{x_k}^2.$$

我们要寻找满足 Armijo 条件的步长 $\alpha$，即要求

$$f(c(\alpha)) \leqslant f(x_k) - r\alpha\|\operatorname{grad} f(x_k)\|_{x_k}^2.$$

将泰勒展开的上界代入，只需满足：

$$-\alpha\|\operatorname{grad} f(x_k)\|_{x_k}^2 + \frac{L}{2}\alpha^2\|\operatorname{grad} f(x_k)\|_{x_k}^2 \leqslant -r\alpha\|\operatorname{grad} f(x_k)\|_{x_k}^2.$$

消去 $\alpha\|\operatorname{grad} f(x_k)\|_{x_k}^2$，不等式化简为：

$$-1+\frac{L}{2}\alpha \leqslant -r \iff \alpha \leqslant \frac{2(1-r)}{L}.$$

这表明，一旦回溯过程中的试探步长 $\alpha$ 减小到 $\frac{2(1-r)}{L}$ 以下，Armijo 条件即成立。由于每次迭代步长乘以因子 $\tau \in (0,1)$，步长序列是一个几何级数，必然在有限步内落入该区间或直接以初始步长满足条件。当在某一步落入该区间时，说明上一步 $\alpha' > \frac{2(1-r)}{L}$，因此 $\alpha = \tau\alpha' > \frac{2\tau(1-r)}{L}$. 所以饭回的步长 $\alpha$ 满足

$$\alpha \geqslant \min\left\{\bar{\alpha}, \frac{2\tau(1-r)}{L}\right\} > 0. \tag{1}$$

注意到返回的 $\alpha = \bar{\alpha}\tau^{n-1}$，$n$ 是计算收缩映射和目标函数的次数，经过计算有

$$n = 1 + \log_{\tau^{-1}} \frac{\bar{\alpha}}{\alpha} \leqslant 1 + \max\left\{1, 1 + \log_{\tau^{-1}} \frac{\bar{\alpha}L}{2\tau(1-r)}\right\},$$

其中最后一步带入了不等式 (1). □

**定理 4: 带回溯线搜索的黎曼梯度下降的收敛定理**

如果下列条件满足：

- 存在 $f_{low} \in \mathbb{R}$ 使得 $f(x) \geqslant f_{low}$, $\forall x \in \mathcal{M}$.
- Lipschitz 型正则性条件 (假设 2)
- 回溯线搜索的参数 $\tau, r \in (0,1)$，初始步长设置为 $\bar{\alpha}_0, \bar{\alpha}_1, \bar{\alpha}_2, \ldots$，并满足
  - $\forall k \in \mathbb{N}, \{(x_k, -\alpha \operatorname{grad} f(x_k)) \in T\mathcal{M} : \alpha \in [0, \bar{\alpha}_k]\} \subset S$
  - $\liminf_{k\to\infty} \bar{\alpha}_k > 0$.

那么对于任意整数 $K \geqslant 1$，黎曼梯度下降产生的序列满足

$$\min_{0 \leqslant k \leqslant K-1} \| \operatorname{grad} f(x_k) \|_{x_k} \leqslant \sqrt{\frac{f(x_0) - f_{\text{low}}}{rCK}},$$

其中

$$C = \min \left\{ \overline{\alpha}_0, \ldots, \overline{\alpha}_{K-1}, \frac{2\tau(1-r)}{L} \right\}.$$

证明. 由回溯线搜索的终止条件（Armijo 条件）可知，每一步目标函数的下降量满足：

$$f(x_k) - f(x_{k+1}) \geqslant r\alpha_k \| \operatorname{grad} f(x_k) \|_{x_k}^2.$$

根据定理 3 的结论，步长 $\alpha_k$ 有下界。结合正则性假设，存在一个常数 $c_k = r \min\{\overline{\alpha}_k, \frac{2\tau(1-r)}{L}\} > 0$，使得充分下降条件成立：

$$f(x_k) - f(x_{k+1}) \geqslant c_k \| \operatorname{grad} f(x_k) \|_{x_k}^2.$$

对上述不等式从 $k = 0$ 到 $K-1$ 求和，得到：

$$\sum_{k=0}^{K-1} c_k \| \operatorname{grad} f(x_k) \|_{x_k}^2 \leqslant \sum_{k=0}^{K-1} (f(x_k) - f(x_{k+1})) = f(x_0) - f(x_K).$$

由于 $f$ 有下界 $f_{\text{low}}$，则 $f(x_0) - f(x_{K+1}) \leqslant f(x_0) - f_{\text{low}}$. 并且注意到左侧大于等于 $K \cdot \min_k c_k \min_k \| \operatorname{grad} f(x_k) \|_{x_k}^2$. 整理后得证。 □

