对称正定流形上的低复杂度子空间下降

Yogesh Darmwal YOGESHD@IITK.AC.IN
电气工程系，印度坎普尔理工学院，
北方邦，邮编 *208016*，印度

Ketan Rajawat KETAN@IITK.AC.IN
电气工程系，印度坎普尔理工学院，
北方邦，邮编 *208016*，印度

arXiv:2305.02041v4 [stat.ML] 18 Dec 2023

摘要

本文提出了用于在对称正定（SPD）流形上最小化函数的低复杂度黎曼子空间下降算法。与现有的黎曼梯度下降变体不同，所提方法采用精心选择的子空间，使更新可以表示为当前迭代点的 Cholesky因子与一个稀疏矩阵的乘积。由此得到的更新避免了矩阵指数运算和稠密矩阵乘法等代价高昂的矩阵运算，而这些运算通常是几乎所有其他 SPD流形上黎曼优化算法所必需的。我们还确定了一大类源自多种应用的函数，包括核矩阵学习、高斯分布的协方差估计、椭圆等高分布的最大似然参数估计，以及高斯混合模型问题中的参数估计；对于这类函数，可以高效地计算黎曼梯度。所提的单向和多向黎曼子空间下降变体的每次迭代复杂度分别为 $\mathcal{O}(n)$ 和 $\mathcal{O}(n^2)$，而现有所有黎曼梯度下降变体的复杂度则为 $\mathcal{O}(n^3)$ 或更高。通过针对大规模协方差估计和矩阵平方根问题的数值测试，也展示了所提算法在运行时间和每次迭代复杂度方面的优势。MATLAB 代码实现已在 GitHub 上公开：
https://github.com/yogeshd-iitk/subspace descent over SPD manifold

关键词： 子空间下降算法、对称正定（SPD）流形、测地凸性、矩阵平方根、黎曼自适应算法。

1 引言

我们考虑以下优化问题。

$$\min_{\mathbf{X}\in\mathbb{P}^n} f(\mathbf{X}) \tag{1}$$

其中， $\mathbb{P}^n$是由 $n\times n$实对称正定（SPD）矩阵构成的集合，且 $f\colon\ \mathbb{P}^n\to\mathbb{R}$是定义在 $\mathbb{P}^n$ 上的测地凸光滑函数（Zhang and Sra（2016））；参见定义3。此类问题出现在核矩阵学习（Li等（2009））、高斯分布的协方差估计（Wiesel（2012）；Wiesel等（2015）；Zhang等（2013））以及椭圆等高分布（ECD）的最大似然参数估计中。

（Sra 和 Hosseini（2013年））、高斯混合模型中的参数估计（Hosseini 和 Sra（2015年））、矩阵平方根（Higham（1997年）、Jain 等（2017年）、Sra（2015年）、Gawlik（2019年）、Oviedo 等（2020年））以及矩阵几何均值估计（Moakher（2005年）；Bhatia 和 Holbrook（2006年）；Bhatia（2013年）；Weber 和 Sra（2020年）；Zhang 等（2016年）；Liu 等（2017年）；Zhang and Sra（2016年））。我们旨在利用 f 和 $\mathbb{P}^n$ 的结构，以设计计算高效的流形优化算法来求解（1）。

求解 (1) 的标准方法依赖于对 $\mathbb{P}^n$ 的投影或内点法（Benson 和 Vanderbei（2003年）），随着 n 增大，实施起来会越来越困难。近年来，黎曼优化方法不断发展，旨在利用 $\mathbb{P}^n$ 的几何结构，提高效率和可扩展性（Zhang and Sra（2016年）；Zhang 等（2016年）；Liu 等（2017年））。然而，大多数最先进的黎曼优化方法，包括黎曼梯度下降及其变体，在每次迭代中都涉及矩阵指数运算、矩阵求逆、矩阵平方根和稠密矩阵乘法。因此，这些算法的每次迭代复杂度为 $\mathcal{O}(n^3)$，在大规模场景下可能难以承受，例如前文列举的各种应用场景。现代计算机在计算能力和内存方面的局限，促使人们亟需开发大规模优化算法，使其每次迭代复杂度相对于 n 呈二次甚至线性增长。

坐标（子空间）下降法已广泛应用于求解欧几里得空间中的大规模问题（Bertsekas（2016年）；Luenberger 和 Ye（2015年））。对于超大规模问题，按坐标进行运算尤其具有吸引力，因为全维向量运算的计算和内存开销高得难以承受（Nesterov（2012年）；Saha 和 Tewari（2013年）；Beck 等（2015年）；Karimi 等（2016年）；Richt´arik 和 Tak´ac（2014年）；Fercoq 和 Richtarik（2015年））。仔细观察可知，在欧几里得空间中，子空间下降算法之所以如此简单，是因为其基向量的选择（欧几里得空间中的典范基）以及更新方程具有线性特征。

将坐标下降算法直接应用于黎曼流形时，它们便失去了简洁性。由于流形不具备向量空间结构，更新的非线性特征使我们无法将坐标划分为可分别更新的块。在此背景下，Gutman 和 Ho-Nguyen（2022年）的研究针对一般流形提出了一种黎曼坐标下降算法，但该方法并不总能使求解（1）的算法在计算上更加简单。我们提出了黎曼子空间下降算法，通过维护并直接更新迭代点的 Cholesky因子，降低每次迭代的计算成本。所提算法的不同变体允许随机或贪心地选择一个或多个方向。

任何黎曼一阶优化算法的每次迭代复杂度都可分为两个部分：计算黎曼梯度的成本，以及执行更新的成本；后者可能涉及矩阵指数运算等复杂操作。因此，为了降低每次迭代复杂度，我们聚焦于以下这类函数：

$$\mathcal{F} = \left\{ f \middle| f(\mathbf{X}) = g\left(\begin{array}{l} \left\{\mathrm{tr}\left(\mathbf{C}_p\mathbf{X}^{-1}\right)\right\}_{1\le p\le P}, \left\{\mathrm{tr}\left(\mathbf{D}_q\mathbf{X}\right)\right\}_{1\le q\le Q}, \log\det\mathbf{X}, \\ \left\{\mathrm{tr}\left(\mathbf{X}\mathbf{A}_r\mathbf{X}\mathbf{H}_r\right)\right\}_{1\le r\le R}, \left\{\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\right)\right\}_{1\le s\le S}, \\ \left\{\mathrm{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\right)\right\}_{1\le m\le M} \end{array} \right) \right\}. \tag{2}$$

有趣的是， $\mathcal{F}$ 足以涵盖各种测地凸函数和非凸函数，并包含此前列出的所有应用，但 N 个对称正定矩阵的矩阵几何平均问题除外（适用于 $N > 2$）。对于 $N = 2$，两个矩阵 $\mathbf{W}_1$ 和 $\mathbf{W}_2$ 的矩阵几何平均等于该函数的最小化元（见附录 C）

$$f(\mathbf{X}) = \sum_{i=1}^{N} \operatorname{tr}\left(\mathbf{W}_i \mathbf{X}^{-1} + \mathbf{W}_i^{-1} \mathbf{X}\right) \tag{3}$$

其属于函数类 $\mathcal{F}$。

为简化更新方程，我们首先确定给定点处切空间的一组“规范”正交基向量。随后，精心选择更新方向，以避免在更新方程中执行稠密矩阵指数运算。由此得到的更新形式为：将迭代点的 Cholesky因子与一个稀疏矩阵相乘，这种形式便于高效实现。

所提算法的基本版本是单向随机黎曼子空间下降（RRSD）算法，它在每次迭代中随机选择一个（在可能的$\mathcal{O}(n^2)$个子空间方向中），且每次迭代的复杂度仅为 $\mathcal{O}(n)$。其多向变体在每次迭代中选择 $\mathcal{O}(n)$维子空间，每次迭代的复杂度为$O(n^2)$。我们还提出了贪心子空间选择规则，由此得到黎曼贪心子空间下降（RGSD）算法，其单向和多向变体的每次迭代复杂度分别为 $\mathcal{O}(n^2)$和 $\mathcal{O}(n^2\log(n))$。

关键结果总结于表1中。在表中， t 表示迭代总次数， μ 表示强凸参数， L 表示 Lipschitz光滑参数， c 是依赖于直径和截面曲率下界的常数（Zhang and Sra（2016））。在我们的情形中，因子$\binom{1}{n}$ 源于所提算法是子空间下降算法，它们在一个 $\mathcal{O}$（n）维子空间中运行；而黎曼梯度下降（RGD）和黎曼加速梯度下降（RAGD）则在流形$\mathbb{P}^n$的完整$\binom{n(n+1)}{2}$维切空间中运行。

算法	收敛率	每次迭代 复杂度	子空间- 维度
RRSD（提出）	$\mathcal{O}\left(\left(1-\frac{\mu}{4nL}\right)^t\right)$	$\mathcal{O}(n^2)$	$\mathcal{O}(n)$
RGSD（提出）	$\mathcal{O}\left(\left(1-\frac{\mu}{8nL}\right)^t\right)$	$\mathcal{O}(n^2\log(n))$	$\mathcal{O}(n)$
RGD Zhang 等人（2016年）	$\mathcal{O}\left(\left(1-\text{最小化}\ \left\{\frac{1}{c},\frac{\mu}{L}\right\}\right)^t\right)$	$\mathcal{O}$（n^3）	$\mathcal{O}$（n^2）
RAGD Zhang 等人（2018a）	$\mathcal{O}\left(\left(1-\frac{9}{10}\sqrt{\frac{\mu}{L}}\right)^t\right)$	$\mathcal{O}$（n^3）	$\mathcal{O}$（n^2）

表 1：对于式（2）中的强凸函数类，所提出的多向黎曼子空间下降算法与 RGD 和 RAGD 算法的比较。

1.1 相关工作

Sra 和 Hosseini（2015年）对流形 $\mathbb{P}^n$ 上的测地凸函数进行了全面分析。Zhang and Sra（2016年）首次给出了在阿达马流形上运行的若干一阶黎曼优化算法的复杂度。尤其重要的是 Zhang and Sra（2016年）提出的黎曼随机梯度下降（RSGD）算法，该算法针对以下形式的问题

$$\min_{\mathbf{X}\in\mathbb{P}^n} \sum_{s=1}^{S} f_p(\mathbf{X}) \tag{4}$$

使用更新

$$\mathbf{X}_{t+1} = \mathbf{X_t}^{1/2} \exp\left(-\alpha_t S \mathbf{X}_t^{-1/2} \mathrm{grad}^R f_s\left(\mathbf{X}_t\right) \mathbf{X}_t^{-1/2}\right) \mathbf{X}_t^{1/2} \tag{5}$$

其中， s 是从索引集 $\{1, 2, \ldots, S\}$ 中均匀选取的索引，$\mathrm{grad}^R f_s$ 是函数 f_s的黎曼梯度；参见定义 2。尽管 RSGD 算法的复杂度低于 RGD 算法，但它仍在每次迭代中使用矩阵指数运算和矩阵乘法，因此每次迭代的复杂度为 $\mathcal{O}(n^3)$。

Liu 等人（2017）将 Nesterov 的加速思想推广到非线性空间，提出了首个针对测地凸函数的加速一阶算法。正如 Zhang 等人（2018a）所指出的，所提算法的一个局限是每次迭代都依赖于非线性方程的精确解。为解决这一问题，Zhang 等人（2018a）针对测地线强凸函数类提出了一种易于实现的加速算法。Alimisis 等人（2021）还提出了向更广泛的测地凸函数类和弱拟凸函数类的扩展。最后，加速 RGD 的随机变体也已得到发展（Bonnabel，2013年；Zhang 等人，2018b、2016年；Kasai 等人，2016年；Tripuraneni 等人，2018年；Babanezhad 等人，2019年；Hosseini 和 Sra，2020年；Ahn 和 Sra，2020年）。由于流形的非线性，以及点$\mathbf{X}$ 处的切空间 $T_{\mathbf{X}}\mathcal{M}$ 不同于$\mathbf{Y} \neq \mathbf{X}$处的切空间$T_{\mathbf{Y}}\mathcal{M}$ 这一事实，所有加速 RGD 变体在每次迭代中都需要执行计算量巨大的矩阵指数运算和平行移动步骤（Sra 和 Hosseini，2015年；Zhang 等人，2018a）。

将坐标下降的思想扩展到 Stiefel流形的最初尝试见于 Celledoni 和 Fiori（2008）、Shalit 和 Chechik（2014）以及 Gao 等人（2018）。沿着将坐标下降算法扩展到一般流形这一方向开展的首项研究是 Gutman 和 Ho-Nguyen（2022），其研究动机源于欧几里得空间上的块坐标下降（BCD）算法。与 BCD 类似，Gutman 和 Ho-Nguyen（2022）提出的切空间子空间下降（TSD）算法首先选择一个合适的切子空间，将梯度投影到该子空间上，然后沿所得方向执行一次下降更新步。Gutman 和 Ho-Nguyen（2022）提出的方法在计算上可能代价较高，因为在确定性情形下，它需要执行多次平行移动操作；而在随机化情形下，则需要使用任意的子空间分解。此外，无论选择何种子空间，在 SPD流形的情形下，完成更新步骤仍然需要进行矩阵指数运算，因此每次迭代复杂度仍为 $\mathcal{O}(n^3)$。然而，Gutman 和 Ho-Nguyen（2022）将 Shalit 和 Chechik（2014）针对正交集合提出的降复杂度子空间选择方法进行了扩展。

矩阵到一般 Stiefel 流形。在这方面，我们的工作与 Shalit 和 Chechik（2014年）以及 Gutman 和 Ho-Nguyen（2022年）的研究类似，因为我们提出了一种针对 SPD 流形的降复杂度子空间选择方法。其他研究，例如 Huang 等人（2021年）以及 Firouzehtarash 和 Hosseini（2021年），则将 Gutman 和 Ho-Nguyen（2022年）提出的类似思想应用于乘积流形，通过分块更新变量，每次更新与乘积流形的一个流形分量相对应的变量。

2 符号与背景

本节说明本文所采用的符号，并讨论理解所提出优化算法推导过程所需的相关背景。

2.1 符号

我们用粗体小写（大写）字母表示向量（矩阵）。迹和转置运算分别记为 tr (·) 和 (·)$^\mathsf{T}$。当条件 $\mathcal{C}$ 成立时，指示函数 $\mathbf{1}_{\mathcal{C}}$ 取值为 1，否则取值为 0。矩阵 A 的第 (i, j) 个元素、第 i 行和第 i 列分别记为 $[\mathrm{A}]_{ij}$、$[\mathrm{A}]_{i:}$ 和 $[\mathrm{A}]_{:i}$。我们用 I_n 表示$n \times n$ 单位矩阵；当其维数可由上下文确定时，也记为 I。同理，I_{ij} 表示一个方阵，其第 (i, j) 个元素为 1，其余元素均为 0；矩阵的维数由上下文推知。对称正定矩阵 A 的下三角 Cholesky 因子记为 $\mathcal{L}(\mathrm{A})$，因此 $\mathrm{A} = \mathcal{L}(\mathrm{A})\mathcal{L}(\mathrm{A})^\mathsf{T}$，其最小特征值记为 $\lambda_{\min}(\mathrm{A})$。函数 $f\colon \mathbb{R}^{n\times n} \to \mathbb{R}$ 的梯度记为 $\mathrm{grad} f(\mathrm{X})$，其黎曼梯度记为 $\mathrm{grad}^R f(\mathrm{X})$。欧氏Hessian和黎曼Hessian分别记为 $Hf(\mathrm{X})$ 和 $H^R f(\mathrm{X})$。对于配备黎曼联络 ∇ 的黎曼流形 $\mathcal{M}$，黎曼Hessian $H^R f(\mathrm{X})$ 是线性映射 $H^R f(\mathrm{X})\colon T_{\mathbf{X}}\mathcal{M} \to T_{\mathbf{X}}\mathcal{M}$，定义为 $H^R f(\mathrm{X})[\mathrm{V}] = \nabla_{\mathbf{V}} \mathrm{grad}^R f(\mathrm{X})$（Boumal，2023，第90页）。欧氏范数和Frobenius范数分别记为 $\|\cdot\|_2$ 和 $\|\cdot\|_F$。点 $\mathrm{X} \in \mathbb{P}^n$ 处的切空间记为 $T_{\mathbf{X}}\mathbb{P}^n$，切向量用粗体希腊小写字母表示，例如 $\boldsymbol{\xi}$。给定两个切向量 $\boldsymbol{\xi}$、$\boldsymbol{\eta} \in T_{\mathbf{X}}\mathbb{P}^n$，它们的内积为 $\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\mathbf{X}}$，相应的范数为 $\|\boldsymbol{\xi}\|_{\mathbf{X}} := \sqrt{\langle \boldsymbol{\xi}, \boldsymbol{\xi} \rangle_{\mathbf{X}}}$。给定任意 X、$\mathrm{Y} \in \mathbb{P}^n$，以及连接它们的测地线 $\gamma(\lambda)$，满足 $\gamma(0) = \mathrm{X}$ 且 $\gamma(1) = \mathrm{Y}$，则 X 处的切向量记为 $\boldsymbol{\xi}_{\mathbf{XY}} := \gamma'(0)$。具有特征值分解 UDU^T 的对称正定矩阵 W 的幂映射和指数映射定义为

$$\mathbf{W}^k = \mathbf{U}\mathbf{D}^k\mathbf{U}^\mathsf{T} \qquad\qquad \exp(\mathbf{W}) = \mathbf{U}\exp(\mathbf{D})\mathbf{U}^\mathsf{T} = \sum_{k=0}^{\infty} \frac{\mathbf{W}^k}{k!} \tag{6}$$

对于 $k \in \{0, 1, 2, \dots\}$，其中 $[\mathbf{D}^k]_{ii} = [\mathbf{D}]_{ii}^k$ 且 $[\exp(\mathbf{D})]_{ii} = \exp([\mathbf{D}]_{ii})$ 对所有 $1 \le i \le n$ 均成立。

2.2 流形优化背景

本研究涉及黎曼流形，即配备黎曼度量的流形（Lee，2018）。对于本文所研究的对称正定矩阵流形，其切空间 $T_{\mathbf{X}}\mathbb{P}^n$ 与对称矩阵集合 $\mathbb{S}^n$ 自然同构（Bridson

以及 H¨afliger（2011）），并赋予黎曼度量

$$\langle \boldsymbol{\xi}, \boldsymbol{\eta} \rangle_{\mathbf{X}} = \operatorname{tr}\left(\mathbf{X}^{-1}\boldsymbol{\xi}\mathbf{X}^{-1}\boldsymbol{\eta}\right). \tag{7}$$

式 (7) 中度量的选择使 $\mathbb{P}^n$ 成为阿达马流形，即截面曲率非正的流形。阿达马流形在设计优化算法时很有用，因为任意两点之间都存在唯一的距离最短曲线（测地线）。具体而言，起点为 $\mathbf{X} \in \mathbb{P}^n$、方向为切向量 $\gamma'(0) = \boldsymbol{\xi}$ 的测地线 $\gamma : [0,1] \to \mathbb{P}^n$ 由 Bhatia (2007) 给出：

$$\gamma(\lambda) = \mathbf{X}^{1/2} \exp\left(\lambda \mathbf{X}^{-1/2}\boldsymbol{\xi}\mathbf{X}^{-1/2}\right)\mathbf{X}^{1/2}. \tag{8}$$

等价地，给定 Cholesky分解 $\mathbf{X} = \mathbf{B}\mathbf{B}^{\mathsf{T}}$，测地线也可以写为

$$\gamma(\lambda) = \mathbf{B} \exp\left(\lambda \mathbf{B}^{-1}\boldsymbol{\xi}\mathbf{B}^{-\mathsf{T}}\right)\mathbf{B}^{\mathsf{T}}. \tag{9}$$

最后，Expx：$T_{\mathbf{X}}\mathbb{P}^n \to \mathbb{P}^n$ 是指数映射，使得 Expx($\boldsymbol{\xi}$) $= \gamma(1)$.

