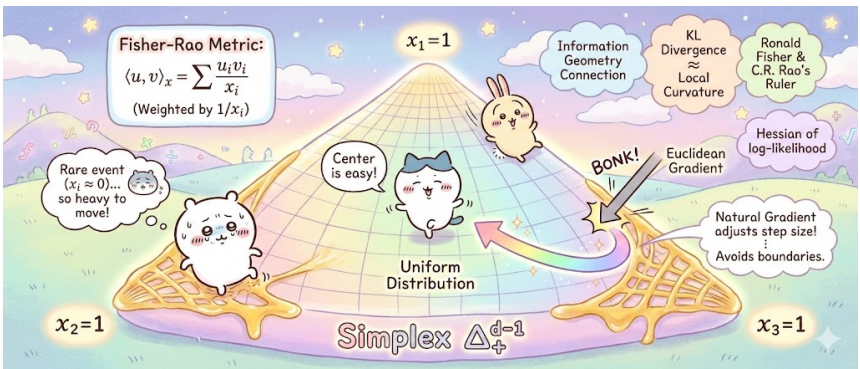
黎曼流形与黎曼梯度 (二)

北极甜虾 (南半球版)

2026.1.12

Manifold Optimization Notes



1 Fisher-Rao 度量下的黎曼梯度

我们来做书上的 Exercise 3.65，通过这个习题我们会学习到：当嵌入子流形上的度量不是从外在空间直接继承的诱导度量时，黎曼梯度就不再仅仅是欧几里得梯度的简单投影了。

定义 1: Δ_+^{d-1} 和 Fisher-Rao 度量

我们定义**单纯形的相对内部:**

$$\Delta_+^{d-1} = \left\{ x \in \mathbb{R}^d : x_1, \ldots, x_d > 0 \text{ 并且 } \sum_{i=1}^{d} x_i = 1 \right\}.$$

注意到 Δ_+^{d-1} 是 $\mathbb{R}^d$ 的嵌入子流形，其在 $x \in \Delta_+^{d-1}$ 处的切空间为

$$T_x\Delta_+^{d-1} = \left\{ v \in \mathbb{R}^d : \sum_{i=1}^{d} v_i = 0 \right\}.$$

Fisher–Rao 度量 $\langle \cdot, \cdot \rangle_x : T_x\Delta_+^{d-1} \times T_x\Delta_+^{d-1} \to \mathbb{R}$ 定义为

$$\langle u, v \rangle_x = \sum_{i=1}^{d} \frac{u_i v_i}{x_i}.$$

命题 2

配备了 Fisher-Rao 度量的 Δ_+^{d-1} 是一个黎曼流形。

证明. 由于黎曼度量要求度量随点 x 的变化是平滑的。在局部坐标下，这等价于度量矩阵的各分量是 x 的平滑函数。我们可以将该度量看作一个对角矩阵 $G(x)$，其分量为：

$$G_{ij}(x) = \begin{cases} \frac{1}{x_i} & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

由于在 $\mathcal{M}$ 上 $x_i > 0$，函数 $f(x) = \frac{1}{x_i}$ 在其定义域上是任意阶可导的。因此，映射 $x \mapsto \langle \cdot, \cdot \rangle_x$ 是平滑的。所以 Fisher-Rao 度量是一个黎曼度量。因此配备了 Fisher-Rao 度量的 Δ_+^{d-1} 是一个黎曼流形。 □

直观理解 我们讨论一下 Fisher-Rao 度量的意义：

- **加权的直观意义：**与标准的欧几里得度量（直接计算 $\sum u_i v_i$）不同，它为每一项乘以了 $1/x_i$ 作为权重。注意到 Δ_+^{d-1} 中的每一个向量 x 都是一个概率向量，这意味着，如果某个事件发生的概率 x_i 已经非常小（稀有事件），那么该分量的微小变化 u_i 会在度量中产生巨大的贡献。从几何上讲，在概率接近 0 的区域移动，比在概率接近 0.5 的区域移动“更费劲”。

- **梯度计算的变化：**在黎曼流形上，梯度取决于我们选择的度量。由于这里的度量不再是直接继承自欧几里得空间的“诱导度量”，黎曼梯度也就不再是欧几里得梯度的简单投影如果在此度量下运行梯度下降，算法会根据当前位置 x_i 的大小自动调整各分量的步长。

1.1 Fisher-Rao 度量下的黎曼梯度推导

在该练习中，我们现在来推导一个光滑函数 $f : \Delta_+^{d-1} \to \mathbb{R}$ 在 Fisher-Rao 度量下的黎曼梯度 $\operatorname{grad} f(x)$。

