Deep Learning Architectures

A Mathematical Approach

作者：康豪
组织：SMS, UESTC
时间：Aug 31, 2023
版本：1.0
类型：学习笔记

Live long and prosper! ——Vulcans

目录

前言

笔记的 LaTeX 模板来自 *ElegantLatex* 团队编写的作品 *elegantbook*，该系列风格优雅、功能齐全，被各类讲义、笔记编著者广泛采纳，实为佳作。

封面图片为船底座大星云，由韦伯望远镜拍摄，图片来自 NASA 官网。

第 1 章 引言

第 2 章 激活函数

第 3 章 损失函数

第 4 章 优化理论与算法

监督学习中的学习过程包括调整网络参数（权值和偏差），直到一定的代价函数被最小化。由于参数的数量相当大（它们很容易成数千），需要一个鲁棒的最小化算法. 本章介绍了许多不同口味的最小化算法，并强调了它们的优缺点.

4.1 最小值的一般性质

本节回顾关于具有一个实变量或几个实变量的最小函数值的基本概念. 这些理论上可行的技术只有在变量的数量不是太大的情况下，在实践中才是有效的。然而，在机器学习中，变量的数量是成千上万或更多的，所以这些经典的寻找最小值的理论方法对于这些应用来说并不是有利可图的。我们在这里包含它们只是为了完整性，并有一个构建未来更复杂的方法的基础.

4.1.1 单变量实值函数

设 $f:[a,b]\to\mathbb{R}$, 是定义在紧区间 $[a,b]$ 上的一个连续函数，在这里紧区间就是有界闭区间。根据微积分的知识，有界闭区间上的连续函数必有界，且一定存在最大最小值. 也就是说，存在 $c\in[a,b]$，使得 $f(c)=\min_{x\in[a,b]}f(x)$. 这个点被称为 f 的全局最小值点，当然 f 可能也有一些局部极小值点. 此外，根据 Fermat 引理，若 c 是函数 f 的极小值点，且 f 在 c 处可导，那么 $f'(c)=0$. 注意，这是一个必要但不充分条件，当补充条件 $f''(x)\geq 0)$ 时，上述条件就变成了充要条件.

4.1.2 多变量实值函数

设 K 是 $\mathbb{R}^n$ 上的紧集，即 K 是一个有界闭集，与单变量实值函数类似，当 K 中有满足条件的全局或局部极值点 c 时，必然满足

$$\frac{\partial f}{\partial x_i}(c)=0,\quad i=1,\ldots,n. \tag{4.1}$$

上式可以等价地描述为 $\nabla f(c)=0$，其中 $\nabla f=\sum_{i=1}^{n}\frac{\partial f}{\partial x_i}\tilde{e}_i$. 系统4.1等价于曲面 z = f (x) 处（c，f (c)）的切平面是水平的，即平行于 x 超平面.

在 $x=c$ 的邻域中，$f(x)$ 的二阶泰勒近似给出如下:

$$\begin{aligned}f(x)&=f(c)+\sum_i\frac{\partial f}{\partial x_i}(c)(x_i-c_i)\\&\quad+\frac{1}{2}\sum_{i,j}\frac{\partial^2 f}{\partial x_j\partial x_k}(c)(x_j-c_j)(x_k-c_k)+o(\|x-c\|^2)\\&=f(c)+(x-c)^T\nabla f(c)+\frac{1}{2}(x-c)^T H_f(c)(x-c)+o(\|x-c\|^2),\end{aligned}$$

其中，$H_f=\frac{\partial^2 f}{\partial x_j\partial x_k}$，是 f 的 Hessian 矩阵.

命题 4.1

设 c 是 $\nabla f(c)=0$ 的解，假设 H_f 在 c 的邻域上是正定的。那么 c 是 f 的局部最小值点.

证明 设 $x(t)$ 为具有 $x(0)=c$ 的 f 的结构域上的任意固定曲线，并考虑复合函数 $g(t)=f(x(t))$. 为了证明 c 是 $f(x)$ 的局部最小值，等价于证明 $t=0$ 是 $g(t)$ 的局部最小值，对于任何曲线 $x(t)$ 的局部最小值。

例题 4.1 二维空间中的正定 Hessian 矩阵 考虑一个在 $\mathcal{R}^2$ 上具有连续导数的二次可微函数 f, 它的 Hessian 矩阵

是一个 2×2 矩阵:

$$H_f(x,y) = \begin{pmatrix} f_{xx} & f_{xy} \\ f_{yx} & f_{yy} \end{pmatrix},$$

其中，我们使用符号 $f_{xx} = \frac{\partial^2 f}{\partial x^2}$ 来表示 x 中的二次偏导数. 对于任何向量，$u = (ab)$ 提供了二次形式

$$u^T H_f u = f_{xx}a^2 + 2f_{xy}ab + f_{yy}b^2,$$

改写一下形式，凑出完全平方后得到：

$$\begin{aligned} u^T H_f u = & \quad f_{xx}\left(a + \frac{f_{xy}}{f_{xx}}b\right)^2 + \frac{b^2}{f_{xx}}(f_{xx}f_{yy} - f_{xy}^2) \\ = & \quad f_{xx}\left(a + \frac{f_{xy}}{f_{xx}}b\right)^2 + \frac{b^2}{f_{xx}} \det H_f. \end{aligned}$$

根据之前的计算，我们可以得到:

- 如果 $f_{xx} > 0$ 和 $detH_f > 0$，那么对于任何 $u \in \mathbb{R}^2$，$u^t H_f u > 0$，即 Hessian 矩阵 H_f 是正定的.
- 如果 $f_{xx} < 0$ 和 $detH_f > 0$，那么对于任何 $u \in \mathbb{R}^2$，$u^t H_f u < 0$，即 Hessian 矩阵 H_f 是负定的.