对于流形 $\mathbb{P}^n$，沿式 (8) 中测地线的切向量 $\boldsymbol{\eta} \in T_{\mathbf{X}}\mathbb{P}^n$ 的平行移动由 Sra 和 Hosseini (2015) 给出：

$$P_{\mathbf{X}}(\lambda) = \mathbf{X}^{1/2} \exp\left(\lambda \frac{1}{2}\mathbf{X}^{-1/2}\boldsymbol{\xi}\mathbf{X}^{-1/2}\right)\mathbf{X}^{-1/2}\eta\mathbf{X}^{-1/2} \exp\left(\lambda \frac{1}{2}\mathbf{X}^{-1/2}\boldsymbol{\xi}\mathbf{X}^{-1/2}\right)\mathbf{X}^{1/2}. \tag{10}$$

我们指出，在加速梯度算法中，计算动量需要平行移动（Bonnabel (2013); Zhang 等人 (2018b, 2016); Kasai 等人 (2016); Tripuraneni 等人 (2018); Babanezhad 等人 (2019); Hosseini 和 Sra (2020); Ahn 和 Sra (2020)），而且这是这些算法中计算开销最大的步骤。不过，所提出的 RRSD 和 RGSD 算法不会使用平行移动步骤。接下来，我们介绍黎曼流形 $\mathbb{P}^n$ 的一些重要定义，其黎曼度量如式 (7) 所指定。

我们首先考察方向导数和黎曼梯度 grad$^R f$(**X**) 的定义（Absil 等，2008年，第40页；Boumal，2014年，第24页）。

定义 1（方向导数）设 $f : \mathbb{P}^n \to \mathbb{R}$ 是一个光滑函数，且 $\gamma(\lambda){:}\mathbb{R} \to \mathbb{P}^n$ 是满足 $\gamma(0) = \mathbf{X}$ 和 $\gamma'(0) = \boldsymbol{\xi}$的光滑曲线。$f$ 在$\mathbf{X}$ 处沿方向 $\boldsymbol{\xi} \in T_{\mathbf{X}}\mathbb{P}^n$ 的方向导数是标量（*Absil* 等，*2008*，第 *40* 页）（*Boumal*，*2014*年，第 *24* 页）：

$$Df_{\mathbf{X}}(\boldsymbol{\xi}) = \frac{d}{dt} f(\gamma(\lambda))\Big|_{\lambda=0} \tag{11}$$

定义 2（黎曼梯度）可微函数 $f : \mathbb{P}^n \to \mathbb{R}$ 在 $\mathbf{X} \in \mathbb{P}^n$ 处的黎曼梯度定义为唯一的切向量 *grad* $^R f$（$\mathbf{X}$）$\in T_{\mathbf{X}}\mathbb{P}^n$，满足（*Absil* 等人，*2008*年，第 *46* 页）（*Boumal*，*2014*年，第 *26* 页）：

$$Df_{\mathbf{X}}(\boldsymbol{\xi}) = \langle grad^R f(\mathbf{X}), \boldsymbol{\xi} \rangle_{\mathbf{X}} \tag{12}$$

$\mathbb{P}^n$ 的黎曼梯度和欧几里得梯度之间的关系为（Sra 和 Hosseini，2015年，第 722 页）（Ferreira 等人，2020年，第 506 页）

$$\mathrm{grad}^R f(\mathbf{X}) = \mathbf{X}\mathrm{grad} f(\mathbf{X})\mathbf{X}. \tag{13}$$

我们对 f 的光滑性作出如下假设。

A1 函数 f： $\mathbb{P}^n \to \mathbb{R}$ 的梯度向量场是 L-利普希茨光滑的。

假设 **A1** 的一个推论是，给定任意两个点 $\mathbf{X}$、$\mathbf{Y} \in \mathbb{P}^n$，以及一条从 $\mathbf{X}$ 出发、满足 $\gamma'(0) = \boldsymbol{\xi}_{\mathbf{XY}}$ 的连接测地线，以下不等式成立（Zhang 和 Sra（2016年）；Liu 等人（2017年））：

$$f(\mathbf{Y}) \leq f(\mathbf{X}) + \langle \mathrm{grad}^R f(\mathbf{X}), \boldsymbol{\xi}_{\mathbf{XY}} \rangle_{\mathbf{X}} + \frac{L}{2} \|\boldsymbol{\xi}_{\mathbf{XY}}\|_{\mathbf{X}}^2. \tag{14}$$

一个具有 Lipschitz光滑梯度的函数示例为（Ferreira 等人（2019年））

$$f(\mathbf{X}) = a \log(\det(\mathbf{X})^{b_1} + b_2) - c \log(\det \mathbf{X}) \tag{15}$$

对于正数 a、b_1、b_2 和 c，常数为 $L < ab_1^2 n$。有趣的是，函数 $\mathrm{tr}\left(\mathbf{C}\mathbf{X}^{-1}\right)$ 和 $\mathrm{tr}(\mathbf{CX})$ 的梯度向量场并非利普希茨光滑，除非将其限制在紧集上。

接下来，我们将讨论函数在 $\mathbb{P}^n$ 上的（强）凸性概念。

定义 3（测地凸函数）若函数 $f : \mathbb{P}^n \to \mathbb{R}$ 在所有测地线上的限制都是凸的，则称其为测地凸函数（*g*-凸）（*Rapcs´ak，1997*年，第 *64* 页）。

定义 3 意味着，对于 $0 \leq \lambda \leq 1$，连接任意两个点 $\mathbf{X},\mathbf{Y} \in \mathbb{P}^n$ 的每条测地线$\gamma(\lambda)$，以下不等式均成立：

$$f(\gamma(\lambda)) \leq (1-\lambda) f(\mathbf{X}) + \lambda f(\mathbf{Y}). \tag{16}$$

例如，对于 $\mathbf{C} \succcurlyeq 0$，$\mathrm{tr}\left(\mathbf{C}\mathbf{X}^{-1}\right)$ 和 $\mathrm{tr}(\mathbf{CX})$ 都是 $\mathbb{P}^n$ 中的测地凸函数。

定义 4（测地 μ-强凸函数）若函数 f: $\mathbb{P}^n \to \mathbb{R}$ 满足：对于任意两点 $\mathbf{X},\mathbf{Y} \in \mathbb{P}^n$，以及从 $\mathbf{X}$ 出发且满足 $\gamma'(0) = \boldsymbol{\xi}_{\mathbf{XY}}$ 的连接测地线，以下不等式成立，则称该函数为测地 μ-强凸（μ-强 *g*-凸）（*Zhang and Sra*（*2016*年）；*Liu* 等（*2017*年））：

$$f(\mathbf{Y}) \geq f(\mathbf{X}) + \langle grad^R f(\mathbf{X}), \boldsymbol{\xi}_{\mathbf{XY}} \rangle_{\mathbf{X}} + \frac{\mu}{2} \|\boldsymbol{\xi}_{\mathbf{XY}}\|_{\mathbf{X}}^2. \tag{17}$$

定义了黎曼空间中的强凸性概念后，我们提出以下假设。

A2 函数 f： $\mathbb{P}^n \to \mathbb{R}$ 是 μ-强 *g*-凸的。

与欧几里得情形一样， $\mathbb{P}^n$ 中的距离函数的平方由

$$d(\mathbf{X},\mathbf{Y})^2 = \|\log(\mathbf{X}^{-\frac{1}{2}}\mathbf{Y}\mathbf{X}^{-\frac{1}{2}})\|_F^2, \tag{18}$$

是 μ-强 g-凸的，且满足 $\mu = 2$。另一个同样属于式（2）中定义的函数类 $\mathcal{F}$ 的 μ-强 g-凸函数示例如下

$$f(\mathbf{X}) = \operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{D}\mathbf{X}\right) \tag{19}$$

对于$\mathbf{C} \succ 0$、$\mathbf{D} \succ 0$，且 $\mu = \min(\lambda_{\min}(\mathbf{C})\ \lambda_{\min}(\mathbf{D}))$（参见引理10）。

沿下降方向 $-\text{grad}^R f(\mathbf{X})$ 在黎曼流形 $\mathbb{P}^n$ 上的 RGD 更新方程为

$$\mathbf{X}_{t+1} = \mathbf{B}_t \exp\left(-\alpha_t \mathbf{B}_t^{-1}\text{grad}^R f\left(\mathbf{X}\right)\mathbf{B}_t^{-\mathsf{T}}\right)\mathbf{B}_t^{\mathsf{T}} \tag{20}$$

其中， α_t 是步长，$\mathbf{B}_t$ 是$\mathbf{X}_t$的 Cholesky因子。

所提算法的 RGD 和更新矩阵都通过函数 $\mathbf{F(X)} = \mathbf{B}^{-1}\text{grad}^R f(\mathbf{X})\mathbf{B}^{-\mathsf{T}}$依赖黎曼梯度 $\text{grad}^R f(\mathbf{X})$，如我们将在第3节中看到的那样。RGD 依赖于$\mathbf{F(X)}$的全部内容，而所提算法的更新矩阵只依赖于其中少数几个条目 $\{[\mathbf{F(X)}]_{ij}\}_{(i,j)\in\mathcal{E}}$，对于 $|\mathcal{E}| \leq n$。为使讨论具有一般性，我们假设计算完整矩阵$\mathbf{F(X)}$的复杂度为 $\mathcal{O}(M)$，而计算单个条目 $[\mathbf{F(X)}]_{ij}$ 的复杂度为 $\mathcal{O}(m)$，其中 $m \leq M \leq {}^{mn}{}_{(n+1)\ 2}$。一般而言，我们注意到 $M = m = \mathcal{O}(n^3)$，除非该函数具有特殊结构。

3 黎曼随机子空间下降算法

本节详细介绍所提 RRSD 算法，并讨论其相较于其他黎曼一阶算法的计算优势。首先需要注意，与欧几里得空间中的坐标下降算法不同，通常无法以低复杂度更新迭代点$\mathbf{X}_t$ 中的某个特定元素。相反，我们必须找出一组张成$\mathbf{X}_t$ 处切空间的标准正交基向量，并沿其中一个或多个基向量进行更新。有趣的是，所选子空间和确定的基向量将使我们能够根据$\mathbf{X}_t$的 Cholesky因子$\mathbf{B}_t$ 直接计算$\mathbf{X}_{t+1}$ 的 Cholesky分解$\mathbf{B}_{t+1}$。因此，所提算法的更新形式为 $\mathbf{B}_{t+1} = \mathbf{B}_t\mathbf{B}_{t+1}^{\text{up}}$，其中更新矩阵$\mathbf{B}_{t+1}^{\text{up}}$ 是稀疏矩阵，能够高效计算。

3.1 单方向更新

我们首先确定给定 $\mathbf{X}$ 处切空间 $T_{\mathbf{X}}\mathbb{P}^n$ 的一个典范基，并说明沿此基进行的更新可以直接通过 $\mathbf{X}$ 的 Cholesky因子 $\mathbf{B}$ 来完成。因此，每次迭代沿单个基向量进行更新，即可得到所谓的黎曼坐标下降，其每次迭代复杂度为 $\mathcal{O}(m+n)$。回顾一下，计算 $\mathbf{F(X)} = \mathbf{B}^{-1}$ $\text{grad}^R f(\mathbf{X})\mathbf{B}^{-\mathsf{T}}$ 中单个元素的复杂度记为 $\mathcal{O}(m)$。

3.1.1 Canonical Basis

注意，$\mathbf{X}$ 处的切空间 $T_{\mathbf{X}}\mathbb{P}^n$ 由以下标准正交基向量张成

$$\mathbf{G}_{ij}^R(\mathbf{X}) = \mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T} \tag{21}$$

对于所有 $1 \le j \le i \le n$，其中回顾 $\mathbf{B} = \mathcal{L}(\mathbf{X})$，且

$$\mathbf{E}_{ij} = (\mathbf{I}_{ij} + \mathbf{I}_{ji})\sqrt{2}^{-1-\mathbf{1}_{i=j}} \tag{22}$$

其中，指示函数 $\mathbf{1}_{i=j}$ 在 $i = j$ 时为 1，否则为 0。(22) 中的定义确保 $\mathbf{E}^{ij}$ 在 (i, j)-th 和 (j, i)-th 位置包含 $1/\sqrt{2}$，并且对于 $i \neq j$，其他位置均为 0；而 $\mathbf{E}_{ii}$ 在 (i, i)-th 位置为 1，其他位置均为 0。通过观察以下关系，可以验证基向量的正交归一性：

$$\begin{aligned}
\langle \mathbf{G}_{ij}^R(\mathbf{X}), \mathbf{G}_{k\ell}^R(\mathbf{X})\rangle_{\mathbf{X}} &= \operatorname{tr}\left(\mathbf{G}_{ij}^R(\mathbf{X})\mathbf{X}^{-1}\mathbf{G}_{k\ell}^R(\mathbf{X})\mathbf{X}^{-1}\right) && (23)\\
&= \operatorname{tr}\left(\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{B}\mathbf{E}_{k\ell}\mathbf{B}^\mathsf{T}\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\right) && (24)\\
&= \operatorname{tr}\left(\mathbf{E}_{ij}\mathbf{E}_{k\ell}\right) = \mathbf{1}_{i=k}\mathbf{1}_{j=\ell} && (25)
\end{aligned}$$

对于所有 $1 \le j \le i \le n$ 和 $1 \le \ell \le k \le n$。因此，f 的黎曼梯度可以写为

$$\operatorname{grad}^R f(\mathbf{X}) = \sum_{1\le j\le i\le n} \beta_{ij}(\mathbf{X})\mathbf{G}_{ij}^R(\mathbf{X}) \tag{26}$$

其中

$$\begin{aligned}
\beta_{ij}(\mathbf{X}) &= \langle \operatorname{grad}^R f(\mathbf{X}), \mathbf{G}_{ij}^R(\mathbf{X})\rangle_{\mathbf{X}} && (27)\\
&= \operatorname{tr}\left(\operatorname{grad}^R f(\mathbf{X})\mathbf{X}^{-1}\mathbf{G}_{ij}^R(\mathbf{X})\mathbf{X}^{-1}\right) && (28)\\
&= \operatorname{tr}\left(\mathbf{B}^{-1}\operatorname{grad}^R f(\mathbf{X})\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right) && (29)\\
&= \operatorname{tr}\left(\mathbf{F}(\mathbf{X})\mathbf{E}_{ij}\right) && (30)\\
&= \sqrt{2}^{\mathbf{1}_{i\neq j}}[\mathbf{F}(\mathbf{X})]_{ij}. && (31)
\end{aligned}$$

今后，我们将省略系数 $\beta_{ij}(\mathbf{X})$ 的自变量，并将其简记为 β_{ij}，对于所有 $1 \le j \le i \le n$。按照我们的记号，对任意 $1 \le j \le i \le n$ 计算 β_{ij} 的复杂度为 $\mathcal{O}(m)$。

3.1.2 Updating the Cholesky factors

确定合适的典范基后，我们现在可以写出黎曼版本的随机坐标下降，其中我们仅沿一个随机选取的基向量更新 $\mathbf{X}_t$。注意，沿 $\mathbf{G}_{ij}^R(\mathbf{X}_t)$ 的更新为

$$\begin{aligned}
\mathbf{X}_{t+1} &= \operatorname{Exp}_{\mathbf{X}_t}\left(-\alpha_t\beta_{ij}\mathbf{G}_{ij}^R(\mathbf{X}_t)\right) && (32)\\
&= \mathbf{B}_t \exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij})\mathbf{B}_t^\mathsf{T}. && (33)
\end{aligned}$$

$\mathbf{E}^{ij}$ 的特殊形式使我们能够高效地计算 Cholesky因子 $\mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}^{ij}))$。接下来将结果分为两种情况：(a) $i \neq j$ 和 (b) $i = j$。**情况** $i \neq j$。记 $w = \exp(-\alpha_t\beta_{ij}\sqrt{2}^{-\mathbf{1}_{i\neq j}})$，注意到

$$\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij}) = \mathbf{I} + (\mathbf{I}_{ii} + \mathbf{I}_{jj})(\frac{w}{2} + \frac{1}{2w} - 1) + (\mathbf{I}_{ij} + \mathbf{I}_{ji})(\frac{w}{2} - \frac{1}{2w});$$

对于 $i \neq j$。因此，我们有

$$\mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij})) = \mathbf{I} + \mathbf{I}_{jj}(u-1) + \mathbf{I}_{ii}(\frac{1}{u} - 1) + \mathbf{I}_{ij}\frac{w - 1/w}{2u}; \tag{34}$$

其中 $u = \sqrt{\frac{w+1/w}{2}}$。换言之， $\mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}^{ij}))$ 的主对角线上元素均为 1，除了第 (j,j) 个和第 (i,i) 个元素分别为 u 和 $1/u$，并且在 (i,j) 位置的元素为 $\frac{w-1/w}{2u}$。

Case $i = j$. 对于这种情况，我们有 $\exp(-\alpha_t\beta_{ii}\mathbf{E}_{ii}) = \mathbf{I} + \mathbf{I}_{ii}(w-1)$，因此

$$\mathcal{L}(\exp(-\alpha_t\beta_{ii}\mathbf{E}_{ii})) = \mathbf{I} + \mathbf{I}_{ii}(\sqrt{w} - 1). \tag{35}$$

综合这两种情况，我们有

$$\begin{aligned}\mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij})) = \mathbf{I} &+ \mathbf{1}_{i\neq j}\left[\mathbf{I}_{jj}(u-1) + \mathbf{I}_{ii}(\frac{1}{u} - 1) + \mathbf{I}_{ij}\frac{w - 1/w}{2u}\right] \\ &+ \mathbf{1}_{i=j}\mathbf{I}_{ii}(\sqrt{w} - 1)\end{aligned} \tag{36}$$

因此，可以按如下方式进行更新

$$\mathbf{B}_{t+1} = \mathbf{B}_t\mathbf{B}_{t+1}^{\text{up}} \tag{37}$$

其中，对任意 $1 \leq j \leq i \leq n$，$\mathbf{B}_{t+1}^{\text{up}} = \mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}^{ij}))$ 如式 (36) 所示。注意，由于 β_{ij} 可在 $\mathcal{O}(m)$ 时间内计算得出，而式 (36) 中的其他计算需要 $\mathcal{O}(1)$ 时间，因此，执行式 (37) 中更新的总体复杂度为 $\mathcal{O}(m+n)$。

3.2 多方向更新

在本节中，我们将第3.1节的坐标下降思想推广为沿多个随机选取的基向量进行更新。然而，若不加注意，这类更新未必高效。事实上，沿多个基向量更新可能会使$\mathbf{B}_{t+1}^{\text{up}}$ 通常变成稠密矩阵，这又会导致每次更新需要进行 $\mathcal{O}(n^3)$ 次计算，无法带来显著的计算优势。

为了确保可以直接使用稀疏更新矩阵 $\mathbf{B}_{t+1}^{\text{up}}$更新乔列斯基因子，我们必须沿着互不重叠的基进行更新。若 $i \neq k$, $i \neq \ell$, $j \neq k$ 和 $j \neq \ell$，或等价地， $\{i,j\} \cap \{k,\ell\} = \emptyset$（令 $\{i,i\} := \{i\}$），则称两个基向量 $\mathbf{G}_{ij}^R(\mathbf{X})$ 和 $\mathbf{G}_{k\ell}^R(\mathbf{X})$ 互不重叠。对于这类基向量对，可以看出，相应 $\mathbf{E}^{ij}$ 和 $\mathbf{E}_{k\ell}$ 的非零项位于不同的行和列。因此，沿每个这样的方向对 $\mathbf{B}_t$ 执行的更新可以并行进行，再将所得矩阵相加即可得到$\mathbf{B}_{t+1}$。

更准确地说，对于不重叠基$\mathbf{G}_{ij}^R(\mathbf{X})$ 和 $\mathbf{G}_{k\ell}^R(\mathbf{X})$，有以下结论：

$$\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij} - \alpha_t\beta_{k\ell}\mathbf{E}_{k\ell}) = \exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij}) + \exp(-\alpha_t\beta_{k\ell}\mathbf{E}_{k\ell}) - \mathbf{I} \tag{38}$$