## 3 仿射不变性与启发式步长

在实际优化中，如果我们将目标函数 $f(x)$ 替换为其平移或正缩放的版本，例如 $g(x) = cf(x) + d$（其中 $c > 0$），优化问

题的本质并没有改变。因此，一个合理的优化算法在 $x_0$ 处初始化后，无论是最小化 $f$ 还是 $g$，都应当产生相同的迭代序列。

黎曼梯度下降结合回溯线搜索具备这种**仿射不变性**，前提是初始试探步长 $\bar{\alpha}_k$ 的选择必须使得步长向量 $-\bar{\alpha}_k \operatorname{grad} f(x_k)$ 在 $f$ 的正缩放变换下保持不变。由于 $\operatorname{grad}(cf) = c \operatorname{grad} f$，这意味着 $\bar{\alpha}_k$ 必须与 $f$ 的缩放成反比。以下是一种启发式方法：

**初始步长 $\bar{\alpha}_k$ 的启发式选择**

1. **首次迭代 ($k = 0$)：**

$$\bar{\alpha}_0 = \frac{\ell_0}{\| \operatorname{grad} f(x_0) \|_{x_0}}$$

其中 $\ell_0$ 是某个常数，可根据搜索空间的尺度或预期到最优解的距离来设定（无需非常精确）。

2. **后续迭代 ($k > 0$)：** 利用基于二次插值的启发式公式（参考 [Noc06] 的 Section 3.5, Eq. (3.60)）：

$$\bar{\alpha}_k = 2 \frac{f(x_{k-1}) - f(x_k)}{\| \operatorname{grad} f(x_k) \|_{x_k}^2} \tag{2}$$

**仿射不变性验证：** 若 $f$ 变为 $c \cdot f$，则分子变为 $c$ 倍，分母（梯度范数平方）变为 $c^2$ 倍，因此 $\bar{\alpha}_k$ 变为 $1/c$ 倍。这正好抵消了梯度的缩放，使得步长向量 $-\bar{\alpha}_k \operatorname{grad} f(x_k)$ 保持不变。

**工程实现细节** 在实际代码实现中，通常会结合以下技巧：

- **略微放大：** 通常将公式 (2) 计算出的值稍微放大（例如乘以 $1/\tau$），以便线搜索从一个稍大的步长开始尝试。

- **下界保护：**为了满足全局收敛性的要求（$\liminf \bar{\alpha}_k > 0$），应将 $\bar{\alpha}_k$ 设为上述启发式计算值与某个微小参考值之间的最大值，以确保初始步长不会趋近于零。

**公式 (2) 的推导**　该公式的核心思想是：**假设函数在搜索方向上局部表现为抛物线，并预期本次下降量与上一次相当。**

在点 $x_k$ 处沿方向 $\eta_k = -\operatorname{grad} f(x_k)$ 的拉回映射为 $\phi(\alpha) = f(R_{x_k}(\alpha\eta_k))$。用二次函数 $q(\alpha) = a\alpha^2 + b\alpha + c$ 近似 $\phi(\alpha)$：

1. **利用当前点信息确定参数：**
   - $q(0) = \phi(0) = f(x_k) \implies c = f(x_k)$.
   - $q'(0) = \phi'(0) = \langle \operatorname{grad} f(x_k), \eta_k \rangle_{x_k} = -\| \operatorname{grad} f(x_k) \|_{x_k}^2 \implies b = -\| \operatorname{grad} f(x_k) \|_{x_k}^2$.

2. **引入下降量一致性假设：**假设二次模型 $q(\alpha)$ 在其极小值点 $\alpha^*$ 处达到的下降量，等于上一次迭代实际观测到的下降量 $\Delta f_{k-1} = f(x_{k-1}) - f(x_k)$：

$$q(0) - q(\alpha^*) = f(x_{k-1}) - f(x_k).$$

对于二次函数 $q(\alpha)$，其顶点处的下降量为 $\frac{b^2}{4a}$，极小值点为 $\alpha^* = -\frac{b}{2a}$。由 $\frac{b^2}{4a} = \Delta f_{k-1}$ 得 $2a = \frac{b^2}{2\Delta f_{k-1}}$。代入 $\alpha^*$ 得

$$\bar{\alpha}_k = \alpha^* = \frac{-b}{2a} = \frac{-b}{b^2/(2\Delta f_{k-1})} = \frac{2\Delta f_{k-1}}{-b} = \frac{2(f(x_{k-1}) - f(x_k))}{\| \operatorname{grad} f(x_k) \|_{x_k}^2}.$$

**注解 5: 隐式曲率**

尽管黎曼梯度下降是一阶算法，但公式 (2) 通过利用函数值的历史差值，隐式地捕获了二阶信息（曲率）。在曲率较大的区域，步长会自动减小，从而提高回溯搜索的效率。

# 参考文献

[Bou23] Nicolas Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[Noc06] Jorge Nocedal. Numerical optimization, 2006.