例题 4.2 二次函数 设 A 是一个对称的、正定的、非退化的 $n \times n$ 矩阵，并考虑以下 n 个变量的实值二次函数

$$f(x) = x^T Ax - 2b^T x + d, \quad x \in \mathbb{R}^n,$$

其中 $b \in \mathbb{R}^n$，$d \in R$，其中 $x^T Ax = \sum_{i,j} a_{ij}x_i x_j$ 和 $b^T x = \sum_k b_k x_k$.。我们有梯度 $\nabla f(x) = 2Ax - 2b$ 和 Hessian 矩阵 $H_f = 2A$，所以这个函数是凸的。$\nabla f(x) = 0$ 的解，即 $c = A^{-1}b$，是 f 的局部最小值。由于该解决方案是唯一的，因此它实际上是一个全局最小值。A 的可逆性是由行列式非零的 Hessian 矩阵的性质得出的.

如果前面的二次函数 $f(x)$ 仅定义在一个紧集 K 上，有时解 $c = a^1 b$ 可能不属于 K，因此我们不能通过求解相关的欧拉系统得到 f 的所有最小值。另一方面，我们知道 f 在紧致域 K 上达到了一个全局最小值。因此，这个最小值必须属于 K 的边界，而找到它需要对最小值进行边界搜索.

例题 4.3 谐函数 如果 $\Delta_x f = 0, \forall x \in D,$，则一个函数 $f(x)$ 在结构域 $D \subset \mathbb{R}^n$ 上被称为谐波，其中 $\Delta_x f = \sum_i \frac{\partial^2 f(x)}{\partial x_i^2}$ 是 f 的拉普拉斯算子. 拉普拉斯的最小（最大）性质表明一个调和函数在域 d 的边界上达到最小（最大值）. 换句话说，一个调和函数的极值总是在域的边界上达到.

由于任何仿射函数 $f(x) = b^T x + db \in R^n d \in R$ 都是调和的，所以 f 的最小值（最大值）在 d 的边界上达到。如果 D 是一个凸多边形的内部，那么在多边形的顶点上达到最小值（极大值）. 寻找解顶点是单纯形算法的基本思想.

我们已经看到，通过求解系统 $\nabla f(x) = 0$，不能总是找到最小值. 即使这是可能的，这种寻找最小值的分析方法在实践中也并不总是可行的，因为涉及到大量的变量. 下面，我们将介绍一些用于使给定函数近似为最小值的鲁棒迭代方法.

4.2 梯度下降法

梯度下降算法是一种通过通过相关的水平集进入最大成本降低的方向来找到函数的最小值的过程。我们将首先介绍水平集的成分. 对令人生畏的细节不感兴趣的读者可以直接跳过到下一节.

4.2.1 水平集

考虑函数 $z = f(x), x \in D \subset \mathbb{R}^n$, 其中 $n \geq 2$，并定义集合 $\mathcal{S}_c = f^{-1}(\{c\}) = \{x \in \mathbb{R}^n; f(x) = c\}$. 假设函数 f 与非零梯度可微，$\nabla f(x) \neq 0, x \in \mathcal{S}_c$. 在此条件下，$S_c$ 成为 $\mathbb{R}^n$ 中的一个 $(n1)$ 维超曲面。族 $\{\mathcal{S}_c\}_c$ 称为函数 f 的水平超曲面. 当 $n = 2$ 时，它们被称为水平曲线。几何上，将 $z = f(x)$ 的图与水平平面 $z = c$ 相交，得到水平超曲面，见图 4.1.

命题 4.2

梯度 ∇f 垂直于水平集 S_c.

♠

证明 ...

在等价的表示中，对于任意的 $x \in S_c$，向量 $\nabla f(x)$ 正交于 S_c 在 x 处的切面 $T_x S_c$，见图 4.2. 超曲面的方向由使 ∇f 指向向外的方向来确定. 切平面 $T_x S_c$ 作为 S_c 内外 x 左右点的无穷小分隔符，见图 4.3 a

假设函数 $z = f(x)$ 在 $x^* \in D$ 处有一个 (局部) 最小值，即对 $x \in \mathcal{V} \backslash \{x^*\}$, 有 $f(x^*) < f(x)$，其中 v 是以 x^* 的邻域 (我们可以假设 V 是一个以 x 为中心的球). 用 $z = f(x)$ 表示 f 在 x 处的局部最小值. 然后存在 $\epsilon > 0$，使得对任何 $c \in [z^*, z^* + \epsilon)$, 有 $\mathcal{S}_c \subset \mathcal{V}$，见图 4.1. 对于 $c = z$，超曲面退化到一个点，$S_c = x$. 对于足够小的，族 $\{\mathcal{S}_c\}_{c \in [z^*, z^* + \epsilon)}$ 是嵌套的，即如果 $c_1 < c_2$，那么 $\mathcal{S}_{c_1} \subset \mathrm{Int}(\mathcal{S}_{c_2})$.