这意味着

$$\mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij} - \alpha_t\beta_{k\ell}\mathbf{E}_{k\ell})) = \mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij})) + \mathcal{L}(\exp(-\alpha_t\beta_{k\ell}\mathbf{E}_{k\ell})) - \mathbf{I}. \tag{39}$$

因此，沿不重叠基进行更新时，累积更新矩阵$\mathbf{B}_{t+1}^{\text{up}}$ 是对应于每个基向量的单个更新矩阵之和。

在一般情况下，设 $\mathcal{E}_t := \{i_t^k, j_t^k\}_{k=1}^{K_t}$ 为一组 $1 \le K_t \le n$ 索引对，使得 $\{i_t^k, j_t^k\} \cap \{i_t^\ell, j_t^\ell\} = \emptyset$ 对所有 $1 \le k,\ \ell \le K_t$均成立。由此可知，每一对 $(\mathbf{G}_{i_t^k j_t^k}^R,\ \mathbf{G}_{i_t^\ell j_t^\ell}^R)$ 都是不重叠的。在第 t 次迭代中，我们考虑方向

$$\mathbf{S}(\mathbf{X}_t) = \sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{G}_{ij}^R(\mathbf{X}_t) = \mathbf{B}_t\Big(\sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{E}_{ij}\Big)\mathbf{B}_t^\mathsf{T} \tag{40}$$

由此得到更新

$$\mathbf{X}_{t+1} = \text{Exp}_{\mathbf{X}_t}(-\alpha_t\mathbf{S}(\mathbf{X}_t)) \tag{41}$$

$$= \mathbf{B}_t \exp\Big(-\alpha_t \sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{E}_{ij}\Big)\mathbf{B}_t^\mathsf{T} \tag{42}$$

与（33）类似。如前所述，对于不重叠基，有

$$\exp\Big(-\alpha_t \sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{E}_{ij}\Big) = \sum_{(i,j)\in\mathcal{E}_t} \exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij}) - (K_t - 1)\mathbf{I} \tag{43}$$

$$\Rightarrow \mathcal{L}\Big(\exp\Big(-\alpha_t \sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{E}_{ij}\Big)\Big) = \sum_{(i,j)\in\mathcal{E}_t} \mathcal{L}(\exp(-\alpha_t\beta_{ij}\mathbf{E}_{ij})) - (K_t - 1)\mathbf{I} =: \mathbf{B}_{t+1}^{\text{up}} \tag{44}$$

所需的多向更新为

$$\mathbf{B}_{t+1} = \mathbf{B}_t\mathbf{B}_{t+1}^{\text{up}}. \tag{45}$$

我们观察到，式 (37) 中的单向更新对应于 $K_t = 1$ （见式 (45)）。一般而言，总可以选择不重叠基，使得 $K_t \ge \lfloor \frac{n}{2} \rfloor$。在一种极端情况下，我们可以选择 $\mathcal{E}_t$，使得对于所有 $1 \le k \le K_t$，都有 $i_t^k \neq j_t^k$，从而 $K_t = \lfloor \frac{n}{2} \rfloor$。在另一种极端情况下，如果 $\mathcal{E}_t$ 满足对于所有 $1 \le k \le K_t$，都有 $i_t^k = j_t^k$，那么我们有 $K_t = n$。

(45) 中的更新要求我们计算 $\mathbf{F}(\mathbf{X})$ 的 K_t 个项，因此对于 $K_t = \mathcal{O}(n)$ 会产生 $\mathcal{O}(mn)$ 的复杂度。接下来，由于 (44) 中的每个求和项都可以在 $\mathcal{O}(1)$ 时间内计算，$\mathbf{B}_{t+1}^{\text{up}}$ 可以在 $\mathcal{O}(n)$ 时间内计算，并且具有 $\mathcal{O}(n)$ 个非零项。最后，(45) 中的计算需要 $\mathcal{O}(n^2)$ 时间，从而得到总体复杂度为 $\mathcal{O}(mn + n^2)$。我们注意到，这些复杂度界适用于基向量随机或任意选取的情况。然而，也可以采用其他基向量选取方式，特别是，下一节将讨论贪心方法。

4 $\mathcal{F}$的黎曼子空间下降变体

如第3节所述，对于单向更新，所提算法的每次迭代复杂度为 $\mathcal{O}(m+n)$；对于多方向更新，其每次迭代复杂度为 $\mathcal{O}(mn+n^2)$。在本节中，我们将证明，对于式（2）中的函数类 $\mathcal{F}$，可以通过维护中间变量，以 $m=\mathcal{O}(n)$ 的复杂度高效地计算$\mathbf{F(X)}$的各个元素。因此，将所提RRSD算法应用于 $\mathcal{F}$ 中的函数时，其最坏情况下的每次迭代复杂度在单向情形下变为 $\mathcal{O}(n)$，在多向情形下变为 $\mathcal{O}(n^2)$。

我们首先分析 $\mathcal{F}$ 中函数的一些性质。为简洁起见，本文其余部分均假设所有矩阵都是对称的。 $f\in\mathcal{F}$ 的梯度为

$$\begin{aligned}\operatorname{grad} f(\mathbf{X}) = &-\sum_{p=1}^{P}\frac{dg}{dg_{1,p}}\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}+\sum_{q=1}^{Q}\frac{dg}{dg_{2,q}}\mathbf{D}_q+\frac{dg}{dg_3}\mathbf{X}^{-1}\\ &+\sum_{r=1}^{R}\frac{dg}{dg_{4,r}}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r+\mathbf{H}_r\mathbf{X}\mathbf{A}_r\right)-\sum_{s=1}^{S}\frac{dg}{dg_{5,s}}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}+\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\right)\\ &+\frac{1}{2}\sum_{m=1}^{M}\frac{dg}{dg_{6,m}}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m+\mathbf{P}_m\mathbf{X}^{-1}\mathbf{Q}_m-\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}-\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\right)\end{aligned} \tag{46}$$

其中，$g_{1,p}=\mathrm{tr}\left(\mathrm{C}_p\mathrm{X}^{-1}\right)$、$g_{2,q}=\mathrm{tr}\left(\mathrm{D}_q\mathrm{X}\right)$、$g_3=\log\det\mathrm{X}$、$g_{4,r}=\mathrm{tr}\left(\mathrm{A}_r\mathrm{X}\mathrm{H}_r\mathrm{X}\right)$、$g_{5,s}=\mathrm{tr}\left(\mathrm{X}^{-1}\mathrm{F}_s\mathrm{X}^{-1}\mathrm{G}_s\right)$ 以及 $g_{6,m}=\mathrm{tr}\left(\mathrm{P}_m\mathrm{X}\mathrm{Q}_m\mathrm{X}^{-1}\right)$。将其代入 β_{ij} 的表达式，可得

$$\begin{aligned}\beta_{ij}&=\sqrt{2}^{\mathbf{1}_{i\neq j}}[\mathbf{F}(\mathbf{X})]_{ij}=\operatorname{tr}\left(\operatorname{grad}f(\mathbf{X})\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right) \qquad (47)\\ &=-\sum_{p=1}^{P}\frac{dg}{dg_{1,p}}\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)+\sum_{q=1}^{Q}\frac{dg}{dg_{2,q}}\operatorname{tr}\left(\mathbf{D}_q\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)+\frac{dg}{dg_3}\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)\\ &+\sum_{r=1}^{R}\frac{dg}{dg_{4,r}}\operatorname{tr}\left(\left[\mathbf{A}_r\mathbf{X}\mathbf{H}_r+\mathbf{H}_r\mathbf{X}\mathbf{A}_r\right]\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)\\ &-\sum_{s=1}^{S}\frac{dg}{dg_{5,s}}\operatorname{tr}\left(\left[\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}+\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\right]\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)\\ &+\frac{1}{2}\sum_{m=1}^{M}\frac{dg}{dg_{6,m}}\operatorname{tr}\left(\begin{pmatrix}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m-\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\\ +\mathbf{P}_m\mathbf{X}^{-1}\mathbf{Q}_m-\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\end{pmatrix}\mathbf{B}\mathbf{E}_{ij}\mathbf{B}^\mathsf{T}\right)\\ &=-\sum_{p=1}^{P}\frac{dg}{dg_{1,p}}\operatorname{tr}\left(\mathbf{B}^{-1}\mathbf{C}_p\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right)+\sum_{q=1}^{Q}\frac{dg}{dg_{2,q}}\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{D}_q\mathbf{B}\mathbf{E}_{ij}\right)+\frac{dg}{dg_3}\operatorname{tr}\left(\mathbf{E}_{ij}\right)\\ &+\sum_{r=1}^{R}\frac{dg}{dg_{4,r}}\left(\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{A}_r\mathbf{B}\mathbf{B}^\mathsf{T}\mathbf{H}_r\mathbf{B}\mathbf{E}_{ij}\right)+\operatorname{tr}\left(\mathbf{B}^\mathsf{T}\mathbf{H}_r\mathbf{B}\mathbf{B}^\mathsf{T}\mathbf{A}_r\mathbf{B}\mathbf{E}_{ij}\right)\right)\end{aligned}$$

$$
\begin{aligned}
&-\sum_{s=1}^{S}\frac{dg}{dg_{5,s}}\left(\mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right)+\mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right)\right)\\
&+\frac{1}{2}\sum_{m=1}^{M}\frac{dg}{dg_{6,m}}\begin{pmatrix}\mathrm{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\mathbf{E}_{ij}\right)-\mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right)\\ +\mathrm{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{P}_m\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{Q}_m\mathbf{B}\mathbf{E}_{ij}\right)-\mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{Q}_m\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{P}_m\mathbf{B}^{-\mathsf{T}}\mathbf{E}_{ij}\right)\end{pmatrix}\\
&=-\sqrt{2}^{\mathbf{1}_{i\neq j}}\sum_{p=1}^{P}\frac{dg}{dg_{1,p}}[\mathbf{B}^{-1}\mathbf{C}_p\mathbf{B}^{-\mathsf{T}}]_{ij}+\sqrt{2}^{\mathbf{1}_{i\neq j}}\sum_{q=1}^{Q}\frac{dg}{dg_{2,q}}[\mathbf{B}^{\mathsf{T}}\mathbf{D}_q\mathbf{B}]_{ij}+\frac{dg}{dg_3}\mathbf{1}_{i=j}\\
&+\sqrt{2}^{\mathbf{1}_{i\neq j}}\sum_{r=1}^{R}\frac{dg}{dg_{4,r}}\left(\left[\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_r\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{H}_r\mathbf{B}\right]_{ij}+\left[\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_r\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{H}_r\mathbf{B}\right]_{ji}\right)\\
&-\sqrt{2}^{\mathbf{1}_{i\neq j}}\sum_{s=1}^{S}\frac{dg}{dg_{5,s}}\left(\left[\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\right]_{ij}+\left[\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\right]_{ji}\right)\\
&\frac{1}{\sqrt{2}}\sum_{m=1}^{M}\frac{dg}{dg_{6,m}}\begin{pmatrix}\left[\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\right]_{ij}+\left[\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\right]_{ji}\\ -\left[\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\right]_{ij}-\left[\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\right]_{ji}\end{pmatrix}
\end{aligned}
\tag{48}
$$

这里需要注意，β_{ij} 通过 $\mathbf{X}_t$ 取决于

$$
\mathbf{M}_{1,p}(\mathbf{X}_t)=\mathbf{B}_t^{-1}\mathbf{C}_p\mathbf{B}_t^{-\mathsf{T}} \qquad \mathbf{M}_{2,q}(\mathbf{X}_t)=\mathbf{B}_t^{\mathsf{T}}\mathbf{D}_q\mathbf{B}_t \tag{49a}
$$

$$
\mathbf{M}_{4,1,r}(\mathbf{X}_t)=\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_r\mathbf{B}_t \qquad \mathbf{M}_{4,2,r}(\mathbf{X}_t)=\mathbf{B}_t^{\mathsf{T}}\mathbf{H}_r\mathbf{B}_t \tag{49b}
$$

$$
\mathbf{M}_{5,1,s}(\mathbf{X}_t)=\mathbf{B}_t^{-1}\mathbf{F}_s\mathbf{B}_t^{-\mathsf{T}} \qquad \mathbf{M}_{5,2,s}(\mathbf{X}_t)=\mathbf{B}_t^{-1}\mathbf{G}_s\mathbf{B}_t^{-\mathsf{T}} \tag{49c}
$$

$$
\mathbf{M}_{6,1,m}(\mathbf{X})=\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B} \qquad \mathbf{M}_{6,2,m}(\mathbf{X})=\mathbf{B}^{-1}\mathbf{Q}_m\mathbf{B} \tag{49d}
$$

以及通过 $g_{1,\ p}$, $g_{2,\ q}$, g_3, $g_{4,\ r}$, $g_{5,\ s}$ 和 $g_{6,\ m}$，而这些又可以计算为

$$
g_{1,p}(\mathbf{X}_t)=\mathrm{tr}\left(\mathbf{C}_p\mathbf{X}_t^{-1}\right)=\mathrm{tr}\left(\mathbf{B}_t^{-1}\mathbf{C}_p\mathbf{B}_t^{-\mathsf{T}}\right)=\mathrm{tr}\left(\mathbf{M}_{1,p}(\mathbf{X}_t)\right) \tag{50}
$$

$$
g_{2,q}(\mathbf{X}_t)=\mathrm{tr}\left(\mathbf{D}_q\mathbf{X}_t\right)=\mathrm{tr}\left(\mathbf{B}_t^{\mathsf{T}}\mathbf{D}_q\mathbf{B}_t\right)=\mathrm{tr}\left(\mathbf{M}_{2,q}(\mathbf{X}_t)\right) \tag{51}
$$

$$
g_3(\mathbf{X}_t)=\log\det\mathbf{X}_t=2\log\det\left(\mathbf{B}_t\right)=2\log\det\left(\mathbf{B}_{t-1}\right)+2\log\det\left(\mathbf{B}_t^{\mathrm{up}}\right) \tag{52}
$$

$$
g_{4,r}(\mathbf{X}_t)=\mathrm{tr}\left(\mathbf{X}\mathbf{A}_r\mathbf{X}\mathbf{H}_r\right)=\mathrm{tr}\left(\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_r\mathbf{B}_t\mathbf{B}_t^{\mathsf{T}}\mathbf{H}_r\mathbf{B}_t\right)=\mathrm{tr}\left(\mathbf{M}_{4,1,r}(\mathbf{X}_t)\mathbf{M}_{4,2,r}(\mathbf{X}_t)\right) \tag{53}
$$

$$
g_{5,s}(\mathbf{X}_t)=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\right)=\mathrm{tr}\left(\mathbf{B}_t^{-1}\mathbf{F}_s\mathbf{B}_t^{-\mathsf{T}}\mathbf{B}_t^{-1}\mathbf{G}_s\mathbf{B}_t^{-\mathsf{T}}\right)=\mathrm{tr}\left(\mathbf{M}_{5,1,s}(\mathbf{X}_t)\mathbf{M}_{5,2,s}(\mathbf{X}_t)\right) \tag{54}
$$

$$
g_{6,m}(\mathbf{X})=\mathrm{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\right)=\mathrm{tr}\left(\mathbf{B}^{-1}\mathbf{P}_m\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_m\mathbf{B}^{-\mathsf{T}}\right)=\mathrm{tr}\left(\mathbf{M}_{6,1,m}(\mathbf{X})\left[\mathbf{M}_{6,2,m}(\mathbf{X})\right]^{\mathsf{T}}\right) \tag{55}
$$

由于 $\mathbf{B}$ 和 $\mathbf{B}_t^{\mathrm{up}}$ 均为下三角矩阵，因此可以在$\mathcal{O}(n)$时间内计算它们的行列式。此外，我们可以通过观察以下内容来维护 $\{\mathbf{M}_{1,p}(\mathbf{X}_t)\}_{1\leq p\leq P}$, $\{\mathbf{M}_{4,1,r}(\mathbf{X}_t), \mathbf{M}_{4,2,r}(\mathbf{X}_t)\}_{1\leq r\leq R}$, $\{\ \ _{2\,q\ \ t}\}_{1\leq q\leq Q}$ $\{\ \ _{5\,1\,s\ \ t},\ \ _{5\,2\,s\ \ t}\}_{1\leq s\leq S}$ $\{\ \ _{6\,1\,m\ \ t},\ \ _{6\,2\,m\ \ t}\}_{1\leq m\leq M}$ $\mathbf{M}_{,,}(\mathbf{X})$, $\mathbf{M}_{,,}(\mathbf{X})$ $\mathbf{M}_{,,}(\mathbf{X})$ and $\mathbf{M}_{,,}(\mathbf{X})$ $\mathbf{M}_{,,}(\mathbf{X})$

$$
\mathbf{M}_{1,p}(\mathbf{X}_{t+1})=\mathbf{B}_{t+1}^{-1}\mathbf{C}_p\mathbf{B}_{t+1}^{-\mathsf{T}}=\left[\mathbf{B}_{t+1}^{\mathrm{up}}\right]^{-1}\mathbf{B}_t^{-1}\mathbf{C}_p\mathbf{B}_t^{-\mathsf{T}}\left[\mathbf{B}_{t+1}^{\mathrm{up}}\right]^{-\mathsf{T}}
$$

$$= \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{M}_{1,p}(\mathbf{X}_t) \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-\mathsf{T}} \tag{56}$$

$$\mathbf{M}_{2,q}(\mathbf{X}_{t+1}) = \mathbf{B}_{t+1}^{\mathsf{T}} \mathbf{D}_q \mathbf{B}_{t+1} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{B}_t^{\mathsf{T}} \mathbf{D}_q \mathbf{B}_t \mathbf{B}_{t+1}^{\text{up}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{M}_{2,q}(\mathbf{X}_t) \mathbf{B}_{t+1}^{\text{up}} \tag{57}$$

$$\mathbf{M}_{4,1,r}(\mathbf{X}_t) = \mathbf{B}_{t+1}^{\mathsf{T}} \mathbf{A}_r \mathbf{B}_{t+1} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{B}_t^{\mathsf{T}} \mathbf{A}_r \mathbf{B}_t \mathbf{B}_{t+1}^{\text{up}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{M}_{4,1,r}(\mathbf{X}_t) \mathbf{B}_{t+1}^{\text{up}} \tag{58}$$

$$\mathbf{M}_{4,2,r}(\mathbf{X}_t) = \mathbf{B}_{t+1}^{\mathsf{T}} \mathbf{H}_r \mathbf{B}_{t+1} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{B}_t^{\mathsf{T}} \mathbf{H}_r \mathbf{B}_t \mathbf{B}_{t+1}^{\text{up}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{\mathsf{T}} \mathbf{M}_{4,2,r}(\mathbf{X}_t) \mathbf{B}_{t+1}^{\text{up}} \tag{59}$$

$$\begin{aligned}\mathbf{M}_{5,1,s}(\mathbf{X}_{t+1}) &= \mathbf{B}_{t+1}^{-1} \mathbf{G}_s \mathbf{B}_{t+1}^{-\mathsf{T}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{B}_t^{-1} \mathbf{G}_s \mathbf{B}_t^{-\mathsf{T}} \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-\mathsf{T}} \\ &= \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{M}_{5,1,s}(\mathbf{X}_t) \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-\mathsf{T}} \end{aligned} \tag{60}$$

$$\begin{aligned}\mathbf{M}_{5,\ s}(\mathbf{X}_{t+1}) &= \mathbf{B}_{t+1}^{-1} \mathbf{H}_s \mathbf{B}_{t+1}^{-\mathsf{T}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{B}_t^{-1} \mathbf{H}_s \mathbf{B}_t^{-\mathsf{T}} \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-\mathsf{T}} \\ &= \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{M}_{5,2,s}(\mathbf{X}_t) \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-\mathsf{T}} \end{aligned} \tag{61}$$

$$\mathbf{M}_{6,1,m}(\mathbf{X}_{t+1}) = \mathbf{B}_{t+1}^{-1} \mathbf{P}_m \mathbf{B}_{t+1} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{B}_t^{-1} \mathbf{P}_m \mathbf{B}_t \mathbf{B}_{t+1}^{\text{up}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{M}_{6,1,m}(\mathbf{X}_t) \mathbf{B}_{t+1}^{\text{up}} \tag{62}$$

$$\mathbf{M}_{6,2,m}(\mathbf{X}_{t+1}) = \mathbf{B}_{t+1}^{-1} \mathbf{Q}_m \mathbf{B}_{t+1} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{B}_t^{-1} \mathbf{Q}_m \mathbf{B}_t \mathbf{B}_{t+1}^{\text{up}} = \left[\mathbf{B}_{t+1}^{\text{up}}\right]^{-1} \mathbf{M}_{6,1,m}(\mathbf{X}_t) \mathbf{B}_{t+1}^{\text{up}} \tag{63}$$