命题 3: Fisher-Rao 梯度的显式表达

对于定义在单纯形 Δ_+^{d-1} 上的光滑标量场 $f : \Delta_+^{d-1} \to \mathbb{R}$，其在 Fisher-Rao 度量下的黎曼梯度 $\operatorname{grad} f(x)$ 为:

$$\forall i \in [d], \quad (\operatorname{grad} f(x))_i = x_i \left(\frac{\partial f}{\partial x_i} - \sum_{j=1}^{d} x_j \frac{\partial f}{\partial x_j} \right).$$

证明. 记 $g = \operatorname{grad} f(x)$。根据黎曼梯度的定义，对于任何切向量 $u \in T_x \Delta_+^{d-1}$，必须满足

$$\langle g, u \rangle_x = Df(x)[u] = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} u_i. \tag{1}$$

将 Fisher-Rao 度量的定义代入 (1) 左侧：

$$\sum_{i=1}^{d} \frac{g_i u_i}{x_i} = \sum_{i=1}^{d} \frac{\partial f}{\partial x_i} u_i \implies \sum_{i=1}^{d} \left(\frac{g_i}{x_i} - \frac{\partial f}{\partial x_i} \right) u_i = 0.$$

由于该等式对所有满足 $\sum_{i=1}^{d} u_i = 0$ 的切向量 u 都成立，这意味着向量 v (分量为 $v_i = \frac{g_i}{x_i} - \frac{\partial f}{\partial x_i}$) 必然与切空间正交。在欧几里得空间中，这意味着 v 是常数向量，即存在 $\lambda \in \mathbb{R}$ 使得

$$\frac{g_i}{x_i} - \frac{\partial f}{\partial x_i} = \lambda \implies g_i = x_i \left(\frac{\partial f}{\partial x_i} + \lambda \right). \tag{2}$$

为了确定 λ，利用黎曼梯度 g 必须属于切空间 $T_x \Delta_+^{d-1}$ 的性质，即 $\sum_{i=1}^{d} g_i = 0$：

$$\sum_{i=1}^{d} x_i \left(\frac{\partial f}{\partial x_i} + \lambda \right) = 0 \implies \sum_{i=1}^{d} x_i \frac{\partial f}{\partial x_i} + \lambda \sum_{i=1}^{d} x_i = 0.$$

由于 $\sum x_i = 1$，解得 $\lambda = -\sum_{j=1}^{d} x_j \frac{\partial f}{\partial x_j}$。将 λ 代回 (2) 即得证。 □

结论与直观理解 该梯度表达式具有极佳的几何与统计解释性：

- **重加权 (Reweighting)：** 分量 x_i 的出现反映了度量的非均匀性。当概率 x_i 接近 0 时，该方向的更新步长会相应减小，防止越界。
- **中心化 (Centering)：** 括号内项 $\frac{\partial f}{\partial x_i} - E_x[\nabla f]$ 实际上是分量偏导数减去其期望值。这确保了 $\sum g_i = 0$，使得梯度迭代始终保持在单纯形内。

- **自然梯度:** 在信息几何中, 这种形式的梯度下降被称为**自然梯度下降 (Natural Gradient Descent)**。在演化博弈论中，该公式对应于著名的**复制子方程 (Replicator Dynamics)**。

2 Fisher-Rao 度量与信息几何的联系

Fisher-Rao 度量的名字来自于两位统计学大师：Ronald Fisher 和 C.R. Rao。

- Fisher 最初提出这个概念是为了衡量样本数据中包含多少关于未知参数的信息。直观上，如果概率密度函数随参数变化剧烈，那么这个参数就容易被估计，Fisher 信息量就大。

- 1945 年，Rao 意识到这个正定的 Fisher 信息矩阵可以被看作是黎曼流形上的度量, 开启了信息几何的大门：我们将概率分布看作流形上的点，而 Fisher 信息矩阵就是这片空间里的“尺子”。

Fisher-Rao 度量的用处 在欧几里得度量下，从概率 0.01 变到 0.02 的距离，和从 0.50 变到 0.51 是一样的（都是 0.01）。但在 Fisher–Rao 度量下，前者被视为更大的变化，因为它代表了信息量的巨大飞跃。

而且无论我们如何重新定义概率的参数（比如用对数坐标），使用 Fisher-Rao 度量的**自然梯度下降法 (Natural Gradient Descent)** 的更新步长在物理意义（分布的变化程度）上是保持一致的。

当前概率接近 0 或 1 时，普通的随机梯度下降会产生的“病态曲率”问题，但自然梯度下降能够避免在这些区域步长过小或产生剧烈振荡，从而让算法沿着最快降低信息差异（通常是 KL 散度）的方向前进。