下一个结果说明了存在由 x 发出任意初始方向曲线，它们演化垂直于族 S_c，见图 4.3 b. 回想一下，如果一个函数 $\phi(x)$ 和常数 $K > 0$，使得 $|\phi(x) - \phi(y)| \leq K\|x - y\|$,，则称 ϕ 为 *Lipschitz* 连续. 下面的两个存在性结果将用到这个假设.

引理 4.1

假设 ∇f 是 Lipschitz 连续的。对于任何向量 $v \in \mathcal{R}^n$，存在 $\delta > 0$ 和一个可微曲线 $\alpha : [0, \delta) \to \mathbb{R}^n$，使:

- $\alpha(0) = x^*$;
- $\dot{\alpha}(0) = v$;
- $\dot{\alpha}(t)$ 垂直于 $\mathcal{S}_{f(\alpha(t))}$, 对所有的 $t \in [0, \delta)$.

♡

证明 ...

为了将来参考，当我们也表示初始方向 v 时，作为上述 ode 系统解的曲线将用 $\alpha_v(t)$ 表示。

值得注意的是，在重新参数化时，曲线是唯一的，即曲线可以改变速度，而其几何图像保持不变。

一个显著的参数化是在水平差参数 $\tau = cz$ 上。这是由于对于较小的 t 值，我们有 $\dot{c}(t) > 0$. 如果新参数化中的曲线用 $\beta(\tau)$ 表示，则 $\beta(0) = x = \alpha(0)$ 和 $\beta'(\tau) = \dot{\alpha}(t(\tau))t'(\tau)$，那么 $\beta'(0) = t'(0)v$。我们也有方便的关联关系 $\beta(\tau) = \beta(c - z^*) \in \mathcal{S}_c$，它也可以写成 $f(\beta(\tau)) = c$.

现在我们陈述并证明以下局部连通性结果，这将在稍后的梯度下降法中使用:

定理 4.1

假设 ∇f 是连续的。对于任何一个足够接近 x 的点 x_0，都有一个可微曲线 $\gamma : [0, \delta] \to \mathbb{R}^n$，使得

- $\gamma(0) = x^0$;
- $\gamma(\delta) = x^*$;
- $\dot{\gamma}(t)$ 垂直于 $\mathcal{S}_{f(\gamma(t))}$, 对所有的 $t \in [0, \delta)$.

♡

我们将提供一个非建设性的证据。我们先准备好一些符号. 设 $\alpha_v : [0, \delta_v) \to \mathbb{R}^n$ 为引理 4.2.2 提供的曲线。假设初始 subunitary 向量，$\|v\| \leq 1$。由于结束值 δ_v 关于 v 是连续的，所以它在单位球上达到其最小值，即 $\delta = \min_{\|v\| \leq 1} \delta_v$。现在用 $A_\delta = \{\alpha_\epsilon(t); t \in [0, \delta), \forall v \in \mathbb{R}^n, \|v\| \leq 1\}$. 表示. 集合 A_δ 表示由曲线 $\alpha_v(t)$ 交换的实心域，该曲线从 x^* 发出到所有方向 v，直到时间 δ.δ 值的选择使得 A_δ 的定义有意义. 集合 A_δ 是非空的，因为显然是 $x \in A_\delta$. 下一个结果表明，实际上 A_δ 包含一个以 x 为中心的非空球，见图 4.4 a. 这实际上是一个等价于定理 4.2.3 的陈述的结果.

定理 4.2

存在 $\epsilon > 0$，使得 $B(x^*, \epsilon) \subset A_\delta$.

♡

证明 我们将首先提供一个经验证明.$B(x^*, \epsilon) \subset A_\delta$ 意味着 x 是 A_δ 的一个内点. 如果，由反证法，我们假设 x 不是一个内点，那么有一个序列的点 $(x_k)_k$ 收敛到 x，使得 $x_k \notin A_\delta$. 这里是经验假设的地方：假设 x_k 位于一条光

滑的曲线 $x(s)$ 上，从 $x(0)=x$ 开始，满足 $x(s_k)=x_k s_k$ 为负数递减序列，见图 4.4 b. 设 $v^0=x'(0)$ 为曲线 $x(s)$ 接近于点 x 的方向. 引理 4.2.2 产生了一个从这个方向开始的曲线 α_{v0}，它将与一个邻域上的 $x(s)$ 重合. 这是由于曲线 $x(s)$ 和 α_{v0} 都具有相同的初始点和速度. 因为在这种情况下，对 $-s<\delta$ 我们会有 $x(s)=\alpha_{v^0}(-s)\in A_\delta$，这导致了一个矛盾.

我们有两点注记:

- 所有可以通过满足上述性质的曲线 γ 连接到 x 的点都构成了 x 的吸引盆地。这个定理可以等价地说，引力盆地包含一个以 x 为中心的球.
- 方向 $v^0=x'(0)$ 是一个退化的方向. 先前的证明表明不存在退化方向. 对这一事实的正式证明可以使用逆函数定理 (见附录中的定理 F.1) 来完成，如下所示.

证明 ...

证明 ...