其中，$\mathbf{B}_{t+1}^{\text{up}}$ 是一个取决于更新所使用的基向量的稀疏矩阵。$\mathbf{B}_{t+1}^{\text{up}}$ 的逆矩阵可以通过对角元素取倒数，并将非对角元素替换为$\mathbf{B}_{t+1}^{\text{up}}$中对应值的负数来获得，即，

$$[[\mathbf{B}_{t+1}^{\text{up}}]^{-1}]_{ij} = \frac{\mathbf{1}_{i=j}}{[\mathbf{B}_{t+1}^{\text{up}}]_{ij}} - \mathbf{1}_{i \neq j} [\mathbf{B}_{t+1}^{\text{up}}]_{ij} \tag{64}$$

对于所有的 $1 \leq j \leq i \leq n$。更新的总体复杂度取决于更新方向的数量 K_t。

4.1 单向更新

对于单方向更新情况，在每次迭代中，中间变量 $\mathbf{M}1,_p(\mathbf{X}_t)$、$\mathbf{M}2,_q(\mathbf{X}_t)$、$\mathbf{M}4,1,_r(\mathbf{X}_t)$、$\mathbf{M}4,2,_r(\mathbf{X}_t)$、$\mathbf{M}5,1,_s(\mathbf{X}_t)$、$\mathbf{M}5,2,_s(\mathbf{X}_t)$、$\mathbf{M}6,1,_m(\mathbf{X}_t)$、$\mathbf{M}6,2,_m(\mathbf{X}_t)$ 以及 $\mathbf{B}_t$ 至多更新两行和两列，因此，式 (56)–(63) 中更新的复杂度为 $\mathcal{O}(n)$。给定变量 $\mathbf{M}1,_p(\mathbf{X}_t)$、$\mathbf{M}2,_q(\mathbf{X}_t)$，函数 $f_1,_p(\mathbf{X}_t)$、$f_2,_q(\mathbf{X}_t)$ 以及 $f_3(\mathbf{X}_t)$ 可以在 $O(n)$ 内计算完成。一个有趣的观察是，给定中间变量 $\mathbf{M}4,1,_r(\mathbf{X}_t)$、$\mathbf{M}4,2,_r(\mathbf{X}_t)$、$\mathbf{M}5,1,_s(\mathbf{X}_t)$、$\mathbf{M}5,2,_s(\mathbf{X}_t)$、$\mathbf{M}6,1,_m(\mathbf{X}_t)$ 和 $\mathbf{M}6,2,_m(\mathbf{X}_t)$，通过在函数计算中调整修改后的 $O(n)$元素的贡献，$g_4,_r(\mathbf{X}_t)$、$g_5,_s(\mathbf{X}_t)$ 和 $g_6,_m(\mathbf{X}_t)$ 的计算同样可以在 $O(n)$ 内完成。因此可以推得，β_{ij} 可以在 $\mathcal{O}(n)$ 时间内计算完成。最后，式 (37) 中的更新需要进行 $\mathcal{O}(n)$ 次计算，因此，对于 $f \in \mathcal{F}$，单向更新的总体复杂度为 $\mathcal{O}(n)$，显著优于 RGD 的 $\mathcal{O}(n^3)$ 复杂度。

我们指出，所提算法的所有变体，以及当前最先进的梯度下降算法及其变体，都需要 $\mathcal{O}(n^2)$的内存复杂度。然而，有趣的是，在单向情形下，每次迭代仅更新 $\mathbf{B}^{t+1}$ 和其他中间变量中的 $\mathcal{O}(n)$ 个条目。因此，我们可以在单向情形下执行一次更新，同时仅使用随机存取存储器（RAM）中的 $\mathcal{O}(n)$ 空间。

4.2 多方向更新

对于多向情况，可以看出，在每次迭代中，M1,$_p$(X$_t$)、M2,$_q$(X$_t$)、M4,1,$_r$(X$_t$)、M4,2,$_r$(X$_t$)、M5,1,$_s$(X$_t$)、M5,2,$_s$(X$_t$)、M6,1,$_m$(X$_t$)、M6,2,$_m$(X$_t$) 和 B$_t$ 的 $\mathcal{O}(n)$ 行和列都会更新。因此，(56)-(63) 中更新的复杂度为$\mathcal{O}(n^2)$。因此，计算 β_{ij} 的计算复杂度也变为 $\mathcal{O}(n^2)$。最后，对于这种情况，(45) 中更新计算的复杂度为 $\mathcal{O}(n^2)$，整体复杂度为 $\mathcal{O}(n^2)$。

4.3 贪心子空间下降

到目前为止，所提出的黎曼子空间下降框架的所有版本都使用随机选取的基向量，这与随机坐标下降算法类似（Nesterov（2012年））。我们指出，也可以依据某些准则或启发式方法来选择基向量。例如，可以选择与 $|\beta_{ij}|$的最大值相对应的基向量，这些基向量对应于最速下降方向。然而，为了选出对应于 $|\beta_{ij}|$最大值的索引（$i,\ j$），必须针对每个 t计算 **F(X**$_t$**)**的所有元素。因此，单方向更新的复杂度会从 $\mathcal{O}(m+n)$增加到 $\mathcal{O}(M+n)$；即使 $f \in \mathcal{F}$ with$M = \mathcal{O}(n^2)$，这也会带来显著影响。因此，在采用单方向更新时，所提出的启发式方法并无计算优势。

在多方向情形下，要选择互不重叠的基向量以确保选出 $|\beta_{ij}|$ 的最大值，则更具挑战性。为此，我们提出一种贪心启发式方法，按如下方式选择基向量。首先，按照绝对值从大到小对**F(X)**下三角部分的元素排序，并保存所得顺序。第一步选取最大的元素，并移除与之重叠的元素。对剩余元素重复这一过程，直至所有元素均处理完毕，且所有行和列均已移除。可以看出，贪心启发式方法最终会选出**F(X)**中 $\lfloor n/2 \rfloor \leq K_t \leq n$ 个对应于互不重叠基向量的元素。

所提出的 RGSD 算法的计算复杂度并不比 RRSD 高出太多。一般而言，计算完整矩阵 **F(X)**需要 $\mathcal{O}(M)$ 时间，随后可在 $\mathcal{O}(n^2\log(n))$ 时间内按绝对值对所得元素进行排序，而用于选择不重叠基的贪心启发式还需要额外的 $\mathcal{O}(n^2)$ 时间。因此，RGSD 的总体复杂度为 $\mathcal{O}(M + n^2\log(n))$；当 $f \in \mathcal{F}$ 时，复杂度则为 $\mathcal{O}(n^2\log(n))$。

表 2 总结了梯度下降以及所提出的子空间下降算法每次迭代的时间复杂度。

4.4 有限和问题的黎曼随机子空间下降 **s**

最后，我们将拟议的 RRSD 算法与 RSGD 算法（5）进行比较，并讨论其对于有限和问题的适用性。考虑以下有限和问题

$$\min_{\mathbf{X}\in\mathbb{P}^n} f(\mathbf{X}) := \sum_{s=1}^{S} f_s(\mathbf{X}); \qquad f_s(\mathbf{X}) \in \mathcal{F} \tag{65}$$

算法	M	m	复杂度/迭代
RGD	$\mathcal{O}(n^3)$	$\mathcal{O}\ (n^3)$	$\mathcal{O}(n^3)$
RRSD（单向，一般）	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$
RRSD（多向，一般）	$\mathcal{O}(n^3)$	$\mathcal{O}(n^3)$	$\mathcal{O}\ (n^3)$
RRSD（单向，$f \in \mathcal{F}$）	$\mathcal{O}\ (n^2)$	$\mathcal{O}(n)$	$\mathcal{O}(n)$
RGSD（单向，$f \in \mathcal{F}$）	$\mathcal{O}\ (n^2)$	$\mathcal{O}(n)$	$\mathcal{O}\ (n^2)$
RRSD（多向，$f \in \mathcal{F}$）	$\mathcal{O}\ (n^2)$	$\mathcal{O}(n)$	$\mathcal{O}\ (n^2)$
RGSD（多向，$f \in \mathcal{F}$）	$\mathcal{O}\ (n^2)$	$\mathcal{O}(n)$	$\mathcal{O}\ (n^2 \log n)$

表 2：算法每次迭代的时间复杂度比较

假设每个 f_s 具有以下函数形式：

$$f_s(\mathbf{X}) = g\left(\begin{array}{c} \operatorname{tr}\left(\mathbf{C}_s\mathbf{X}^{-1}\right), \operatorname{tr}\left(\mathbf{D}_s\mathbf{X}\right), \log\det\mathbf{X}, \operatorname{tr}\left(\mathbf{X}\mathbf{A}_s\mathbf{X}\mathbf{H}_s\right), \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\right), \\ \operatorname{tr}\left(\mathbf{P}_s\mathbf{X}\mathbf{Q}_s\mathbf{X}^{-1}\right) \end{array}\right) \tag{66}$$

在实现 RSGD (5) 时，我们从索引集合 $\{1, 2, \ldots, S\}$ 中均匀随机选取一个索引 s，并沿函数 $f_s(\mathbf{X})$ 的黎曼梯度的负方向迈出一步。$f_s(\mathbf{X})$ 的黎曼梯度为

$$\begin{aligned} \operatorname{grad} f(\mathbf{X}) = &-\frac{dg}{dg_1}\mathbf{X}^{-1}\mathbf{C}_s\mathbf{X}^{-1} + \frac{dg}{dg_2}\mathbf{D}_s + + \frac{dg}{dg_3}\mathbf{X}^{-1} + \frac{dg}{dg_4}\left(\mathbf{A}_s\mathbf{X}\mathbf{H}_s + \mathbf{H}_s\mathbf{X}\mathbf{A}_s\right) \\ &- \frac{dg}{dg_5}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\right) \\ &+ \frac{1}{2}\frac{dg}{dg_6}\left(\mathbf{Q}_s\mathbf{X}^{-1}\mathbf{P}_s + \mathbf{P}_s\mathbf{X}^{-1}\mathbf{Q}_s - \mathbf{X}^{-1}\mathbf{P}_s\mathbf{X}\mathbf{Q}_s\mathbf{X}^{-1} - \mathbf{X}^{-1}\mathbf{Q}_s\mathbf{X}\mathbf{P}_s\mathbf{X}^{-1}\right) \end{aligned} \tag{67}$$

其中，dg/dg_i 表示 g 对其第 i 个参数的导数。最后，更新可写为

$$\mathbf{X}_{t+1} = \mathbf{B}_t \exp\left(\alpha_t \begin{bmatrix} \frac{dg}{dg_1}\mathbf{B}_t^{-1}\mathbf{C}_s\mathbf{B}_t^{-\mathsf{T}} - \frac{dg}{dg_2}\mathbf{B}_t^{\mathsf{T}}\mathbf{D}_s\mathbf{B}_t - \frac{dg}{dg_3}\mathbf{I} \\ -\frac{dg}{dg_4}\left(\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_r\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{H}_r\mathbf{B} + \mathbf{B}_t^{\mathsf{T}}\mathbf{H}_r\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{A}_r\mathbf{B}\right) \\ +\frac{dg}{dg_5}\left(\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}} + \mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\right) \\ -\frac{1}{2}\frac{dg}{dg_6}\begin{pmatrix} \mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B} + \mathbf{B}^{\mathsf{T}}\mathbf{P}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{Q}_s\mathbf{B} \\ -\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}} - \mathbf{B}^{-1}\mathbf{Q}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{P}_s\mathbf{B}^{-\mathsf{T}} \end{pmatrix} \end{bmatrix}\right)\mathbf{B}_t^{\mathsf{T}} \tag{68}$$

这会产生 $\mathcal{O}(n^3)$ 的时间复杂度，并且需要对矩阵 $\mathbf{B}_t$ 求逆。相比之下，对于 RRSD 算法，我们有：

$$\beta_{ij}(\mathbf{X}_t) = \sum_{s=1}^{S}\begin{pmatrix} -\sqrt{2}^{\mathbf{1}_{i\neq j}}\frac{dg}{dg_1}[\mathbf{B}_t^{-1}\mathbf{C}_s\mathbf{B}_t^{-\mathsf{T}}]_{ij} + \sqrt{2}^{\mathbf{1}_{i\neq j}}\frac{dg}{dg_2}[\mathbf{B}^{\mathsf{T}}\mathbf{D}_s\mathbf{B}]_{ij} + \frac{dg}{dg_3}\mathbf{1}_{i\neq j} \\ +\sqrt{2}^{\mathbf{1}_{i\neq j}}\frac{dg}{dg_{4,r}}\left(\left[\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{H}_s\mathbf{B}\right]_{ij} + \left[\mathbf{B}_t^{\mathsf{T}}\mathbf{A}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{H}_s\mathbf{B}\right]_{ji}\right) \\ -\sqrt{2}^{\mathbf{1}_{i\neq j}}\frac{dg}{dg_5}\left(\left[\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\right]_{ij} + \left[\mathbf{B}^{-1}\mathbf{F}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{G}_s\mathbf{B}^{-\mathsf{T}}\right]_{ji}\right) \\ +\frac{1}{\sqrt{2}^{\mathbf{1}_{i\neq j}}}\frac{dg}{dg_6}\begin{pmatrix} \left[\mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B}\right]_{ij} + \left[\mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}}\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B}\right]_{ji} \\ -\left[\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}}\right]_{ij} - \left[\mathbf{B}^{-1}\mathbf{P}_s\mathbf{B}\mathbf{B}^{\mathsf{T}}\mathbf{Q}_s\mathbf{B}^{-\mathsf{T}}\right]_{ji} \end{pmatrix} \end{pmatrix} \tag{69}$$

回顾式（49）中 $\mathrm{M}1,s(\mathrm{X}_t)$、$\mathrm{M}2,s(\mathrm{X}_t)$、$\mathrm{M}4,1,s(\mathrm{X}_t)$、$\mathrm{M}4,2,s(\mathrm{X}_t)$、$\mathrm{M}5,1,s(\mathrm{X}_t)$、$\mathrm{M}5,2,s(\mathrm{X}_t)$、$\mathrm{M}6,1,s(\mathrm{X}_t)$ 和 $\mathrm{M}6,2,s(\mathrm{X}_t)$ 的定义，我们可以写出与式（56）–（63）类似的递归更新规则。此外，按照第3节中的说明，可以计算 $\mathrm{B}_{t+1}^{\mathrm{up}}$ ，从而得到更新 $\mathrm{B}_{t+1} = \mathrm{B}_t \mathrm{B}_{t+1}^{\mathrm{up}}$。对于单向和多向两种情况，在每次迭代中，维护这些中间变量并更新 B_{t+1} 的总复杂度分别为 $\mathcal{O}(Sn)$ 和 $\mathcal{O}(Sn^2)$。因此，与 RSGD 相比，当 $S = o(n^2)$ 时，所提出的单向 RRSD 具有更低的每次迭代复杂度，而多向 RRSD 在 $S = o(n)$ 时具有更低的复杂度。进一步指出，对于单向更新，在每次迭代中，仅更新 $\{\mathrm{M}1,s(\mathrm{X}_t)、\mathrm{M}2,s(\mathrm{X}_t)、\mathrm{M}4,1,s(\mathrm{X}_t)、\mathrm{M}4,2,s(\mathrm{X}_t)、\mathrm{M}5,1,s(\mathrm{X}_t)、\mathrm{M}5,2,s(\mathrm{X}_t)、\mathrm{M}6,1,s(\mathrm{X}_t)、\mathrm{M}6,2,s(\mathrm{X}_t)\}_{s=1}^{S}$ 的两行和两列，因此该更新在 RAM 中仅需要 $\mathcal{O}(Sn)$ 的空间。

5 收敛分析

本节将刻画所提算法达到 ϵ 精度所需的迭代复杂度。一般而言，$\mathbf{X}_t$ 如果满足 $f(\mathbf{X}_t) \leq f(\mathbf{X}^\star) + \epsilon$，则称其为 ϵ 精确。然而，对于所提算法的随机化变体，我们只要求 $\mathbb{E}[f(\mathbf{X}_t)] \leq f(\mathbf{X}^\star) + \epsilon$。我们证明，在假设 **A1** 和 **A2** 下，所有所提算法都呈现线性收敛。然而，RRSD 和 RGSD 的单向变体通常需要其多方向对应算法所需迭代次数的 $\mathcal{O}(n)$ 倍。后续界将取决于条件数 $\kappa := L/\mu$ 以及通过 $D_0 := f(\mathbf{X}_0) - f(\mathbf{X}^\star)$ 给出的初始化。对于所有算法，我们设步长为 $\alpha = 1/L$。

5.1 单方向更新

定理 5 在假设 ***A1*** 和 ***A2*** 下，单向 *RRSD* 算法的迭代复杂度为 $\mathcal{O}(n^2 \kappa \log(\frac{D_0}{\epsilon}))$。

—

证明 在第 t 次迭代中，$\mathbf{G}^R_{i_t j_t}(\mathbf{X}_t)$ 从 $d = \frac{n(n+1)}{2}$ 个可能的典范基方向中均匀随机选取。回顾第 2 节中 $\boldsymbol{\xi}_{\mathbf{XY}}$ 的定义，根据 (32) 中的更新可知 $\boldsymbol{\xi}_{\mathbf{X}_t \mathbf{X}_{t+1}} = -\frac{1}{L} \beta_{i_t j_t} \mathbf{G}^R_{i_t j_t}(\mathbf{X}_t)$。因此，由 (14) 可得

$$f(\mathbf{X}_{t+1}) \leq f(\mathbf{X}_t) - \frac{\beta^2_{i_t j_t}}{L} + \frac{\beta^2_{i_t j_t}}{L^2} \frac{L}{2} = f(\mathbf{X}_t) - \frac{\beta^2_{i_t j_t}}{2L}$$

由于 (i_t, j_t) 独立同分布，给定 $\{\mathbf{X}_\tau\}_{\tau=1}^t$ 取条件期望，我们得到。

$$\mathbb{E}_t[f(\mathbf{X}_{t+1})] \leq f(\mathbf{X}_t) - \frac{1}{2dL} \sum_{1 \leq j \leq i \leq n} \beta^2_{ij}(\mathbf{X}_t) \tag{70}$$

$$= f(\mathbf{X}_t) - \frac{1}{2dL} \|\mathrm{grad}^R f(\mathbf{X}_t)\|^2_{\mathbf{X}_t} \tag{71}$$

其中， $\mathbb{E}_t[\cdot]$ 表示关于（i_t, j_t）的期望，式（71）由式（26）中 $\text{grad}^R f(\mathbf{X}_t)$ 的正交分解得到。将 $\mathbf{Y} = \mathbf{X}^\star$ 代入式（17）并重新整理，可得

$$\langle \text{grad}^R f(\mathbf{X}), -\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star} \rangle_{\mathbf{X}} \geq f(\mathbf{X}) - f(\mathbf{X}^\star) + \frac{\mu}{2} \|\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star}\|_{\mathbf{X}}^2. \tag{72}$$

由于 $f(\mathbf{X}) - f(\mathbf{X}^\star) \geq 0$，由柯西-施瓦茨不等式可得

$$\|\text{grad}^R f(\mathbf{X})\|_{\mathbf{X}} \|\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star}\|_{\mathbf{X}} \geq \frac{\mu}{2} \|\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star}\|_{\mathbf{X}}^2$$

这意味着

$$\|\text{grad}^R f(\mathbf{X})\|_{\mathbf{X}} \geq \frac{\mu}{2} \|\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star}\|_{\mathbf{X}} \tag{73}$$