2.1 统计推导：从似然函数到 $1/x_i$

定理 4

对于离散概率分布 $x = (x_1, \ldots, x_d) \in \Delta_+^{d-1}$，其 Fisher 信息矩阵 $G(x) = (G_{ij}(x)) \in \mathbb{R}^{d \times d}$：

$$G_{ij}(x) = \mathbb{E}_x \left[-\frac{\partial^2}{\partial x_i \partial x_j} \log p(Y|x) \right] = \sum_{k=1}^{d} x_k \left(-\frac{\partial^2 \log x_k}{\partial x_i \partial x_j} \right)$$

在单纯形内部恰好给出 Fisher-Rao 度量。

证明. 设随机变量 Y 有 d 个可能结果，其概率分布由参数 $x = (x_1, \ldots, x_d) \in \Delta_+^{d-1}$ 决定。观测到结果 k 的对数似然函数为：

$$\ell_k(x) = \log p(Y = k|x) = \log x_k$$

Fisher 信息矩阵 $G(x)$ 的分量定义为对数似然 Hessian 期望的负值：

$$G_{ij}(x) = \mathbb{E}_x \left[-\frac{\partial^2}{\partial x_i \partial x_j} \log p(Y|x) \right] = \sum_{k=1}^{d} x_k \left(-\frac{\partial^2 \log x_k}{\partial x_i \partial x_j} \right)$$

计算括号内的二阶导数：

$$\frac{\partial \log x_k}{\partial x_i} = \frac{1}{x_k} \delta_{ki}, \quad \frac{\partial^2 \log x_k}{\partial x_i \partial x_j} = \frac{\partial}{\partial x_j} \left(\frac{1}{x_k} \delta_{ki} \right) = -\frac{1}{x_k^2} \delta_{ki} \delta_{kj}$$

其中 δ_{ij} 是 Kronecker delta。代入求和式：

$$G_{ij}(x) = \sum_{k=1}^{d} x_k \left(\frac{1}{x_k^2} \delta_{ki} \delta_{kj} \right) = \begin{cases} 0 & i \neq j \\ \frac{1}{x_i} & i = j \end{cases}$$

因此，度量张量为对角阵 $G(x) = \mathrm{diag}(1/x_1, \ldots, 1/x_d)$. 这证明了 Fisher-Rao 度量 $\langle u, v\rangle_x = \sum \frac{u_i v_i}{x_i}$ 正是概率空间的自然统计度量。 □

2.2 几何本质：KL 散度与局部曲率

Fisher-Rao 度量的物理意义在于：它是 Kullback-Leibler (KL) 散度的二阶近似。这解释了为什么它能衡量“分布之间的距离”。

考虑两个邻近的分布 x 和 $x + \epsilon u$（其中 $u \in T_x\Delta$ 且 $\|\epsilon\|$ 极小）。对 $D_{KL}(x\|x + \epsilon u)$ 关于 ϵ 在 0 处进行 Taylor 展开：

$$D_{KL}(x\|x + \epsilon u) = \sum_{i=1}^{d} x_i \log \frac{x_i}{x_i + \epsilon u_i}$$

- **零阶项：** $D_{KL}(x\|x) = 0$。
- **一阶项：** 计算 $\frac{d}{d\epsilon} D_{KL}\big|_{\epsilon=0}$：

$$\begin{aligned}\frac{d}{d\epsilon} D_{KL} &= \sum x_i \frac{d}{d\epsilon}(\log x_i - \log(x_i + \epsilon u_i)) \\ &= \sum x_i \left(-\frac{u_i}{x_i + \epsilon u_i}\right)\end{aligned}$$

  当 $\epsilon = 0$ 时，该项为 $-\sum u_i = 0$（由于 u 是切向量）。

- **二阶项：** 计算 $\frac{d^2}{d\epsilon^2} D_{KL}\Big|_{\epsilon=0}$：

$$\begin{aligned}\frac{d^2}{d\epsilon^2} D_{KL} &= \sum x_i \frac{d}{d\epsilon}\left(-\frac{u_i}{x_i + \epsilon u_i}\right) \\ &= \sum x_i \frac{u_i^2}{(x_i + \epsilon u_i)^2} \xrightarrow{\epsilon=0} \sum \frac{u_i^2}{x_i}\end{aligned}$$

由此得到：$D_{KL}(x\|x + \epsilon u) \approx \frac{1}{2}\epsilon^2 \langle u, u\rangle_x$，所以 Fisher-Rao 度量衡量了在特定方向上移动时 KL 散度增加的速率，其曲率反映了局部敏感度。

2.3 算法应用：自然梯度下降

在优化中，标准梯度下降假设空间是平直的。而**自然梯度**通过在 KL 散度约束下寻找最速下降方向：

$$\text{minimize } Df(x)[u] \quad \text{s.t. } \frac{1}{2}\langle u, u\rangle_x \leqslant \epsilon$$