先前的非构造证明提供了从 x_0 到 x 的最速下降曲线 γ 的存在性，它通常与水平集族 S_c 相交。然而，这种连续的结果在涉及到计算机实现时并没有什么用处. 为了实现曲线的构造，我们需要用一条多边形线 $\mathfrak{P}_m=[x^0x^1\dots x^m]$ 来近似曲线 γ，以满足以下属性:

- $x^k\in\mathcal{S}_{c_k}$;
- $c_{k+1}<c_k$, 对任意的 $k=0,\dots,m-1$;
- 线段 x_jx_{j+1} 正交于 S_{c_j}

构造算法如下图所示。我们从点 x_0 开始，沿着 S_{c_0} 的法线向内走 η，或者等价地，在点 x_0 的 $-\nabla f$ 方向. 因此，我们得到了点 $x_1\in S_{c_1}$. 我们继续，沿着 S_{c_1} 处的法线向向内的方向再走一段距离 η，得到点 x_2. 经过 m 步，我们得到点 x_m，我们希望它接近 x，因此它是一个很好的近似.

但是我们如何选择 m，或者换句话说，我们如何知道什么时候停止这个程序呢？只要 ck+1 < ck，即着陆超曲面嵌套。对于任何一个先验固定的 η > 0，都有一个最小的 m，具有性质 (i)-（iii）。这意味着当 xm+1 降落在一个超表面 Scm+1 上时，我们就会停止，该 +1 没有嵌套在之前的超表面 Scm 中。步长 η 越小，停止阶数 m 就越大，并且它所期望得到的越接近 x∗。当 η→ 为 0 时，多边形线 Pm 趋向于曲线 γ

如果算法在 m 步后停止，则上、下误差界为

$$\|x^0-x^*\|-m\eta\le\|x^m-x^*\|\le \mathrm{dia}(\mathcal{S}_{c_m}), \tag{4.2}$$

其中，直径（Sc）为 Sc 的直径，即 Sc 的任意两个元素之间的最大距离. 考虑一个有限数量的步骤 $m<\frac{\|x^0-x^*\|}{\eta}$. 也很有意义.

左不等式源于任何多边形线都大于连接其端点的线段

$$\begin{aligned}\|x^0-x^*\|\le\ & \|x^0-x^1\|+\|x^1-x^2\|+\dots+\|x^{m-1}-x^m\|+\|x^m-x^*\|\\ &=m\eta+\|x^m-x^*\|.\end{aligned}$$

如果多边形线是一条直线，这就变成了恒等式，如果超曲面是以 x 为中心的超球体，就会出现这种情况。

（4.2.2）右边的不等式来自于 xm 的构造，它属于 Scm。因此，我们有了估计

$$\|x^*-x^m\|\le\max_{y\in\mathcal{S}_{cm}}\|x^*-y\|\le\sup_{x,y\in\mathcal{S}_{cm}}\|x-y\|=\mathrm{dia}(\mathcal{S}_{c_m}).$$

收缩条件 Scm→x∗ 允许直径直径（Scm）尽可能小，前提是 η 足够小。

关于 η 的大小，我们有一些评论:

- 如果 η 较大，算法停止过早，然后达到点 x∗ 的良好近似，见图 4.6 a;
- 如果 η 太小，停止阶数 m 很大，在计算机实现的情况下可能没有时间有效，见图 4.6 b。

在实际应用中，步长 η 的大小是一个正在运行的应用程序的误差幅度和时间有效性之间的权衡。我们将在第 4.2.4 节中进一步正式阐述这一想法.

4.2.2 方向导数

后来使用的另一个概念是方向导数，它测量一个函数在给定方向上的某一点上的瞬间变化率。更准确地说，假设 v 是 Rn 中的一个酉向量，并考虑可微函数 f：U Rn→R。f 在第 x0 U 点处的方向导数定义为

$$\frac{\partial f}{\partial v}(x^0) = \lim_{t \searrow 0} \frac{f(x^0 + tv) - f(x^0)}{t}.$$

注意，关于坐标的偏导数，∂x∂fk，是关于坐标向量 v =（0，...，1，...，0）T 的方向导数。链规则的应用提供了作为标量积的方向导数的计算：

$$\begin{aligned}\frac{\partial f}{\partial v}(x^0) = \frac{d}{dt} f(x^0 + tv)\big|_{t=0+} &= \sum_{k+1}^{n} \frac{\partial f}{\partial x_k}(x^0 + tv)v^k \Big|_{t=0} \\ &= v^T \nabla f(x^0) = \langle \nabla f(x^0), v \rangle.\end{aligned}$$

4.2.3 最速下降法

最速下降法（或梯度下降法）是一种基于贪婪算法的数值方法，该算法通过将给定的一步指向函数值减少最大的方向来搜索函数的最小值。人们可以想象一个被蒙着眼睛的游客想要以最快的方式下山的方法。在每一点上，游客都在检查距离，以找到下降最陡的方向，然后朝那个方向走一步。然后重复这个过程，直到游客最终到达山谷的底部（或者，如果他的脚步太小，就会被困在当地的最小值）.