因为 $\mathbf{X} \neq \mathbf{X}^\star$。再次由 (72) 可得

$$\Rightarrow \|\text{grad}^R f(\mathbf{X})\|_{\mathbf{X}} \|\boldsymbol{\xi}_{\mathbf{X}\mathbf{X}^\star}\|_{\mathbf{X}} \geq f(\mathbf{X}) - f(\mathbf{X}^\star) \tag{74}$$

结合 (73)，可得

$$\frac{2}{\mu} \|\text{grad}^R f(\mathbf{X})\|_{\mathbf{X}}^2 \geq f(\mathbf{X}) - f(\mathbf{X}^\star) \tag{75}$$

结合 (71) 和 (75) 中的界，可得

$$\begin{aligned} \mathbb{E}_t[f(\mathbf{X}_{t+1})] &\leq f(\mathbf{X}_t) - \frac{1}{2dL} \frac{\mu}{2} (f(\mathbf{X}_t) - f(\mathbf{X}^\star)) \\ \Rightarrow \mathbb{E}_t[f(\mathbf{X}_{t+1})] - f(\mathbf{X}^\star) &\leq f(\mathbf{X}_t) - f(\mathbf{X}^\star) - \frac{\mu}{4dL} (f(\mathbf{X}_t) - f(\mathbf{X}^\star)) \qquad (76) \\ &= \left(1 - \frac{\mu}{4dL}\right) \left[f(\mathbf{X}_t) - f(\mathbf{X}^\star)\right]. \qquad (77) \end{aligned}$$

对两边取全期望并继续递推，可得所需的线性收敛界

$$\mathbb{E}[f(\mathbf{X}_{t+1})] - f(\mathbf{X}^\star) \leq \left(1 - \frac{\mu}{4dL}\right)^{t+1} \left[f(\mathbf{X}_0) - f(\mathbf{X}^\star)\right]. \tag{78}$$

等价地，为确保 $\mathbb{E}[f(\mathbf{X}_T)] - f(\mathbf{X}^\star) \leq \epsilon$，我们要求

$$T = \mathcal{O}\left(n^2 \kappa \log\left(\frac{D_0}{\epsilon}\right)\right) \tag{79}$$

对于 $d \gg 1$。 ■

请注意，与梯度下降相比，所提出的单向子空间下降算法所需的迭代次数是其 $\mathcal{O}(n^2)$ 倍，因为在每次迭代中，我们选择沿可能的 $\mathcal{O}(n^2)$ 个方向中的单个方向前进。对于 RGSD 算法，也可以得到类似的结果，如下面的推论所述。

推论 6 在假设 ***A1*** 和 ***A2***下，单向 *RGSD* 算法的迭代复杂度为 $\mathcal{O}(n^2 \kappa \log(\frac{D_0}{\epsilon}))$。

—

证明 回顾一下，在每次迭代中，RGSD 都会选择 (i,j)，使 β_{ij}^2 达到最大值。记 $\beta_{\max}^2 := \max_{(i,j)} \beta_{ij}^2$，则可得 $\boldsymbol{\xi}_{\mathbf{X}_t\mathbf{X}_{t+1}} = -\frac{1}{L}\beta_{\max}\mathbf{G}_{\max}^R(\mathbf{X}_t)$，并且–

$$\begin{aligned} f(\mathbf{X}_{t+1}) &\leq f(\mathbf{X}_t) - \frac{\beta_{\max}^2}{L} + \frac{\beta_{\max}^2}{L^2}\frac{L}{2} \;=\; f(\mathbf{X}_t) - \frac{\beta_{\max}^2}{2L} \\ &\leq f(\mathbf{X}_t) - \frac{1}{2dL}\sum_{1\leq j\leq i\leq n}\beta_{ij}^2 \qquad (80) \\ &= f(\mathbf{X}_t) - \frac{1}{2dL}\|\mathrm{grad}^R f(\mathbf{X}_t)\|_{\mathbf{X}_t}^2 \qquad (81) \end{aligned}$$

这与（71）类似。因此，证明的其余部分也以相同方式进行，从而得到相同的迭代复杂度。 ■

回顾表 2 可知，对于 $f \in \mathcal{F}$，RRSD 的每次迭代复杂度为 $\mathcal{O}(n)$，而 RGSD 的每次迭代复杂度为 $\mathcal{O}(n^2)$。根据定理 5 和推论 6，RRSD 和 RGSD 的单向变体的整体计算复杂度分别变为 $\mathcal{O}(n^3)$ 和 $\mathcal{O}(n^4)$。因此，单向 RRSD 的整体复杂度与 RGD 相同，而单向 RGSD 的复杂度高于 RGD。然而，与 RGD 所需的 $\mathcal{O}(n^2)$ 内存相比，这两种单向变体仅需 $\mathcal{O}(n)$ 内存。

5.2 多维更新

在多向 RRSD 中，如果所选下降方向至少近似是最速下降方向的无偏估计，则证明过程类似。回顾 $\mathcal{E}_t$ 的定义，我们要求的是

$$\mathbb{E}_t[\sum_{(i,j)\in\mathcal{E}_t}\beta_{ij}^2] \approx \frac{1}{n}\sum_{1\leq j\leq i\leq n}\beta_{ij}^2 \qquad (82)$$

其中，$\mathbb{E}_t[\cdot]$ 表示相对于随机集合 $\mathcal{E}_t$ 的期望。实践中，为确保所选方向近似无偏，我们随机排列集合$\{1,2,\ldots,n\}$，然后将其重新整形为一个 $2\times\lfloor n/2\rfloor$ 矩阵 $\mathbf{E}$，如果 n 为奇数，则忽略最后一个元素。随后，针对 $\mathbf{E}$ 的每一列选取相应的基向量。对于 $\mathbf{E}$ 的第 i 列（记为 $[k,l]^\mathsf{T}$），我们以概率 $1/n$ 选择方向 $\mathbf{G}_{kk}^R(\mathbf{X})$ 和$\mathbf{G}_{ll}^R(\mathbf{X})$，并以概率 $1-1/n$ 选择方向 $\mathbf{G}_{kl}^R(\mathbf{X})$。对于该算法，可以看出

$$\mathbb{E}_t[\sum_{(i,j)\in\mathcal{E}_t}\beta_{ij}^2] = \frac{1}{c}\sum_{1\leq j\leq i\leq n}\beta_{ij}^2 \qquad (83)$$

其中，当 n 为偶数时，$c=n$ 为 n；当 n 为奇数时，$c=n^2/(n-1)$。下述定理给出了 (83) 成立时的迭代复杂度结果。

定理 7 在假设 ***A1*** 和 ***A2***成立，且满足 (83) 时，多向 *RRSD* 算法的迭代复杂度为 $\mathcal{O}(n\kappa\log(\frac{D_0}{\epsilon}))$。 —

证明 将所选的下降方向记为

$$\mathbf{S}(\mathbf{X}_t) = \sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{G}_{ij}^R(\mathbf{X}_t) = \mathbf{B}_t\Big(\sum_{(i,j)\in\mathcal{E}_t} \beta_{ij}\mathbf{E}_{ij}\Big)\mathbf{B}_t^\mathsf{T}. \tag{84}$$

将 $\boldsymbol{\xi}_{\mathbf{X}_t\mathbf{X}_{t+1}} = -\frac{1}{L}\mathbf{S}(\mathbf{X}_t)$ 代入式 (14)，得到

$$\begin{aligned} f\left(\mathbf{X}_{t+1}\right) &\leq f\left(\mathbf{X}_t\right) - \sum_{(i,j)\in\mathcal{E}_t} \frac{\beta_{ij}^2}{L} + \sum_{(i,j)\in\mathcal{E}_t} \frac{\beta_{ij}^2}{L^2}\frac{L}{2} \\ &= f\left(\mathbf{X}_t\right) - \sum_{(i,j)\in\mathcal{E}_t} \frac{\beta_{ij}^2}{2L}. \end{aligned} \tag{85}$$

对 $\mathcal{E}_t$ 中的随机索引取期望，并利用式 (83)，得到

$$\mathbb{E}_t[f\left(\mathbf{X}_{t+1}\right)] \leq f\left(\mathbf{X}_t\right) - \frac{1}{2cL}\|\mathrm{grad}^R f(\mathbf{X}_t)\|_{\mathbf{X}_t}^2 \tag{86}$$

这与式 (71) 中的界类似。按照前面的步骤，可得到所需的界

$$\mathbf{E}f\left(\mathbf{X}_{t+1}\right) - f(\mathbf{X}^\star) \leq \left(1 - \frac{\mu}{4cL}\right)^{t+1}\left[f(\mathbf{X}_0) - f(\mathbf{X}^\star)\right] \tag{87}$$

这对应的迭代复杂度为 $\mathcal{O}(n\kappa\log(D_0/\epsilon))$。 ■

与单向情形相比，多方向 RRSD 的迭代复杂度提升了 $\mathcal{O}(n)$ 倍，因为我们选择沿着 $\mathcal{O}(n)$ 个方向推进，而可能的方向共有 $\mathcal{O}(n^2)$ 个。对于多方向 RGSD 算法，也有类似的结果，其证明需要以下引理作为准备。

引理 8 设 $\mathcal{E}$ 为索引对的集合 $\{(i,j)\}$，它们表示由多方向 *RGSD* 算法选取的基方向 $\{\mathbf{G}_{ij}^R(\mathbf{X})\}$，则以下不等式成立

$$\sum_{(i,j)\in\mathcal{E}} \beta_{ij}^2 \geq \frac{1}{2n}\|grad^R f(\mathbf{X})\|_{\mathbf{X}}^2 \tag{88}$$

对于所有$\mathbf{X} \in \mathbb{P}^n$。

证明 稍微滥用一下记号，令 $\{\beta_{ij}\}_{1\leq j\leq i\leq n}$ 表示由线性索引的项 $\{\beta_m\}_{m\in\mathcal{I}}$，其中 $\mathcal{I} = \{1, 2, \cdots, n(n+1)/2\}$。类似地，令 $\mathcal{M}$ 收集贪心算法所选基向量对应的线性索引，使得$\sum_{(i,\ j)\in\mathcal{E}} \beta_{ij}^2 = \sum_{m\in\mathcal{M}} \beta_m^2$。不失一般性，假设 $\beta_{m_1}^2 \geq \beta_{m_2}^2 \geq \cdots \geq \beta_{m_K}^2$对于$m_i \in \mathcal{M}$成立，其中 $K = |\mathcal{M}|$。

在第一步中，RGSD 选择 $m_1 = \arg\max_{m\in\mathcal{I}} \beta_m^2$。令 $\mathcal{M}(m_1)$ 表示包含 m_1 在内的重叠基向量集合，则 $m_1 = \arg\max_{m\in\mathcal{M}(m_1)} \beta_m^2$。对于满足 $i \neq j$ 的基向量$\mathbf{G}_{ij}^R(\mathbf{X})$，有 $|\mathcal{M}(m_1)| = 2n-1$；而对于 $\mathbf{G}_{ii}^R(\mathbf{X})$，有$|\mathcal{M}(m_1)| = n$。因此，根据 β_{m_1} 的定义，我们得到不等式：

$$\beta_{m_1}^2 \geq \frac{\sum_{m\in\mathcal{M}(m_1)} \beta_m^2}{|\mathcal{M}_{m_1}|} \geq \frac{\sum_{m\in\mathcal{M}(m_1)} \beta_m^2}{2n} \tag{89}$$

同理，在第 i 步中，RGSD 选择 $m_i = \arg\max_{m\in\mathcal{M}(m_i)} \beta_m^2$，其中 $\mathcal{M}(m_i)$ 是从集合 $\mathcal{I}\setminus\cup_{j=1}^{i-1}\mathcal{M}(m_j)$ 中取出的、与 m_i 重叠的索引集合（包括 m_i）。此外，可以看出 $|\mathcal{M}(m_i)| \le 2n$，因此

$$\beta_{m_i}^2 \ge \frac{\sum_{m\in\mathcal{M}(m_i)}\beta_m^2}{|\mathcal{M}_{m_i}|} \ge \frac{\sum_{m\in\mathcal{M}(m_i)}\beta_m^2}{2n} \tag{90}$$

因此，将 (90) 对所有 $i = 1、\ldots、K$ 求和，可得

$$\sum_{m\in\mathcal{M}}\beta_m^2 \ge \frac{1}{2n}\sum_{m\in\cup_j\mathcal{M}(m_j)}\beta_m^2 = \frac{1}{2n}\sum_{m\in\mathcal{I}}\beta_m^2 \tag{91}$$

$$= \frac{1}{2n}\|\mathrm{grad}^R f(\mathbf{X})\|_{\mathbf{X}}^2 \tag{92}$$

这就是所需的不等式。(91) 中的最后一个等式由 RGSD 的停止准则得出。

■

推论9 在假设 ***A1*** 和 ***A2***下，多方向 *RGSD* 算法的迭代复杂度为 $\mathcal{O}(n\kappa\log(\frac{D_0}{\epsilon}))$。

—

证明 由 (85) 可得

$$f(\mathbf{X}_{t+1}) \le f(\mathbf{X}_t) - \sum_{(i,j)\in\mathcal{E}_t}\frac{\beta_{ij}^2}{2L} \le f(\mathbf{X}_t) - \frac{1}{4nL}\|\mathrm{grad}^R f(\mathbf{X}_t)\|_{\mathbf{X}_t}^2 \tag{93}$$

其中，(93) 中最后一个不等式由引理 8 得出。结合 (75) 和 (93)，可得

$$\Rightarrow f(\mathbf{X}_{t+1}) - f(\mathbf{X}^\star) \le \left(1 - \frac{\mu}{8nL}\right)^{t+1}\left[f(\mathbf{X}_0) - f(\mathbf{X}^\star)\right]$$

这对应的迭代复杂度为 $\mathcal{O}\left(n\kappa\log\left(\frac{D_0}{\epsilon}\right)\right)$。

—

■

总之，与其单向变体相比，多向 RRSD 和 RGSD 算法的迭代复杂度提升至 $\mathcal{O}\big(n\kappa\log\big(\frac{D_0}{\epsilon}\big)\big)$。然而，从表 2 可以看出，对于 $f\in\mathcal{F}$，多向 RRSD 和 RGSD 算法的总体计算复杂度仍分别为 $\mathcal{O}(n^3\kappa\log(D_0/\epsilon))$ 和 $\mathcal{O}(n^3\log(n)\kappa\log(D_0/\epsilon))$。换言之，与欧几里得情形下的坐标下降算法一样，每种子空间下降变体的总体计算复杂度大致与 RGD 相同。不过，子空间下降算法的每次迭代计算复杂度显著更低，因此可应用于大规模场景，在这类场景中，即使只执行一次 RGD 迭代也可能难以承受。

6 函数类 F 的自适应步长选择 $\mathcal{F}$

在本节中，我们提出一种自适应步长选择规则，可简化在大规模问题上运行所提算法时的超参数调优。尽管自适应步长已用于黎曼优化领域，但本工作着重于在实现自适应的同时不牺牲子空间方法的计算效率。对低复杂度自适应方法的需求，排除了使用非精确线搜索（Ferreira 等人（2019））和自适应梯度方法（Roy 等人（2018）、B´ecigneul 和 Ganea（2018）、Kasai 等人（2019））等技术的可能性。在本工作中，我们采用 Fountoulakis 和 Tappenden（2018）针对欧几里得情形提出的步长选择策略的一个修改版本。具体而言，我们沿着根据第 4 节选定的子空间内的每个基方向，分别构造目标函数的二次近似。随后，选择沿这些方向的步长，使相应的二次近似最小化。沿测地线 $\gamma(\lambda)$ 且 $\gamma'(0) = \mathbf{V}$ 的 f 的泰勒级数展开可写为（Boumal，2023，第 80 页）：

$$\begin{aligned} f(\gamma(\lambda)) &\cong f(\mathbf{X}) + \lambda \left\langle \operatorname{grad}^R f(\mathbf{X}), \mathbf{V} \right\rangle_{\mathbf{X}} + \frac{\lambda^2}{2} \left\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V} \right\rangle_{\mathbf{X}} \\ &= f(\mathbf{X}) + \lambda \operatorname{tr}\left(\operatorname{grad} f(\mathbf{X})\mathbf{V}\right) + \frac{\lambda^2}{2} \left\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V} \right\rangle_{\mathbf{X}}. \end{aligned} \tag{94}$$

回顾一下，grad $f(\mathbf{X})$ 和 $\operatorname{grad}^R f(\mathbf{X})$ 分别表示欧几里得梯度和黎曼梯度，而 $H^R f(\mathbf{X})$ 表示函数 f 的黎曼Hessian。在时刻 t，使测地凸函数 f 沿初始方向 $\mathbf{V}$ 的测地线上的二阶近似 (94) 最小化的 $\lambda_{t,\mathbf{V}}$ 最优值为

$$\lambda^\star_{t,\mathbf{V}} = -\frac{\operatorname{tr}\left(\operatorname{grad} f(\mathbf{X})\mathbf{V}\right)}{\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V}\rangle_{\mathbf{X}}}. \tag{95}$$

对于每个 $\mathbf{G}^R_{ij}$，相应的 $\lambda^\star_{t,\mathbf{G}^R_{ij}}$ 在 (42) 中起到 $-\alpha_t\beta_{ij}$ 的作用。值得注意的是，计算 β_{ij} 本身就是计算 $\lambda^\star_{t,\mathbf{G}^R_{ij}}$ 的一部分。因此，在第 t 次迭代中，下降方向为

$$\mathbf{S}(\mathbf{X}_t) = \sum_{(i,j)\in\mathcal{E}_t} \lambda^\star_{t,\mathbf{G}^R_{ij}} \mathbf{G}^R_{ij}(\mathbf{X}_t) = \mathbf{B}_t \left(\sum_{(i,j)\in\mathcal{E}_t} \lambda^\star_{t,\mathbf{G}^R_{ij}} \mathbf{E}_{ij} \right) \mathbf{B}_t^\mathsf{T} \tag{96}$$

由于每个子空间中的基向量彼此不重叠，由 (43) 可知，沿各个基方向的相应最优步长可以独立选择。因此，更新方程变为

$$\mathbf{X}_{t+1} = \operatorname{Exp}_{\mathbf{X}_t}(\mathbf{S}(\mathbf{X}_t)) = \mathbf{B}_t \exp\left(\sum_{(i,j)\in\mathcal{E}_t} \lambda^\star_{t,\mathbf{G}^R_{ij}} \mathbf{E}_{ij} \right) \mathbf{B}_t^\mathsf{T} \tag{97}$$

值得注意的是，对于函数类而言，计算最优 $\lambda^\star_{t,\mathbf{G}^R_{ij}}$ 的计算复杂度为 $O(n)$，其中每个 $\mathbf{G}^R_{ij}$ 均如此。详细计算见附录 B。若我们

预计 f 的增长速度会超过其二次近似，则可能有必要通过将其除以一个缩放因子来缩小 $\lambda_t^\star$ ，$\mathbf{G}_{ij}^R$ ，从而实现函数值的降低。对于给定问题，必须对该缩放因子进行调节。

值得注意的是，对于黎曼梯度下降，基于二次近似的步长选择在计算上要求较高；另请参见附录 B。因此，所提出的步长选择方法对于黎曼梯度下降并不实用。

7 实验结果

本节我们在 $\mathcal{F}$ 中测试所提子空间下降变体在强 g-凸函数上的性能。我们将所提算法的性能与 RGD 算法进行比较。我们尤其关注大规模场景，在这类场景下，加速方法和二阶方法的复杂度高得难以承受，因而不切实际。我们还通过数值结果确认，RGD 的每次迭代复杂度随 $\mathcal{O}(n^3)$ 增长，而多向 RRSD 和 RGSD 的每次迭代复杂度随 $\mathcal{O}(n^2)$ 增长，单向 RRSD 的每次迭代复杂度则随 $\mathcal{O}(n)$ 增长。所有模拟均在配备 512 GB 内存的系统上使用 MATLAB 完成。为确保大规模场景下有足够的内存，我们不预先分配中间变量，并在每次使用后将其清除。我们仔细确保不同算法的关键步骤均以相似方式实现，从而使其运行时间能够反映各自的计算复杂度，并可直接进行比较。