使用 Fisher-Rao 度量下的黎曼梯度 $\operatorname{grad} f(x) = G(x)^{-1}\nabla f(x)$，其优势在于：

1. **黎曼预条件：** $G(x)^{-1} = \operatorname{diag}(x_1, \ldots, x_d)$ 扮演了预条件矩阵，修正了病态曲率。
2. **自动步长调整：** 当 $x_i \to 0$ 时，自然梯度自动抑制该方向的步长（乘以 x_i），防止点跳出单纯形边界。
3. **收敛性：** 它捕捉了损失函数的二阶结构，在收敛速度上接近牛顿法，但由于 $G(x)$ 的对角结构，计算成本极低。

3 投影算子的光滑性

在流形优化中，算法（如梯度下降）的收敛性分析往往依赖于梯度场的光滑性。对于嵌入子流形，我们证明将点 x 映射到其切空间投影算子 Proj_x 的映射是光滑的，由此保证了光滑函数的黎曼梯度场也是光滑的。

命题 5: 投影映射的光滑性, Exercise 3.66

设 $\mathcal{M}$ 是欧几里得空间 $\mathcal{E}$ 的嵌入子流形。则映射 $P : \mathcal{M} \to \mathcal{L}(\mathcal{E}), x \mapsto \operatorname{Proj}_x$ 是光滑的。

证明. 由于光滑性是局部性质，考虑 $\mathcal{M}$ 在 x 附近的局部定义函数 $h : U \to \mathbb{R}^k$，使得 $\mathcal{M} \cap U = h^{-1}(0)$，且 $Dh(x)$ 在 U 上是满秩的。

此时，x 处的切空间可以表示为 $T_x\mathcal{M} = \ker(Dh(x))$。根据线性代数，从 $\mathcal{E}$ 到其子空间 $\ker(A)$ 的正交投影算子（其中 $A = Dh(x)$）可以显式表示为：

$$\mathrm{Proj}_x = \mathrm{Id} - Dh(x)^* \left(Dh(x) \circ Dh(x)^*\right)^{-1} Dh(x).$$

注意到：

1. 由于 h 是光滑的，其微分 $Dh(x)$ 也是 x 的光滑函数。
2. 由于 $Dh(x)$ 满秩，矩阵 $Dh(x)Dh(x)^*$ 是正定且可逆的。矩阵求逆在正定矩阵集合上是光滑操作。
3. 算子间的复合与伴随运算均保持光滑性。

因此，$x \mapsto \mathrm{Proj}_x$ 是光滑映射。 □

4 乘积流形上的梯度分解

处理具有多变量结构的优化问题（如矩阵分解、字典学习）时，我们经常在乘积流形 $\mathcal{M} \times \mathcal{M}'$ 上进行操作。我们证明在这种结构下，黎曼梯度可以分解为各个分量梯度的简单组合。

命题 6: 乘积流形上的梯度分解, Exercise 3.67

设 $\mathcal{M}, \mathcal{N}$ 为黎曼流形，其乘积流形 $\mathcal{M} \times \mathcal{N}$ 配备乘积度量。对于光滑函数 $f : \mathcal{M} \times \mathcal{N} \to \mathbb{R}$，其黎曼梯度为：

$$\operatorname{grad} f(x, y) = \left(\operatorname{grad}_x f(x, y), \operatorname{grad}_y f(x, y)\right).$$

其中 $\operatorname{grad}_x f$ 表示将 y 固定时关于第一个变量的黎曼梯度。

证明. 设 $(u, v) \in T_x\mathcal{M} \times T_y\mathcal{N}$ 为乘积流形上的任意切向量。根据乘积度量的定义：

$$\langle \operatorname{grad} f, (u, v) \rangle_{(x,y)} = \langle \operatorname{grad}_x f, u \rangle_x + \langle \operatorname{grad}_y f, v \rangle_y.$$

同时，根据黎曼梯度的定义和全微分的线性性质：

$$\langle \operatorname{grad} f, (u, v) \rangle_{(x,y)} = Df(x, y)[(u, v)] = D_1 f(x, y)[u] + D_2 f(x, y)[v].$$

由于上述等式对所有 $(u, 0)$ 和 $(0, v)$ 均成立，我们可以分别得到：

$$\langle \operatorname{grad}_x f, u \rangle_x = D_1 f(x, y)[u], \quad \langle \operatorname{grad}_y f, v \rangle_y = D_2 f(x, y)[v].$$

根据黎曼梯度的唯一性，命题得证。 □

注解 7: 在多变量优化中的应用

当处理复杂问题时，这种分解性质允许我们“分而治之”。例如在**字典学习 (Dictionary Learning)** 中，损失函数定义在单位球面流形（字典单元）与欧几里得空间（稀疏系数）的乘积上。利用这个结论，我们可以独立地计算每个流形分量的梯度，然后直接拼接，这极大简化了算法实现的复杂度。

参考文献

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