为了应用这种方法，我们感兴趣的是找到酉方向 v，其中函数 f 在给定的小步长 η 内尽可能地减小。函数 f 在初始点 x0 处的值和在方向 v 上大小为 η 的一步后的值之间的变化，用线性近似表示为

$$\begin{aligned}f(x^0 + \eta v) - f(x^0) &= \sum_{k=1}^{n} \frac{\partial f}{\partial x_k}(x^0)\eta v^k + o(\eta^2) \\ &= \eta \langle \nabla f(x^0), v \rangle + o(\eta^2).\end{aligned}$$

下面，由于 η 很小，我们将忽略二次项 o（η2）的影响。因此，为了得到 v，使函数的变化具有最大的负值，我们将使用标量积的柯西不等式

$$-\left\|\nabla f(x^0)\right\| \|v\| \le \left\langle \nabla f(x^0), v \right\rangle \le \left\|\nabla f(x^0)\right\| \|v\|.$$

已知道，对于负比例的向量达到左边的不等式。3 由于 v=1，最小值为

$$v = -\frac{\nabla f(x^0)}{\|\nabla f(x^0)\|}.$$

那么这个函数中的最大变化近似等于

$$f(x^0 + \eta v) - f(x^0) = \eta \langle \nabla f(x^0), v \rangle = -\eta \|\nabla f(x^0)\|.$$

常数 η 被称为学习速率。从前面的关系来看，每一步后函数的变化与梯度的大小和学习速率成正比.

该算法由以下迭代组成，它构造了以下序列（xn）：

- 在全局最小值 x∗ 的吸引盆地中选择一个初始点 x0,
- 使用迭代来构造序列 $(xn)n$.

$$x^{n+1} = x^n - \eta \frac{\nabla f(x^n)}{\|\nabla f(x^n)\|}. \tag{4.3}$$

这保证了目标函数的负变化，这是由 f（xn+1）−f（xn）=−η f（xn）<0 给出的。

我们注意到线 xnxn+1 是正常的水平超表面 Sf（xn）。因此，我们从前一节中得到了多边形线 Pm = [x0，...，xm]，它是由定理 4.2.3 提供的曲线 γ 的近似值。

然而，这种结构有一个缺点，它很快就会被修复。由于 xn+1−xn= η > 0，近似序列（xn）n 不收敛，因此很容易错过最小点 x∗，见图 4.7 a。为了克服这个问题，我们将假设学习速率 η 是可调的，即随着函数变化的较慢（当梯度较小时），它会变得更小，见图 4.7 b。我们假设现在有一个正的常数 δ > 0，这样在第 n 次迭代中的学习

率与梯度，ηn = δ f（xn）. 成正比然后，迭代（4.2.3）变为

$$x^{n+1} = x^n - \delta \nabla f(x^n). \tag{4.4}$$

命题 4.3

当且仅当梯度序列收敛于零，f（xn）→0 时，n→∞ 时，由（4.2.4）定义的序列（xn）n 为收敛

证明 ...

例题 4.4

4.2.4 线搜索法

这是具有可调学习率 η 的最速下降方法的一种变体。该速率的选择如下所示。从初始点 x0 开始，考虑由梯度 f（x0）给出的水平超曲面 Sf（x0）上的法线方向。我们需要在这个方向上选择一个点 x1，目标函数 f 达到最小值。这相当于选择值 η0 > 0，这样

$$\eta_0 = \arg\min_{\eta} f(x^0 - \eta \nabla f(x^0)). \tag{4.5}$$

该过程继续执行下一个起点 $x^1 = x^0 - \eta_0 \nabla f(x^0)$.. 通过一次迭代，我们得到了点序列（xn）和递归定义的学习率序列（ηn）

$$\begin{aligned} \eta_n &= \arg\min f(x^n - \eta \nabla f(x^n)) \\ x^{n+1} &= x^n - \eta_n \nabla f(x^n). \end{aligned}$$

刚才描述的线搜索方法具有以下几何意义。考虑函数

$$g(\eta) = f(x^0 - \eta \nabla f(x^0)),$$

并区分它

$$g'(\eta) = -\langle \nabla f(x^0 + \eta \nabla f(x^0), \nabla f(x^0) \rangle.$$

如果选择 η0 来实现最小值（4.2.5），那么 g（η0）= 0，这意味着通过（4.2.6），f（x1）和 f（x0）是法向量。当点 x1 作为线 x0−η f（x0）与水平超表面 Sf（x1）之间的切线接触得到时，就会发生这种情况，见图 4.10。

一般来说，算法继续如下：考虑从 xn 到超曲面 Sf（xn）的法线，并选择 xn+1 作为这条线与水平面相切的点。该算法产生的序列收敛到 x∗ 的速度比最快下降方法要快得多。请注意，多边形线 [x0，x1，x2，...] 是无限的，并且有直角。在进一步讨论之前，我们将提供一些例子。

例题 4.5 单变量线搜索方法

例题 4.6 多变量线搜索方法

4.3 运动学解释

本节讨论最陡下降方法的运动学解释。考虑一个质量为 m = 1 的球，它向山谷的底部向下，没有摩擦。球的状态用对 s =（x，v）来描述，其中 x 和 v 分别表示球的坐标和速度。状态的空间 s 被称为相空间。在相空间中可以追踪球的动态为曲线 s (t) =（x (t)，v (t)），其中 x (t) 是其坐标，v (t)=˙x (t) 是球在 t 时刻的速度，见图 4.11。

山谷的几何形状由一个凸函数 z = f (x) 建模。我们考虑两种作用于球上的能量：

- 由运动引起的动能：$E_k = \frac{1}{2}\|v\|^2$;;
- 由于高度所具有的势能：$E_p = f(x)$.