未纳入与 Gutman 和 Ho-Nguyen（2022年）提出的坐标下降方法的比较，因为该方法未提供用于处理 SPD流形上的函数最小化（参见（1））的具体低复杂度算法。不过，我们注意到，Gutman 和 Ho-Nguyen（2022年）中给出的一般方法可以被视为包含单方向 RRSD 算法这一特殊情况，但不包含其他变体。

考虑由 $f \in \mathcal{F}$ 给出的函数。

$$\begin{aligned} f(\mathbf{X}) &= \operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{D}\mathbf{X}\right) + k \log\det(\mathbf{X}) \\ &= \operatorname{tr}\left(\mathbf{B}^{-1}\mathbf{C}\mathbf{B}^{-\mathsf{T}}\right) + \operatorname{tr}\left(\mathbf{B}^{\mathsf{T}}\mathbf{D}\mathbf{B}\right) + 2k \log\det(\mathbf{B}) \end{aligned} \tag{98}$$

其中，$\mathbf{B} = \mathcal{L}(\mathbf{X})$、$\mathbf{C}$，$\mathbf{D} \succ 0$和 $k \in \mathbb{R}$。此外，f 是具有 $\mu = \min\{\lambda_{\min}(\mathbf{C}), \lambda_{\min}(\mathbf{D})\}$ 的 μ-强 g-凸函数（参见引理10），并且具有梯度

$$\operatorname{grad} f(\mathbf{X}) = \mathbf{D} - \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} + k\mathbf{X}^{-1} \tag{99}$$

$$\operatorname{grad}^R f(\mathbf{X}) = \mathbf{X}\mathbf{D}\mathbf{X} - \mathbf{C} + k\mathbf{X} \tag{100}$$

一般而言，f 的梯度场可能并非利普希茨光滑，除非我们将$\mathbf{X}$ 限制在半径为 R的范数球内。例如，函数$\operatorname{tr}\left(\mathbf{X}^{-1} + \mathbf{X}\right)$ 的梯度向量场并非利普希茨光滑，但当$\mathbf{X}$ 被限制在半径为 R 的范数球内时，其梯度向量场是$(e^R + e^{-R})$-利普希茨光滑的（参见引理11）。对于下降算法，假设满足 μ-强测地凸性，可以看出，所有迭代点都将位于半径为$\sqrt{2(f(\mathbf{X}_0) - f(\mathbf{X}^\star))/\mu}$的范数球内。

(98) 中的函数形式源于其在对具有 Wishart 先验的 n 维高斯随机变量的协方差矩阵进行最大后验（MAP）估计中的应用。

具体而言，针对独立同分布数据点 $\mathbf{x}_1, \ldots, \mathbf{x}_N \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ 的似然函数为

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \{\mathbf{x}_i\}_{i=1}^N) = c_1 \det(\boldsymbol{\Sigma})^{-n/2} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{C}_d \boldsymbol{\Sigma}^{-1}\right)\right) \tag{101}$$

其中，

$$\mathbf{C}_d = \sum_{i=1}^{N} (\mathbf{x}_i - \boldsymbol{\mu})(\mathbf{x}_i - \boldsymbol{\mu})^{\mathsf{T}} \tag{102}$$

且 $\boldsymbol{\Sigma}$ 服从参数为 $(n,\ p,\ \mathbf{S})$ 的 Wishart 分布（Gupta and Nagar，1999，第 87 页）

$$p\left(\boldsymbol{\Sigma} \middle| n, p, \mathbf{S}\right) = c_2 \det(\boldsymbol{\Sigma})^{\frac{1}{2}(n-p-1)} \exp\left(-\frac{1}{2}\mathrm{tr}\left(\mathbf{S}^{-1}\boldsymbol{\Sigma}\right)\right) \tag{103}$$

for $\mathbf{S} \succ \mathbf{0}$ 和 $n \geq p$。因此，$\boldsymbol{\Sigma}$ 的最大后验估计量可通过求解得到。 g

$$\hat{\boldsymbol{\Sigma}}_{\mathrm{MAP}} = \min_{\boldsymbol{\Sigma}} \left\{\mathrm{tr}\left(\mathbf{C}_d \boldsymbol{\Sigma}^{-1}\right) + \mathrm{tr}\left(\mathbf{S}^{-1}\boldsymbol{\Sigma}\right) + c_3 \log\det(\boldsymbol{\Sigma})\right\} \tag{104}$$

我们指出，在对协方差矩阵进行最大后验估计时，通常会改用逆Wishart先验。由于逆Wishart是高斯分布协方差矩阵的共轭先验，因此可以得到闭式最大后验估计。然而，式 (104) 中的表述允许采用Wishart先验，也可以类似地加以修改，以允许使用其他先验，包括逆Wishart先验和正态-Wishart先验。

我们还注意到，对于$\mathbf{C}$、$\mathbf{D}$和 k 的某些特定取值，解可以通过闭式形式求得：

- 对于 $\mathbf{D} = \mathbf{I}$（单位矩阵）和 $k = 0$，我们有 $\mathrm{grad}^R f(\mathbf{X}^\star) = \mathbf{X}^\star\mathbf{X}^\star - \mathbf{C} = 0$，由此得到 $\mathbf{X}^\star = \mathbf{C}^{1/2}$，并且
- 对于$\mathbf{C} = \mathbf{0}$ 和 $k = -1$，有 $\mathrm{grad}^R f(\mathbf{X}^\star) = \mathbf{X}^\star\mathbf{D}\mathbf{X}^\star - \mathbf{X}^\star = 0$，由此得到$\mathbf{X}^\star = \mathbf{D}^{-1}$。

对于 f 在 (98) 中，RGD更新形式如下：

$$\begin{aligned}\mathbf{X}_{t+1} &= \mathbf{B}_t \exp\left(-\alpha_t \mathbf{B}_t^{-1} \mathrm{grad}^R f(\mathbf{X}_t) \mathbf{B}_t^{-\mathsf{T}}\right) \mathbf{B}_t^{\mathsf{T}} \\ &= \mathbf{B}_t \exp\left(\alpha_t \left(\mathbf{B}_t^{-1}\mathbf{C}\mathbf{B}_t^{-\mathsf{T}} - \mathbf{B}_t^{\mathsf{T}}\mathbf{D}\mathbf{B}_t - k \cdot \mathbf{I}\right)\right) \mathbf{B}_t^{\mathsf{T}}.\end{aligned} \tag{105}$$

对于 RRSD 和 RGSD算法，系数 β_{ij} 为

$$\beta_{ij} = \mathrm{tr}\left(\left[\mathbf{B}_t^{\mathsf{T}}\mathbf{D}\mathbf{B}_t - \mathbf{B}_t^{-1}\mathbf{C}\mathbf{B}_t^{-\mathsf{T}} + k \cdot \mathbf{I}\right] \mathbf{E}_{ij}\right). \tag{106}$$

如前文所述，$f(\mathbf{X}_t)$ 和 β_{ij} 可以使用以下中间变量表示：

$$\mathbf{M}_1(\mathbf{X}_t) = \mathbf{B}_t^{\mathsf{T}}\mathbf{D}\mathbf{B}_t \tag{107}$$

$$\mathbf{M}_2(\mathbf{X}_t) = \mathbf{B}_t^{-1}\mathbf{C}\mathbf{B}_t^{-\mathsf{T}} \tag{108}$$

使得

$$\begin{aligned} f(\mathbf{X}_t) &= \operatorname{tr}(\mathbf{M}_1(\mathbf{X}_t)) + \operatorname{tr}(\mathbf{M}_2(\mathbf{X}_t)) + k \cdot \log\det(\mathbf{X}_t) \\ \beta_{ij} &= \sqrt{2}^{\mathbf{I}(i\neq j)} \left[\mathbf{M}_1(\mathbf{X}_t) - \mathbf{M}_2(\mathbf{X}_t) + k \cdot \mathbf{I}\right]_{ij} \end{aligned} \tag{109}$$

中间矩阵 $\mathbf{M}_1(\mathbf{X}_t)$ 和 $\mathbf{M}_2(\mathbf{X}_t)$ 可以按照（56）–（57）中规定的递归方式进行更新。

我们考虑 k 的两种取值，即 -1 和 0，分别对应以下函数：

$$f_1(\mathbf{X}) = \operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{D}\mathbf{X}\right) - \log\det(\mathbf{X}) \tag{110}$$

$$f_2(\mathbf{X}) = \operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{D}\mathbf{X}\right) \tag{111}$$

分别地。我们注意到，$\log\det(\mathbf{X})$ 在 SPD流形上是测地线线性的，因此 f_1 和 f_2 具有相同的条件数。然而，基于梯度的方法的性能通常取决于 $\mathbf{X}^\star$ 邻域内的局部条件数，而 f_1 和 f_2 的局部条件数可能不同（见附录 A.1 ）。例如，若我们取

$$\mathbf{C} = \begin{bmatrix} 5.6667 & 10.0000 & 5.8889 \\ 10.0000 & 26.2222 & 17.5556 \\ 5.8889 & 17.5556 & 12.1111 \end{bmatrix} \tag{112}$$

以及 $\mathbf{D} = \mathbf{I}$，则可以通过数值验证得到，f_1 和 f_2 在各自最优点处的条件数分别为 12.87 和 104.88。事实上，正如我们在模拟中也将看到的，f_1 表现相对良好，各种算法在 f_1 上的收敛速度都快于在 f_2上的速度。

7.1 每次迭代复杂度

我们首先通过实证验证所提出的子空间下降算法的每次迭代时间复杂度。由于我们只关注每次迭代的复杂度及其随 n 的变化，因此为节省仿真时间，我们令$\mathbf{C}$ 和$\mathbf{D}$ 为不一定正定的对称矩阵。图 1(a) 展示了 RGD 和 RRSD 算法每次迭代所需的时间。如图所示，正如预期，多向 RRSD 的每次迭代时间复杂度随 $\mathcal{O}(n^2)$ 增长，而 RGD 的则随 $\mathcal{O}(n^3)$ 增长。

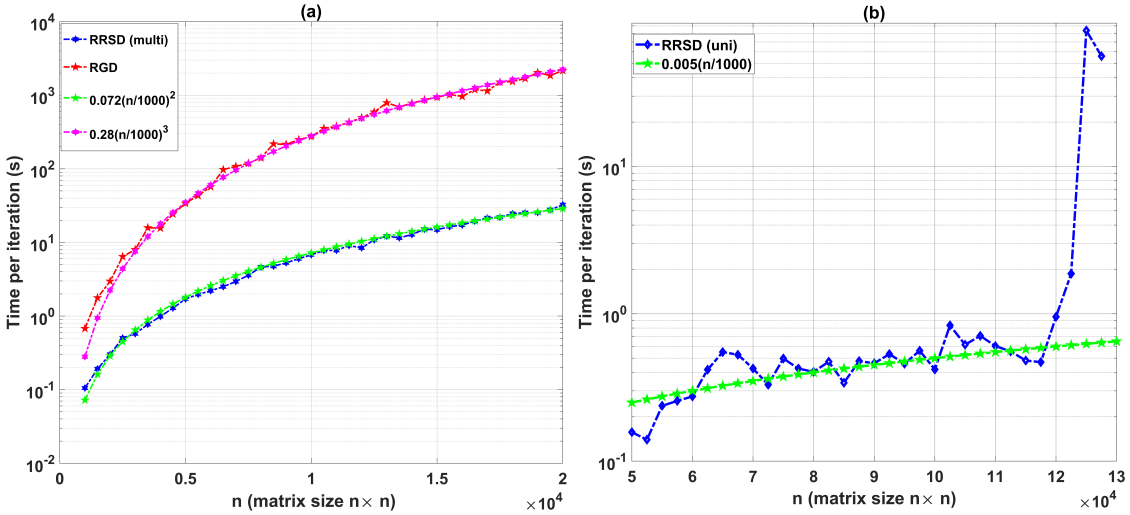


图1：RGD、单向 RRSD（RRSD（单向））和多向 RRSD（RRSD（多向））方法的每次迭代耗时比较

我们还考虑了更大规模的情形，在这种情形下，即使 $\mathcal{O}(n^2)$ 复杂度也难以实际应用，因此有必要实现单向 RRSD，其每次迭代复杂度仅随 n 线性增长。图 1(b) 展示了 RRSD 算法每次迭代所需的时间。图中未显示 RGD 的性能，因为我们发现它在处理 $n = 60000$ 时内存耗尽。可以看出，RRSD 的每次迭代时间复杂度在 $n \leq 122500$ 范围内近乎线性增长。然而，对于更大的 n，由于使用了交换内存，时间复杂度会骤然增加，而系统在处理 $n = 130000$ 时会耗尽内存。我们指出，如 4.1 节所述，通过直接从磁盘读取/写入不同变量的部分内容，或许可以突破这一限制。

7.2 性能比较

接下来，我们比较各种所提算法与 RGD 的实测性能。为评估所提算法的性能，我们针对 $n \in \{100, 500, 1000\}$ 模拟了对称正定矩阵 $\mathbf{C} \in \mathbb{R}^{n\times n}$，令其为 $\mathbf{C} = \mathbf{C}^{temp}\mathbf{C}_{temp}^{\mathsf{T}}/n^2$，其中，$\mathbf{C}^{temp}$ 是从区间 [1,10] 上的离散均匀分布中抽取的 $n \times n$ 伪随机整数矩阵。并且 $\mathbf{D} = \mathbf{I}$。所有算法均以 $\mathbf{X}_0 = \mathbf{I}$ 初始化。RRSD 和 RGSD 的步长根据第 6 节所述的自适应步长选择方法计算，而 RGD 的步长设为 0.1，这是 RGD 能够收敛的最大值。对于当前问题，可以用闭式形式计算最优值。具体而言，若$\mathbf{C}$ 的特征值分解为 $\mathbf{C} = \mathbf{U}\Sigma_{\mathbf{C}}\mathbf{U}^{\mathsf{T}}$，则 f_1 的最小化元为$\mathbf{X}_1^\star = \mathbf{U}\Sigma_{\mathbf{X}_1^\star}\mathbf{U}^{\mathsf{T}}$，其中 $[\Sigma_{\mathbf{X}_1^\star}]_{ii} = \frac{1+\sqrt{1+4[\Sigma_{\mathbf{C}}]_{ii}}}{2}$，而 f_2的最小化元为 $\mathbf{C}^{1/2}$。

回顾一下，各种算法的理论性能是通过计算复杂度来表征的。在实际应用中，有多种方法可以对性能进行基准测试。这里，我们使用以下指标进行性能比较：

1) 基方向数量，$\sum_t K_t$,

2) 所有迭代过程中必须计算的**F(X)**的条目数,

3）浮点运算总次数。

不同 n 值和不同度量下的最优性间隙分别展示于图 2 和图 3 中，对应 f_1 和 f_2。

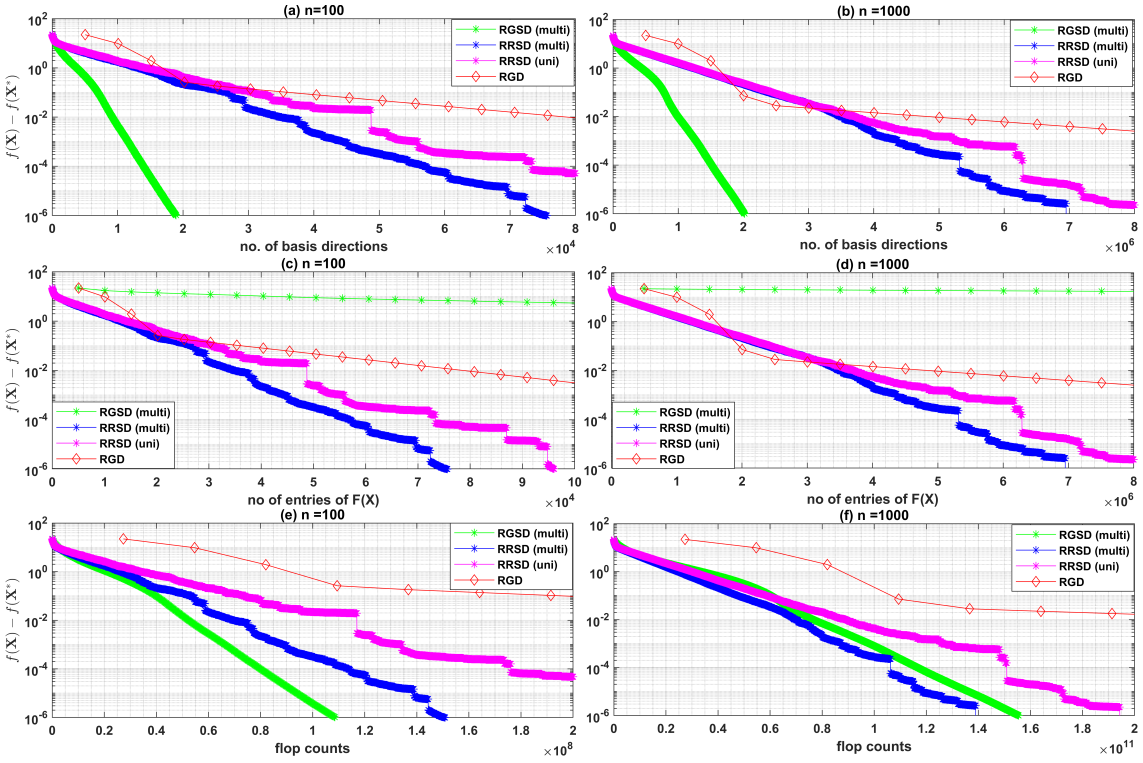


图 2：函数 f_1**(X)** 的 RGD 与子空间下降方法对比；RGSD（多向）：多方向 RGSD，RRSD（多向）：多向 RRSD，RRSD（单向）：单向 RRSD。

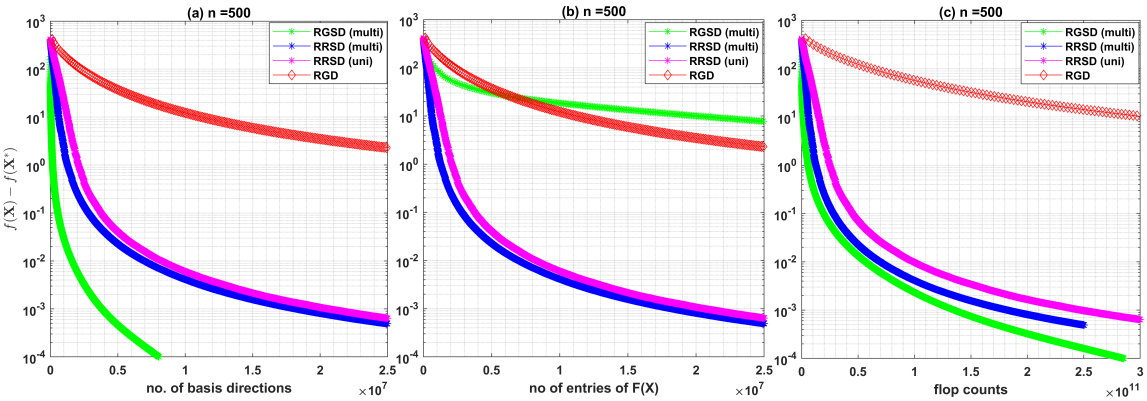


图 3：系 亿博 函数 f_2**(X)** 的梯度下降与子空间下降方法的比较

回顾一下，RRSD 和 RGSD 算法每次迭代最多使用 n 个基方向，而 RGD 在每次迭代中使用全部 $n_{(n+1)\ 2}$个方向。从图 2(a)、2(b) 和图 3(a) 可以看出，RRSD 和 RGSD 算法的表现均优于 RGD。有趣的是，对于这一度量，多方向 RGSD 算法的性能显著优于所有其他算法，因为它力求在每次迭代中选择尽可能好的方向。

回想一下，两种 RRSD 算法变体每次迭代最多需要 n 个 **F(X)** 条目。相比之下，尽管 RGSD 算法每次迭代也最多使用 n 个 **F(X)** 条目，但为了选择贪心方向，它仍然需要 **F(X)** 的所有条目。图 2(c)、2(d) 和 3(b) 印证了这一观察结果：在所有算法中，RGSD 的表现最差。

最后，我们根据不同算法的浮点运算次数进行比较；这一指标应能反映实际运行时间，且在很大程度上不受系统和实现方式的影响。图 2(e)、2(f) 和 3(c) 显示，多向 RRSD 和 RGSD 均优于单向 RRSD，而单向 RRSD 又明显优于 RGD 算法。在这些算法中，仅当 n 较小时，RGSD 才表现更好，因为对于较大的 n，计算并排序 **F(X)** 条目的开销开始占据主导。

8 结论与未来工作

对称正定（SPD）流形上的最小化问题出现在许多领域，例如核矩阵学习、高斯分布的协方差估计、椭圆等高分布的最大似然参数估计，以及高斯混合模型问题中的参数估计。近年来，人们发展了黎曼优化算法，通过利用 SPD 流形的结构，更高效地求解这些问题。然而，黎曼梯度下降及其他相关算法通常需要在每次迭代中执行矩阵指数运算和稠密矩阵乘法等开销较大的矩阵运算，因此每次迭代的复杂度至少为 $\mathcal{O}(n^3)$。受欧几里得情形中常用的坐标下降算法启发，我们提出了黎曼子空间下降算法，其每次迭代的复杂度可降低至 $\mathcal{O}(n)$ 和 $\mathcal{O}(n^2)$。所提算法通过识别特殊子空间，使迭代点的 Cholesky因子能够以较低复杂度更新，并将优化限制在一类可高效计算黎曼梯度的函数上，从而实现了这一复杂度降低。我们在大规模协方差估计问题上评估了所提算法的性能，结果表明其速度快于黎曼梯度下降算法。

所提算法可视为经典块坐标下降（BCD）算法在黎曼流形上的对应版本。自然地，BCD 领域的进展能否延伸至黎曼情形，仍有待探究。例如，在欧几里得情形下，人们已探索了大量变量选择或块选择策略，黎曼情形下也可以探索类似策略。同样，针对其他有趣情形（例如函数为非凸函数和/或满足 Polyak-Lojasiewicz（PL）不等式时），对所提算法进行迭代复杂度分析也非常重要。最后，仍有待探究的是，能否针对第 4 节简要讨论的有限和问题，开发更快版本的所提算法，例如

采用动量或方差缩减技术。

附录 A.

引理10 函数 $f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{D}\mathbf{X}\right)$ 是 μ-强 g-凸的，其中$\mathbf{C} \succ 0$和$\mathbf{D} \succ 0$满足 $\mu = \min(\lambda_{\min}(\mathbf{C})\ \lambda_{\min}(\mathbf{D}))$.

证明 我们首先注意到

$$df(\mathbf{X}) = -\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}d\mathbf{X}\right) + \mathrm{tr}\left(\mathbf{D}d\mathbf{X}\right) \tag{113}$$

$$\Rightarrow \mathrm{grad}f(\mathbf{X}) = \mathbf{D} - \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}. \tag{114}$$

欧几里得海森矩阵可按如下方式计算。

$$\mathbf{G}(\mathbf{X}) = \mathrm{grad}f(\mathbf{X}) = \mathbf{D} - \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} \tag{115}$$

$$d\mathbf{G}(\mathbf{X}) = \mathbf{X}^{-1}d\mathbf{X}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}d\mathbf{X}\mathbf{X}^{-1} \tag{116}$$

$$\Rightarrow Hf(\mathbf{X})[\mathbf{V}] = \mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1} \tag{117}$$

对于黎曼流形 $\mathbb{P}^n$，欧几里得 Hessian 与黎曼 Hessian 之间的关系如下（Ferreira 等人, 2019年）：

$$H^R f(\mathbf{X})[\mathbf{V}] = \mathbf{X}Hf(\mathbf{X})[\mathbf{V}]\mathbf{X} + \frac{1}{2}\left[\mathbf{V}\mathrm{grad}f(\mathbf{X})\mathbf{X} + \mathbf{X}\mathrm{grad}f(\mathbf{X})\mathbf{V}\right] \tag{118}$$

$$\Rightarrow \left\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V}\right\rangle_{\mathbf{X}} = \mathrm{tr}\left(\mathbf{D}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}\right). \tag{119}$$

假设 $f(\mathbf{X})$ 是 μ-强 g-凸的。那么我们必然有

$$\langle Hf(\mathbf{X})[\mathbf{V}], \mathbf{V}\rangle_{\mathbf{X}} \geq \mu\mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) \tag{120}$$

$$\Leftrightarrow \mathrm{tr}\left(\mathbf{D}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}\right) - \mu\mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) \geq 0 \tag{121}$$

$$\Leftrightarrow \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\left(\mathbf{D} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1}\right)\right) \geq 0 \tag{122}$$

当且仅当

$$\mathbf{D} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1} \succcurlyeq 0 \tag{123}$$

或者等价地

$$\lambda_i(\mathbf{D} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1}) \geq 0 \qquad\qquad 1 \leq i \leq n. \tag{124}$$

为了推导满足式 (124) 的 μ 条件，我们将其分为两种情况：(a) $\lambda_i(\mathbf{X}) \leq 1$ 和 (b) $\lambda_i(\mathbf{X}) > 1$。

情况 $\lambda_i(\mathbf{X}) \leq 1$**:** 对于$\mathbf{D} \succ 0$，我们有（Seber，2008，第117页）

$$\mathbf{D} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1} \succeq \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1} = \mathbf{X}^{-\frac{1}{2}}\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}} - \mu\mathbf{I}\right)\mathbf{X}^{-\frac{1}{2}}. \tag{125}$$

对于两个对称矩阵$\mathbf{A}$ 和$\mathbf{B}$，已知对于非奇异$\mathbf{R}$ ，有$\mathbf{A} \succeq \mathbf{B} \Leftrightarrow \mathbf{R}^{\mathsf{T}}\mathbf{A}\mathbf{R} \succeq \mathbf{R}^{\mathsf{T}}\mathbf{B}\mathbf{R}$ （Seber, 2008，第227页）。因此，（125）等价于

$$\mathbf{X}^{\frac{1}{2}}\left(\mathbf{D}+\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}-\mu\mathbf{X}^{-1}\right)\mathbf{X}^{-\frac{1}{2}} \succeq \mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}}-\mu\mathbf{I} \tag{126}$$

这意味着 (Seber, 2008, p.228)

$$\lambda_i\left(\mathbf{X}^{\frac{1}{2}}\left(\mathbf{D}+\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}-\mu\mathbf{X}^{-1}\right)\mathbf{X}^{-\frac{1}{2}}\right) \geq \lambda_i\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}}-\mu\mathbf{I}\right). \tag{127}$$

若选取 μ使得 $\lambda_i\left(\mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}}-\mu\mathbf{I}\right) \geq 0$，则关系式 (127) 保证对于满足 $\lambda_i(\mathbf{X}) \leq 1$ 的指标 i ，条件 (124) 成立。现在，这将使我们能够推导出关于 μ 的条件。注意,

$$\lambda_i\left[\mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}}-\mu\mathbf{I}\right] = \lambda_i\left[\mathbf{X}^{-\frac{1}{2}}\mathbf{C}\mathbf{X}^{-\frac{1}{2}}\right]-\mu = \lambda_i\left(\mathbf{X}^{-1}\mathbf{C}\right)-\mu \tag{128}$$

其中，最后一个等式是由以下事实得出的：对于两个对称正定矩阵$\mathbf{A}$ 和$\mathbf{B}$，$\mathbf{AB}$ 和$\mathbf{BA}$ 的特征值相同。我们还有 $\lambda_i(\mathbf{AB}) \geq \lambda_i(\mathbf{A})\lambda_{\min}(\mathbf{B})$ (Seber, 2008, p.119)，因此

$$\lambda_i\left(\mathbf{X}^{-1}\mathbf{C}\right)-\mu \geq \lambda_i\left(\mathbf{X}^{-1}\right)\lambda_{\min}(\mathbf{C})-\mu. \tag{129}$$

因此，由于 $\lambda_i(\mathbf{X}) \leq 1$，我们有

$$\lambda_i\left(\mathbf{X}^{-1}\right)\lambda_{\min}(\mathbf{C})-\mu \geq 0 \text{ if } \lambda_{\min}(\mathbf{C}) \geq \mu \tag{130}$$

.

情形 $\lambda_i(\mathbf{X}) > 1$**：** 与前一种情形类似，对于$\mathbf{C} \succ 0$ ，我们有$\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} \succ 0$，这意味着

$$\mathbf{D}+\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}-\mu\mathbf{X}^{-1} \succeq \mathbf{D}-\mu\mathbf{X}^{-1} \tag{131}$$

如果选择 μ 使得不等式

$$\lambda_i\left(\mathbf{D}-\mu\mathbf{X}^{-1}\right) \geq 0 \Leftrightarrow \lambda_i(\mathbf{D}) \geq \mu\lambda_i(\mathbf{X}^{-1}) \tag{132}$$

成立，则可确保对于索引 i，当 $\lambda_i(\mathbf{X}) > 1$时，(124) 也成立。当 $\lambda_{\min}(\mathbf{D}) \geq \mu$时，不等式 (132) 成立。

Comb综合这两种情况可知，$\operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1}+\mathbf{D}\mathbf{X}\right)$ 是 g-强 μ-凸的， h

$$\mu = \min\left(\lambda_{\min}(\mathbf{C}), \lambda_{\min}(\mathbf{D})\right) \tag{133}$$

■

引理11 函数 $f\left(\mathbf{X}\right) = \operatorname{tr}\left(\mathbf{X}^{-1}+\mathbf{X}\right)$ 的梯度向量场一般不是利普希茨光滑的，但当$\mathbf{X}$ 位于紧集内时，是利普希茨光滑的。

证明 对于 $f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{X}^{-1} + \mathbf{X}\right)$，由（119）可得以下关系式

$$\left\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V} \right\rangle_{\mathbf{X}} = \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-2}\right) \tag{134}$$

For L梯度向量场的光滑性，必须满足以下不等式

$$\langle H\ f(\mathbf{X})[\mathbf{V}], \mathbf{V} \rangle_{\mathbf{X}} = \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-2}\right) \leq L\|\mathbf{V}\|_{\mathbf{X}}^2 \tag{135}$$

$$\Leftrightarrow \mathrm{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\left(\mathbf{I} + \mathbf{X}^{-2} - L\mathbf{X}^{-1}\right)\right) \leq 0 \tag{136}$$

当且仅当

$$\mathbf{I} + \mathbf{X}^{-2} - L\mathbf{X}^{-1} \preceq 0 \tag{137}$$

$$\Leftrightarrow \lambda_i^2(\mathbf{X}) + 1 - L\lambda_i(\mathbf{X}) \leq 0;\ \ 1 \leq i \leq n \tag{138}$$

$$\Leftrightarrow \frac{L - \sqrt{L^2 - 4}}{2} \leq \lambda_i(\mathbf{X}) \leq \frac{L + \sqrt{L^2 - 4}}{2};\ \ 1 \leq i \leq n \tag{139}$$

这在一般情况下未必成立。对于不满足 (139) 的$\mathbf{X}$ ，条件 (138) 不成立。因此，函数 $f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{X}^{-1} + \mathbf{X}\right)$的梯度向量场不是利普希茨光滑的。

接下来，我们考虑$\mathbf{X}$ 属于紧集的情况。不失一般性，我们可以假设该集合包含在以 $\mathbf{X} = \mathbf{I}$ 为中心、半径为 R 的范数球内，该范数球由下式给出：

$$\mathcal{B} = \{\mathbf{X} : d(\mathbf{X}, \mathbf{I}) \leq R\} \tag{140}$$

其中,

$$d(\mathbf{X}, \mathbf{I})^2 = \|\log(\mathbf{I}^{-1/2}\mathbf{X}\mathbf{I}^{-1/2})\|_F^2 \tag{141}$$

$$= \|\log(\mathbf{X})\|_F^2 = \sum_{i=1}^{n} (\log(\lambda_i(\mathbf{X})))^2 \tag{142}$$

这意味着

$$-R \leq \log(\lambda_i(\mathbf{X})) \leq R \Leftrightarrow e^{-R} \leq \lambda_i(\mathbf{X}) \leq e^R. \tag{143}$$

注意到 $L = e^{-R} + e^R$ 满足（139）和（143）。因此，在最优点 $\mathbf{X} = \mathbf{I}$周围半径为R的范数球内，函数 $f(\mathbf{X}) = \mathrm{tr}\left(\mathbf{X}^{-1} + \mathbf{X}\right)$ 具有（$e^{-R} + e^R$）-Lipschitz光滑梯度向量场。

■

A.1 测地线线性函数对最优点处条件数的影响

首先考虑函数

$$f_2(\mathbf{X}) = \mathrm{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{X}\right). \tag{144}$$

对于 μ-强 g-凸函数，我们有

$$\langle Hf_2(\mathbf{X})[\mathbf{V}],\mathbf{V}\rangle_{\mathbf{X}} = \operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}\right) \geq \mu\|\mathbf{V}\|_{\mathbf{X}}^2 \tag{145}$$

$$\Leftrightarrow \operatorname{tr}\left((\mathbf{V}\mathbf{X}^{-1}\mathbf{V}(\mathbf{I}+\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}-\mu\mathbf{X}^{-1}))\right) \geq 0 \tag{146}$$

$$\Leftrightarrow \mathbf{I} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1} \succcurlyeq 0. \tag{147}$$

在最优点 $\mathbf{X}^\star = \mathbf{C}^{1/2}$ 处，强凸性条件 (147) 简化为

$$\mathbf{I} + \mathbf{I} - \mu\mathbf{C}^{-1/2} \succcurlyeq 0 \Rightarrow \mu = 2\sqrt{\lambda_{\min}(\mathbf{C})}. \tag{148}$$

类似地，对于 f_2 的 L-光滑性，必须满足以下条件。

$$\left\langle H^R f_2(\mathbf{X})[\mathbf{V}],\mathbf{V}\right\rangle_{\mathbf{X}} = \operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) + \operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}\right) \leq L\|\mathbf{V}\|_{\mathbf{X}}^2 \tag{149}$$

$$\Leftrightarrow \operatorname{tr}\left((\mathbf{V}\mathbf{X}^{-1}\mathbf{V}(\mathbf{I}+\mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1}-L\mathbf{X}^{-1}))\right) \leq 0 \tag{150}$$

$$\Leftrightarrow \mathbf{I} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - L\mathbf{X}^{-1} \preccurlyeq 0. \tag{151}$$

在最优点 $\mathbf{X}^\star = \mathbf{C}^{1/2}$ 处， L-利普希茨光滑条件 (151) 简化为

$$\mathbf{I} + \mathbf{I} - L\mathbf{C}^{-1/2} \preccurlyeq 0 \Rightarrow L = 2\sqrt{\lambda_{\max}(\mathbf{C})}. \tag{152}$$

因此，函数 f_2 在最优点 $\mathbf{X}^\star = \mathbf{C}^{1/2}$ 处的条件数为

$$\kappa_2 = \frac{2\sqrt{\lambda_{\max}(\mathbf{C})}}{2\sqrt{\lambda_{\min}(\mathbf{C})}} = \sqrt{\frac{\lambda_{\max}(\mathbf{C})}{\lambda_{\min}(\mathbf{C})}}. \tag{153}$$

现在考虑函数

$$f_1(\mathbf{X}) = \operatorname{tr}\left(\mathbf{C}\mathbf{X}^{-1} + \mathbf{X}\right) - \log\det(\mathbf{X}) \tag{154}$$

$$\operatorname{grad}^R f_1(\mathbf{X}) = \mathbf{X}^2 - \mathbf{X} - \mathbf{C}. \tag{155}$$

f_1 的最小化元满足

$$\mathbf{X}_\star^2 - \mathbf{X}_\star - \mathbf{C} = 0 \Leftrightarrow \mathbf{I} - \mathbf{X}_\star^{-1} = \mathbf{X}_\star^{-1}\mathbf{C}\mathbf{X}_\star^{-1}. \tag{156}$$

如果将$\mathbf{C} = \mathbf{U}\Sigma_{\mathbf{C}}\mathbf{U}^\mathsf{T}$分解为，那么

$$\mathbf{X}_\star = \mathbf{U}\Sigma_{\mathbf{X}_\star}\mathbf{U}^\mathsf{T} \tag{157}$$

$$[\Sigma_{\mathbf{X}_\star}]_{ii} = \frac{1+\sqrt{1+4[\Sigma_{\mathbf{C}}]_{ii}}}{2}. \tag{158}$$

要使 f_1 为 μ-强 g-凸，我们要求

$$\mathbf{I} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - \mu\mathbf{X}^{-1} \succcurlyeq 0 \tag{159}$$

我们观察到，在满足式 (156) 的最优点 $\mathbf{X}_\star$ 处，强凸性条件 (159) 简化为

$$\mathbf{I} + \mathbf{I} - \mathbf{X}_\star^{-1} - \mu\mathbf{X}_\star^{-1} \succcurlyeq 0 \Leftrightarrow \mathbf{X}_\star \succcurlyeq \frac{\mu+1}{2}\mathbf{I} \tag{160}$$

$$\mu = \sqrt{1 + 4\lambda_{\min}(\mathbf{C})}. \tag{161}$$

同理， L-光滑性条件为

$$\mathbf{I} + \mathbf{X}^{-1}\mathbf{C}\mathbf{X}^{-1} - L\mathbf{X}^{-1} \preccurlyeq 0 \tag{162}$$

在最优点处，(162) 简化为

$$\mathbf{I} + \mathbf{I} - \mathbf{X}_\star^{-1} - L\mathbf{X}_\star^{-1} \preccurlyeq 0 \Leftrightarrow \frac{L+1}{2}\mathbf{I} \succcurlyeq \mathbf{X}_\star \tag{163}$$

$$L = \sqrt{1 + 4\lambda_{\max}(\mathbf{C})}. \tag{164}$$

因此，最优点处的条件数$\mathbf{X}_\star$ 为

$$\kappa_1 = \frac{\sqrt{1 + 4\lambda_{\max}(\mathbf{C})}}{\sqrt{1 + 4\lambda_{\min}(\mathbf{C})}} = \sqrt{\frac{1 + 4\lambda_{\max}(\mathbf{C})}{1 + 4\lambda_{\min}(\mathbf{C})}}. \tag{165}$$

如果矩阵$\mathbf{C}$ 的最小特征值为 $\lambda_{\min}(\mathbf{C}) \ll 1$，那么将 $-\log\det(\mathbf{X})$加到函数 $f_2(\mathbf{X})$中，可将 $\mathbf{X}^\star$处的条件数从 $\kappa_2 = \sqrt{\frac{\lambda_{\max}}{\lambda_{\min}}(\mathbf{C})}\,(\mathbf{C})$ 改善为 $\kappa_1 \approx \sqrt{1 + 4\lambda_{\max}(\mathbf{C})}$，但整体条件数保持不变。

附录 B. 函数类的步长选择

函数 $D(f)\mathbf{x}(\mathbf{V})$表示欧几里得几何中，函数 $f(\mathbf{X})$在点$\mathbf{X}$ 沿$\mathbf{V}$ 方向的方向导数。

分量函数$g_{1,p}$,$g_{2,q}$, g_3,$g_{4,r}$,$g_{5,s}$ 和$g_{6,m}$ 的方向导数如下:

$$D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V}) = -\mathrm{tr}\left(\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) = -\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right) \tag{166}$$

$$D(g_{2,q})\mathbf{x}(\mathbf{V}) = \text{迹}(\mathbf{D}^q\mathbf{V}) \tag{167}$$

$$D(g_3)\mathbf{x}(\mathbf{V}) = \text{迹}\left(\mathbf{X}^{-1}\mathbf{V}\right) \tag{168}$$

$$D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V}) = \mathrm{tr}\left(\mathbf{A}_r\mathbf{V}\mathbf{H}_r\mathbf{X}\right) + \mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right) = 2\mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right) \tag{169}$$

$$D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V}) = -2\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \tag{170}$$

$$D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V}) = \mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right) - \mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\right) \tag{171}$$