球的动力学用经典的拉格朗日来描述，它是动能和势能的差：

$$L(x, \dot{x}) = E_c - E_p = \frac{1}{2}\|\dot{x}\|^2 - f(x).$$

运动方程由欧拉-拉格朗日变分方程表示 $\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{x}}\right) = \frac{\partial L}{\partial x}$，也就是

$$\ddot{x}(t) = -\nabla f(x(t)). \tag{4.6}$$

这可以看作是一个单位质量球在由势 f (x) 导出的力下的牛顿运动定律。

球在 t 时刻的总能量被定义为动能和势能之和,

$$E_{tot}(t) = E_k(t) + E_p(t) = \frac{1}{2}\|\dot{x}(t)\|^2 + f(x(t)). \tag{4.7}$$

施加（4.3.7）得到

$$\begin{aligned}\frac{d}{dt}E_{tot}(t) = &\quad \frac{d}{dt}\Big(\frac{1}{2}\dot{x}(t)^T\dot{x}(t) + f(x(t))\Big)\\ = &\quad \dot{x}(t)^T\ddot{x}(t) + \dot{x}(t)^T\nabla f(x)\\ = &\quad \dot{x}(t)^T\big(\ddot{x}(t) + \nabla f(x(t))\big) = 0,\end{aligned}$$

也就是说，总能量沿着轨迹保持不变。我们可以使用这个观察结果来构建级别集，如下所示。考虑一个在相空间上定义的函数

$$H(x, v) = \frac{1}{2}\|v\|^2 + f(x), \tag{4.8}$$

称为哈密顿函数。定义由与函数 H 相关的超曲面 Sc 给出的能级

$$\mathcal{S}_c = H^{-1}(\{c\}) = \{(x, v); H(x, v) = c\}.$$

前面的计算表明，相空间中的任何解轨迹（x (t)、v (t)）的解轨迹都属于这些能级超曲面之一. 查看方程（4.3.7）的一种等价方法是将其写成 ode 的一阶系统

$$\dot{x}(t) \quad = \quad v(t) \tag{4.9}$$

$$\dot{v}(t) \quad = \quad -\nabla f(x(t)). \tag{4.10}$$

给定一些初始条件，x (0) = x0，v (0) = v0，标准 ODE 结果提供了从点（x0，v0）开始的唯一局部解（x (t)，v (t)）的存在性.

为了找到势函数 z = f (x) 的最小值，只要找到球的稳定平衡点，该平衡点在谷的底部实现了。如果球被放置在此时，以零速度，它将永远停留在那里。这一点是 ODE 系统（4.3.10）–（4.3.11）的一个平衡点，并且在相空间中对应于一个退化到一个点的能级。在前一个 ODE 系统中设置 ˙x(t)=0 和 ˙v(t)=0 提供了由给出的平衡点（x∗，v∗）.

$$v^* = 0, \quad \nabla f(x^*) = 0. \tag{4.11}$$

对相空间中能级超曲面 Sc 的最陡下降法，可以得到平衡态 ∗=（x∗，v∗）。定理 4.2.3 提供了一个最陡下降曲线的存在性，它将球的初始状态 s0=（x0，v0）（初始位置和速度）连接到平衡态 s∗=（x∗，v∗），只要 s0 足够接近 s∗。4. 迭代次数（4.2.4）变成

- 设置初始状态 $s_0 = (x^0, v^0)$;
- 递归地构造状态序列

$$s_{n+1} = s_n - \delta\,\nabla_s H(s_n), \quad \forall n \geq 0. \tag{4.12}$$

利用哈密顿函数（4.3.9）的表达式，它的梯度变为

$$\nabla_s H(s) = \left(\frac{\partial H(x, v)}{\partial x}, \frac{\partial H(x, v)}{\partial v}\right) = (\nabla f(x), v). \tag{4.13}$$

因此，表达式（4.3.13）现在可以写入组件为

$$\begin{aligned}x^{n+1} &= x^n - \delta\nabla f(x^n)\\ v^{n+1} &= (1-\delta)v^n.\end{aligned}$$

我们得到了两个可以独立求解的分离方程。第二种是封闭形式的溶液 vn =（1–δ）nv0，对于 δ 小的，我们有 vn→v∗= 0，n→∞。第一个方程只不过是迭代（4.2.4）。因此，除了一个很好的运动学解释外，这种方法并没有对第 4.2.3 节中描述的方法提供任何改进。为了进行改进，我们需要引入一个摩擦项，这将在下一节中完成。下面将讨论这两种方法之间的定性区别。

解流 x (t)、v (t) 满足系统（4.3.10)–(4.3.11)。这可以用哈密顿函数来表示，等价地，如

$$\dot{x}(t) = \frac{\partial H}{\partial v} \tag{4.14}$$

$$\dot{v}(t) = -\frac{\partial H}{\partial x}. \tag{4.15}$$

解流的切向量场由 X=˙x(t)∂∂x+˙v(t)∂∂v 定义。我们使用（4.3.14）–（4.3.15）来计算它的散度

$$\mathrm{div} X = \frac{\partial}{\partial x}\frac{\partial H}{\partial v} - \frac{\partial}{\partial v}\frac{\partial H}{\partial x} = 0.$$

因为向量场的散度代表体积沿流动曲线演变的速度，前面的关系可以解释说解决方案流是不可压缩的，也就是说，任何给定的体积的粒子保持其体积在动力系统的进化，见图 4.12。事实上，所有光滑函数的哈密顿流（H 的系统的解（4.3.14）- （4.3.15））的散度为零。因此，如果球从某些初始的邻近状态开始滚动，那么在系统演化过程的任何时候，这些状态都处于相同的体积接近，不可能收敛到任何平衡状态。一个在凸出的杯子里没有摩擦的球，永远不会停在杯子的底部；它会继续在杯壁上来回反弹，在平衡点附近无限次

4.4 动量法

为了避免陷入成本函数的局部最小值，设计了几种方法。其基本思想是，通过增加额外的速度或能量来震动系统，将使系统通过能量势垒，进入一个较低的能量状态。

我们已经看到，梯度下降的方法可以通过考虑一个球滚入一个杯子的物理模型来理解。球的位置通过一个给定的步骤被更新到所有时间的梯度的负方向，这就是学习速率。

动量法通过引入速度变量并使梯度改变速度而不是位置来修正梯度下降。正是速度的变化才会影响到这个位置。除了学习速率之外，该技术还使用了一个额外的超参数，来模拟摩擦，逐渐降低速度，使球走向一个稳定的平衡，也就是杯的底部。该方法的作用是在实现代价函数的基础上加速梯度下降法的最小化。

经典的动量法（Polyak，1964，见 [98]）提供了以下位置和速度的同步更新

$$x^{n+1} = x^n + v^{n+1} \tag{4.16}$$

$$v^{n+1} = \mu v^n - \eta \nabla f(x^n), \tag{4.17}$$

其中，η > 0 为学习速率，μ (0,1] 为动量系数。

值得注意的是，对于 μ→0，之前的模型恢复了我们熟悉的梯度下降模型，xn+1 = xn−η f（xn）

4.4.1 运动学解释

我们再次描绘一个球滚进一个杯的模型，其方程由 y = f (x) 给出，其中 f 是服从最小化的目标函数。我们将用 Ff 来表示球和杯壁之间的摩擦力。这个力与速度成正比，其方向与速度相反，Ff =−ρ˙x(t)，对于某些阻尼系数 ρ > 0。因此，牛顿运动定律被写成

$$\ddot{x}(t) = -\rho\dot{x}(t) - \nabla f(x(t)). \tag{4.18}$$

左侧是单位质量球的加速度，右侧是球作用的总力，是摩擦力与势 f 提供的力之和。这个方程可以用来表明，在这种情况下，球的总能量（4.3.8）没有沿解保持. 由于

$$\begin{aligned}\frac{d}{dt}E_{tot}(t) = & \frac{d}{dt}\Big(\frac{1}{2}\dot{x}(t)^T\dot{x}(t) + f(x(t))\Big) = \dot{x}(t)^T\ddot{x}(t) + \dot{x}(t)^T\nabla f(x)\\ = & \dot{x}(t)^T\big(\ddot{x}(t) + \nabla f(x(t))\big) = -\rho\dot{x}(t)^T\dot{x}(t) = -\rho\|\dot{x}(t)\|^2\\ = & -\rho\|v(t)\|^2 < 0,\end{aligned}$$

由此可见，总能量以与速度的平方成正比的速度下降。因此，Etot (t) 是一个递减函数，它在系统的平衡点处达到最小值。

该方程（4.4.18）可以等价地写成一个一阶 ode 系统为

$$\dot{x}(t) = v(t) \tag{4.19}$$

$$\dot{v}(t) = -\rho v(t) - \nabla f(x(t)), \tag{4.20}$$

其中，v (t) 表示在 t 时刻的球的速度。解流的切向量场，˙x(t)，˙v(t) 的散度等于

$$\operatorname{div}(\dot{x}, \dot{v}) = -\rho < 0.$$

这意味着解流正在收缩，解的轨迹越来越近，最终收敛到平衡点。这一点是通过将前一个系统的正确项等于零而得到的

$$v = 0 \tag{4.21}$$

$$-\rho v - \nabla f(x) = 0. \tag{4.22}$$

解（v∗，x∗）满足 v∗= 0，f（x∗）=0，这是与（4.3.12）所描述的无摩擦情况下相同的平衡点。这种情况下唯一的区别是，由于能量损失，相空间中的解从较高能级向较低能级移动，螺旋向下向平衡点移动，见图 4.12 b。

为了得到一个可以在计算机上实现的算法，我们将把该 ODE 系统（4.4.19）–（4.4.20）转换为一个有限差分系统。考虑等距时分 0 = t0 < t1 < · · · < tn <∞，并设 Δt = tn+1−tn 为一个常数时间步长。用（xnvn）=x（tn）、v（tn）表示系统的第 n 个状态。这个系统（4.4.19）–（4.4.20）变成了

$$\begin{aligned} x^{n+1} - x^n &= v^n \Delta t \\ v^{n+1} - v^n &= -\rho v^n \Delta t - \nabla f(x^n) \Delta t. \end{aligned}$$

用 μ < 1 代替 = Δt 和 μ=1−ρ，，得到了有限差分系统

$$x^{n+1} = x^n + \epsilon v^n \tag{4.23}$$

$$v^{n+1} = \mu v^n - \epsilon \nabla f(x^n). \tag{4.24}$$

将速度重新调整为 ˜v=v（通过改变测量单位在物理上可行），系统（4.4.23）–（4.4.24）变成其中，η=2 为学习速率（时间步长由 Δt =√η 给出）。

我们注意到方程（4.4.25）和（4.4.16）之间的速度指数差值，或 Polyak 的经典动量方法。这可以通过在我们的分析中将后向方程 xn+1−xn = vnΔt 替换为正向方程 xn+1−xn = vn+1Δt 来解决。

注意，当函数 f 是二次的时，它的梯度 f 是线性的，因此方程（4.4.23）–（4.4.24）形成了一个可以显式求解的线性系统。我们将在下一个例子中这样做.

例题 4.7

注 “动量法”这个名字在这里用得不恰当。对粒子的动量通常是向前推的，而在这种情况下，我们用摩擦力来抑制粒子。我们这样做是为了避免粒子越过平衡点。

然而，有时我们希望得到相反的效果：使粒子避免卡在局部最小值。在这种情况下，摩擦因子被一个动量因子所取代，这意味着给粒子速度一个破裂，以超过局部最小值。通过询问公式（4.4.24）中的条件 μ > 1，可以很容易地实现这一点。

4.4.2 收敛条件

在本节中，我们感兴趣的是研究由动量法方程（4.4.16）–（4.4.17）定义的序列 xn 和 vn 的收敛性。为了完成这项任务，我们将找到序列的精确公式。迭代方程（4.4.16）

$$\begin{aligned} x^n &= x^{n-1} + v^n \\ x^{n-1} &= x^{n-2} + v^{n-1} \\ \dots &= \dots \\ x^1 &= x^0 + v^1 \end{aligned}$$

然后相加，得到了这个位置用速度表示的表达式

$$x^n = x^0 + \sum_{k=1}^{n} v^k. \tag{4.25}$$

为简单起见，表示 bn = f（xn）。通过迭代方程（4.4.17），我们得到了

$$\begin{aligned} v^n = & \quad \mu v^{n-1} - \eta b_{n-1} \\ = & \quad \mu(\mu v^{n-2} - \eta b_{n-2}) - \eta b_{n-1} \\ = & \quad \mu^2 v^{n-2} - \mu\eta b_{n-2} - \eta b_{n-1} \\ = & \quad \mu^2(\mu v^{n-3} - \eta b_{n-3}) - \mu\eta b_{n-2} - \eta b_{n-1} \\ = & \quad \mu^3 v^{n-3} - \mu^2\eta b_{n-3} - \mu\eta b_{n-2} - \eta b_{n-1}. \end{aligned}$$

这可以通过归纳法来证明

$$v^n = \mu^n v^0 - \eta\Big(\mu^{n-1} b_0 + \mu^{n-2} b_1 + \cdots + \mu b_{n-2} + b_{n-1}\Big).$$

在下面，可以更方便地移动索引和使用求和约定来获得表达式

$$v^{n+1} = \mu^{n+1} v^0 - \eta \sum_{i=0}^{n} \mu^{n-i} b_i. \tag{4.26}$$

为了理解 vn+1 的行为，我们需要引入以下概念

定义 4.1 (卷积级数)

两个数值级数的卷积级数 $\sum_{n\geq 0} a_n$ 和 $\sum_{n\geq 0} b_n$ 是级数 $\sum_{n\geq 0} c_n$，用一般术语 $c_n = \sum_{i=0}^{n} a_i b_{n-i} = \sum_{i=0}^{n} a_{n-i} b_i$.

♣

以下两个结果将用于序列（xn）n 和（vn）n 的收敛性分析。

命题 4.4

设（an）n 是一个收敛于 0 的实数序列，n≥0bn 是一个绝对收敛的数值级数。然后

$$\lim_{n\to\infty} \sum_{i=0}^{n} a_i b_{n-i} = 0.$$

♠

定理 4.3

考虑两个数值级数 n≥0an 和 n≥0bn，一个收敛，另一个绝对收敛。那么它们的卷积级数是收敛的，它的和等于给定级数的和的乘积，即.

$$\sum_{n\geq 0} \left(\sum_{i=0}^{n} a_i b_{n-i} \right) = \left(\sum_{n\geq 0} a_n \right) \left(\sum_{n\geq 0} b_n \right).$$

♡

下面的结果提供了序列（xn）n 和（vn）n 的收敛性的一个表征。

命题 4.5

(a)

♠

证明

注 涅斯特罗夫提出了对经典动量法的一种改进。这是通过修改动量法的参数得到的；它不是在当前位置 xn 计算它，而是在修正的值 xn + μvn：

$$x^{n+1} = x^n + v^{n+1} \tag{4.27}$$

$$v^{n+1} = \mu v^n - \eta \nabla f(x^n + \mu v^n). \tag{4.28}$$

Nesterov 加速梯度法（简称 NAG）是一种比梯度下降法收敛速度更好的一阶优化方法。与动量法相比，NAG 对速度 v 的变化响应性更快，使该方法更稳定，特别是在 μ 值较大的情况下。

附录 A