回顾一下，黎曼Hessian $H^R f(\mathbf{X})$ 与欧几里得海森矩阵 $Hf(\mathbf{X})$ 之间满足以下关系

$$H^R f(\mathbf{X})[\mathbf{V}] = \mathbf{X}Hf(\mathbf{X})[\mathbf{V}]\mathbf{X} + \frac{1}{2}\left[\mathbf{V}\mathrm{grad}f(\mathbf{X})\mathbf{X} + \mathbf{X}\mathrm{grad}f(\mathbf{X})\mathbf{V}\right] \tag{172}$$

因此,

$$\left\langle H^R f(\mathbf{X})[\mathbf{V}], \mathbf{V}\right\rangle_{\mathbf{X}} = \mathrm{tr}\left(Hf(\mathbf{X})[\mathbf{V}]\mathbf{V}\right) + \mathrm{tr}\left(\mathbf{V}\mathrm{grad}\ f(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) \tag{173}$$

对于测地凸函数，使二阶近似 (94) 最小化的最优 $\lambda_{\mathbf{V}}$ 需要以下量：$\mathrm{tr}\left(\mathrm{grad}\ f(\mathbf{X})\mathbf{V}\right)$、$\mathrm{tr}\left(Hf(\mathbf{X})[\mathbf{V}]\mathbf{V}\right)$ 和 $\mathrm{tr}\left(\mathbf{V}\mathrm{grad}\ f(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)$。

我们定义以下变量以简化方程:

$$h_{1,p}(\mathbf{X}) = \frac{dg}{dg_{1,p}}(\mathbf{X}); \quad 1 \le p \le P \tag{174}$$

$$h_{2,q}(\mathbf{X}) = \frac{dg}{dg_{2,q}}(\mathbf{X}); \quad 1 \le q \le Q \tag{175}$$

$$h_3(\mathbf{X}) = \frac{dg}{dg_3}(\mathbf{X}) \tag{176}$$

$$h_{4,r}(\mathbf{X}) = \frac{dg}{dg_{4,r}}(\mathbf{X}); \quad 1 \le r \le R \tag{177}$$

$$h_{5,s}(\mathbf{X}) = \frac{dg}{dg_{5,s}}(\mathbf{X}); \quad 1 \le s \le S \tag{178}$$

$$h_{6,m}(\mathbf{X}) = \frac{dg}{dg_{6,m}}(\mathbf{X}); \quad 1 \le m \le M \tag{179}$$

$$\mathbf{F}_{1,\ p}\ (\mathbf{X}) = h_{1,\ p}\ (\mathbf{X})\ \mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1} \tag{180}$$

$$\mathbf{F}_{2,\ q}\ (\mathbf{X}) = h_{2,\ q}\ (\mathbf{X})\ \mathbf{D}_q \tag{181}$$

$$\mathbf{F}_3(\mathbf{X}) = h_3(\mathbf{X})\mathbf{X}^{-1} \tag{182}$$

$$\mathbf{F}_{4,r}(\mathbf{X}) = h_{4,r}(\mathbf{X})\left[\mathbf{A}_r\mathbf{X}\mathbf{H}_r + \mathbf{H}_r\mathbf{X}\mathbf{A}_r\right] \tag{183}$$

$$\mathbf{F}_{5,s}(\mathbf{X}) = h_{5,s}(\mathbf{X})\Big[\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\Big] \tag{184}$$

$$\mathbf{F}_{6,m}(\mathbf{X}) = h_{6,m}(\mathbf{X})\left[\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m + \mathbf{P}_m\mathbf{X}^{-1}\mathbf{Q}_m - \mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1} - \mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\right] \tag{185}$$

函数 $h_{1,p}(\mathbf{X})$、$h_{2,q}(\mathbf{X})$、 $h_3(\mathbf{X})$、$h_{4,r}(\mathbf{X})$、$h_{5,s}(\mathbf{X})$、$h_{6,m}(\mathbf{X})$ 是函数 $g_{1,p}(\mathbf{X})$、$g_{2,q}(\mathbf{X})$、 $g_3(\mathbf{X})$、$g_{4,r}(\mathbf{X})$、$g_{5,s}(\mathbf{X})$、$g_{6,m}(\mathbf{X})$。根据全导数的定义，式（174）—（179）中定义的变量的方向导数为

$$\begin{aligned} D\left(h_{1,p}\right)_{\mathbf{X}}(\mathbf{V}) = &\sum_{p'=1}^{P}\left[\frac{dh_{1,p}}{dg_{1,p'}}(\mathbf{X})\right]\left[D\left(g_{1,p'}\right)_{\mathbf{X}}(\mathbf{V})\right] + \sum_{q=1}^{Q}\left[\frac{dh_{1,p}}{dg_{2,q}}(\mathbf{X})\right]\left[D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\right] \\ &+ \left[\frac{dh_{1,p}}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right] + \sum_{r=1}^{R}\left[\frac{dh_{1,p}}{dg_{4,r}}(\mathbf{X})\right]\left[D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right] \\ &+ \sum_{s=1}^{S}\left[\frac{dh_{1,p}}{dg_{5,s}}(\mathbf{X})\right]\left[D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\right] + \sum_{m=1}^{M}\left[\frac{dh_{1,p}}{dg_{6,m}}(\mathbf{X})\right]\left[D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right] \end{aligned} \tag{186}$$

$$\begin{aligned} D\left(h_{2,q}\right)_{\mathbf{X}}(\mathbf{V}) = &\sum_{p=1}^{P}\left[\frac{dh_{2,q}}{dg_{1,p}}(\mathbf{X})\right]\left[D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right] + \sum_{q'=1}^{Q}\left[\frac{dh_{2,q}}{dg_{2,q'}}(\mathbf{X})\right]\left[D\left(g_{2,q'}\right)_{\mathbf{X}}(\mathbf{V})\right] \\ &+ \left[\frac{dh_{2,q}}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right] + \sum_{r=1}^{R}\left[\frac{dh_{2,q}}{dg_{4,r}}(\mathbf{X})\right]\left[D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right] \end{aligned}$$

$$+\sum_{s=1}^{S}\left[\frac{dh_{2,q}}{dg_{5,s}}(\mathbf{X})\right]\left[D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{m=1}^{M}\left[\frac{dh_{2,q}}{dg_{6,m}}(\mathbf{X})\right]\left[D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right] \tag{187}$$

$$\begin{aligned}D\left(h_3\right)_{\mathbf{X}}(\mathbf{V})=&\sum_{p=1}^{P}\left[\frac{dh_3}{dg_{1,p}}(\mathbf{X})\right]\left[D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{q=1}^{Q}\left[\frac{dh_3}{dg_{2,q}}(\mathbf{X})\right]\left[D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\left[\frac{dh_3}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{r=1}^{R}\left[\frac{dh_3}{dg_{4,r}}(\mathbf{X})\right]\left[D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\sum_{s=1}^{S}\left[\frac{dh_3}{dg_{5,s}}(\mathbf{X})\right]\left[D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{m=1}^{M}\left[\frac{dh_3}{dg_{6,m}}(\mathbf{X})\right]\left[D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right]\end{aligned} \tag{188}$$

$$\begin{aligned}D\left(h_{4,r}\right)_{\mathbf{X}}(\mathbf{V})=&\sum_{p=1}^{P}\left[\frac{dh_{4,r}}{dg_{1,p}}(\mathbf{X})\right]\left[D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{q=1}^{Q}\left[\frac{dh_{4,r}}{dg_{2,q}}(\mathbf{X})\right]\left[D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\left[\frac{dh_{4,r}}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{r'=1}^{R}\left[\frac{dh_{4,r}}{dg_{4,r'}}(\mathbf{X})\right]\left[D\left(g_{4,r'}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\sum_{s=1}^{S}\left[\frac{dh_{4,r}}{dg_{5,s}}(\mathbf{X})\right]\left[D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{m=1}^{M}\left[\frac{dh_{4,r}}{dg_{6,m}}(\mathbf{X})\right]\left[D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right]\end{aligned} \tag{189}$$

$$\begin{aligned}D\left(h_{5,s}\right)_{\mathbf{X}}(\mathbf{V})=&\sum_{p=1}^{P}\left[\frac{dh_{5,s}}{dg_{1,p}}(\mathbf{X})\right]\left[D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{q=1}^{Q}\left[\frac{dh_{5,s}}{dg_{2,q}}(\mathbf{X})\right]\left[D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\left[\frac{dh_{5,s}}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{r=1}^{R}\left[\frac{dh_{5,s}}{dg_{4,r}}(\mathbf{X})\right]\left[D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\sum_{s'=1}^{S}\left[\frac{dh_{5,s}}{dg_{5,s'}}(\mathbf{X})\right]\left[D\left(g_{5,s'}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{m=1}^{M}\left[\frac{dh_{5,s}}{dg_{6,m}}(\mathbf{X})\right]\left[D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right]\end{aligned} \tag{190}$$

$$\begin{aligned}D\left(h_{6,m}\right)_{\mathbf{X}}(\mathbf{V})=&\sum_{p=1}^{P}\left[\frac{dh_{6,m}}{dg_{1,p}}(\mathbf{X})\right]\left[D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{q=1}^{Q}\left[\frac{dh_{6,m}}{dg_{2,q}}(\mathbf{X})\right]\left[D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\left[\frac{dh_{6,m}}{dg_3}(\mathbf{X})\right]\left[D\left(g_3\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{r=1}^{R}\left[\frac{dh_{6,m}}{dg_{4,r}}(\mathbf{X})\right]\left[D\left(g_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right]\\&+\sum_{s=1}^{S}\left[\frac{dh_{6,m}}{dg_{5,s}}(\mathbf{X})\right]\left[D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\right]+\sum_{m'=1}^{M}\left[\frac{dh_{6,m}}{dg_{6,m'}}(\mathbf{X})\right]\left[D\left(g_{6,m'}\right)_{\mathbf{X}}(\mathbf{V})\right]\end{aligned} \tag{191}$$

梯度向量场 $\mathbf{F(X)}$ 可以表示为 $h_{1,\ p}\mathbf{(X)}$、$h_{2,\ q}\mathbf{(X)}$、 $h_3\mathbf{(X)}$、$h_{4,\ r}\mathbf{(X)}$、$h_{5,\ s}\mathbf{(X)}$ 以及 $h_{6,\ m}\mathbf{(X)}$，如下所示：

$$\begin{aligned}\mathbf{F}(\mathbf{X}) = &-\sum_{p=1}^{P} h_{1,p}(\mathbf{X})\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1} + \sum_{q=1}^{Q} h_{2,q}(\mathbf{X})\mathbf{D}_q + \sum_{r=1}^{R} h_{4,r}(\mathbf{X})\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r + \mathbf{H}_r\mathbf{X}\mathbf{A}_r\right)\\ &- \sum_{s=1}^{S} h_{5,s}(\mathbf{X})\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\right) + h_3(\mathbf{X})\mathbf{X}^{-1}\\ &+ \frac{1}{2}\sum_{m=1}^{M} h_{6,m}(\mathbf{X})\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m + \mathbf{P}_m\mathbf{X}^{-1}\mathbf{Q}_m - \mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1} - \mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\right)\end{aligned} \tag{192}$$

利用关系式 (174)–(185)，我们有

$$\mathbf{F}(\mathbf{X}) = -\sum_{p=1}^{P}\mathbf{F}_{1,p}(\mathbf{X}) + \sum_{q=1}^{Q}\mathbf{F}_{2,q}(\mathbf{X}) + \mathbf{F}_3(\mathbf{X}) + \sum_{r=1}^{R}\mathbf{F}_{4,r}(\mathbf{X}) - \sum_{s=1}^{S}\mathbf{F}_{5,s}(\mathbf{X}) + \frac{1}{2}\sum_{m=1}^{M}\mathbf{F}_{6,m}(\mathbf{X}) \tag{193}$$

将欧几里得几何中向量场 $\mathbf{F(X)}$ 关于向量场$\mathbf{V}$ 的协变导数记为 $\nabla^E_{\mathbf{V}}\mathbf{F(X)}$。我们将使用莱布尼茨法则进行微分，以计算沿特定方向 $\mathbf{V}$ 的变化率。因此，欧几里得海森矩阵可以写成

$$\begin{aligned}Hf(\mathbf{X})[\mathbf{V}] = &-\sum_{p=1}^{P}\nabla^E_{\mathbf{V}}\mathbf{F}_{1,p}(\mathbf{X}) + \sum_{q=1}^{Q}\nabla^E_{\mathbf{V}}\mathbf{F}_{2,q}(\mathbf{X}) + \nabla^E_{\mathbf{V}}\mathbf{F}_3(\mathbf{X}) + \sum_{r=1}^{R}\nabla^E_{\mathbf{V}}\mathbf{F}_{4,r}(\mathbf{X})\\ &- \sum_{s=1}^{S}\nabla^E_{\mathbf{V}}\mathbf{F}_{5,s}(\mathbf{X}) + \frac{1}{2}\sum_{m=1}^{M}\nabla^E_{\mathbf{V}}\mathbf{F}_{6,m}(\mathbf{X})\end{aligned} \tag{194}$$

其中，

$$\nabla^E_{\mathbf{V}}\mathbf{F}_{1,p}(\mathbf{X}) = \left[D\left(h_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1} - h_{1,p}(\mathbf{X})\begin{bmatrix}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\\ +\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\end{bmatrix} \tag{195}$$

$$\nabla^E_{\mathbf{V}}\mathbf{F}_{2,\ q}\mathbf{(X)} = \left[D(h_{2,\ q})\mathbf{x}\,\mathbf{(V)}\right]\mathbf{D}_q \tag{196}$$

$$\nabla^E_{\mathbf{V}}\mathbf{F}_3(\mathbf{X}) = \Big[D\left(h_3\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathbf{X}^{-1} - h_3(\mathbf{X})\Big(\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\Big) \tag{197}$$

$$\nabla^E_{\mathbf{V}}\mathbf{F}_{4,r}(\mathbf{X}) = \Big[D\left(h_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\Big]\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r + \mathbf{H}_r\mathbf{X}\mathbf{A}_r\right) + h_{4,r}(\mathbf{X})\Big(\mathbf{A}_r\mathbf{V}\mathbf{H}_r + \mathbf{H}_r\mathbf{V}\mathbf{A}_r\Big) \tag{198}$$

$$\begin{aligned}\nabla^E_{\mathbf{V}}\mathbf{F}_{5,s}(\mathbf{X}) = &\Big[D\left(h_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\Big]\Big[\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\Big]\\ &- h_{5,s}(\mathbf{X})\begin{bmatrix}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\\ +\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\\ +\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1} + \mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\end{bmatrix}\end{aligned} \tag{199}$$

$$\nabla_{\mathbf{V}}^{E}\mathbf{F}_{6,m}(\mathbf{X})=\left[D\left(h_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\right]\left[\begin{array}{l}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m+\mathbf{P}_m\mathbf{X}^{-1}\mathbf{Q}_m-\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\\-\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\end{array}\right]$$
$$+h_{6,m}(\mathbf{X})\left[\begin{array}{l}-\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{P}_m-\mathbf{P}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{Q}_m\\+\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}-\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\mathbf{Q}_m\mathbf{X}^{-1}\\+\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}+\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\\-\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{V}\mathbf{P}_m\mathbf{X}^{-1}+\mathbf{X}^{-1}\mathbf{Q}_m\mathbf{X}\mathbf{P}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\end{array}\right]\tag{200}$$

我们首先定义几个中间变量

$$T_{11p}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right);\qquad T_{12p}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{201}$$
$$T_{21q}=\mathrm{tr}\left(\mathbf{D}_q\mathbf{V}\right);\qquad T_{22q}=\mathrm{tr}\left(\mathbf{D}_q\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{202}$$
$$T_{31}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right);\qquad T_{32}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{203}$$
$$T_{41r}=\mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right);\qquad T_{42r}=\mathrm{tr}\left(\mathbf{A}_r\mathbf{V}\mathbf{H}_r\mathbf{V}\right);\tag{204}$$
$$T_{43r}=\text{平移}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{205}$$
$$T_{51s}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right);\qquad T_{52s}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{206}$$
$$T_{53s}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right);\tag{207}$$
$$T_{61m}=\mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right);\qquad T_{62m}=\mathrm{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right);\tag{208}$$
$$T_{63m}=\mathrm{tr}\left(\mathbf{P}_m\mathbf{V}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right);\qquad T_{64m}=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{209}$$
$$T_{65m}=\mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right);\tag{210}$$

这意味着

$$D\left(g_{1,p}\right)_{\mathbf{X}}(\mathbf{V})=-\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right)=-T_{11p}\tag{211}$$
$$D\left(g_{2,q}\right)_{\mathbf{X}}(\mathbf{V})=\mathrm{tr}\left(\mathbf{D}_q\mathbf{V}\right)=T_{21q}\tag{212}$$
$$D\left(g_{3}\right)_{\mathbf{X}}(\mathbf{V})=\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right)=T_{31}\tag{213}$$
$$D\left(g_{4,\ r}\right)_{\mathbf{X}}(\mathbf{V})=2\mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right)=2T_{41r}\tag{214}$$
$$D\left(g_{5,s}\right)_{\mathbf{X}}(\mathbf{V})=-2\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right)=-2T_{51s}\tag{215}$$
$$D\left(g_{6,m}\right)_{\mathbf{X}}(\mathbf{V})=\mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right)-\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\right)=T_{61m}-T_{62m}\tag{216}$$

由 (194) 可知♀♀♀♀♀♀

$$\begin{aligned}\Rightarrow\mathrm{tr}\left(Hf(\mathbf{X})[\mathbf{V}]\mathbf{V}\right)=&-\sum_{p=1}^{P}\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\right)+\sum_{q=1}^{Q}\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\right)+\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\right)\\&+\sum_{r=1}^{R}\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\right)-\sum_{s=1}^{S}\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\right)\\&+\frac{1}{2}\sum_{m=1}^{M}\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\right)\end{aligned}\tag{217}$$

方程（217）的每个分量都可以写成

$$\mathrm{tr}\left(\nabla_{\mathbf{V}}^{E}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\right)=\left[D\left(h_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\right]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right)-2h_{1,p}(\mathbf{X})\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\tag{218}$$

$$\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\right) = \Big[D\left(h_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{D}_q\mathbf{V}\right) \tag{219}$$

$$\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\right) = \Big[D\left(h_{3}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right) - h_3(\mathbf{X})\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \tag{220}$$

$$\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\right) = 2\Big[D\left(h_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right) + 2h_{4,r}(\mathbf{X})\mathrm{tr}\left(\mathbf{A}_r\mathbf{V}\mathbf{H}_r\mathbf{V}\right) \tag{221}$$

$$\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\right) = 2\Big[D\left(h_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right)$$

$$- h_{5,s}(\mathbf{X})\begin{bmatrix} 4\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \\ +2\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \end{bmatrix} \tag{222}$$

$$\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\right) = 2\Big[D\left(h_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\Big]\left[\mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right) - \mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\right)\right]$$

$$+ 4h_{6,m}(\mathbf{X})\begin{bmatrix} -\mathrm{tr}\left(\mathbf{P}_m\mathbf{V}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) \\ +\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \end{bmatrix} \tag{223}$$

结合 (218)-(223)，我们得到彩神争霸

$$\begin{aligned}
&\mathrm{tr}\left(Hf(\mathbf{X})[\mathbf{V}]\mathbf{V}\right) \\
&= -\sum_{p=1}^{P}\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\right) + \sum_{q=1}^{Q}\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\right) + \mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\right) \\
&+ \sum_{r=1}^{R}\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\right) - \sum_{s=1}^{S}\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\right) + \frac{1}{2}\sum_{m=1}^{M}\mathrm{tr}\left(\nabla^E_{\mathbf{V}}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\right)
\end{aligned}$$

$$= \begin{bmatrix}
-\sum_{p=1}^{P}\left[\Big[D\left(h_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right) - 2h_{1,p}(\mathbf{X})\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\right] \\
+\sum_{q=1}^{Q}\left[\Big[D\left(h_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{D}_q\mathbf{V}\right)\right] \\
+\left[\Big[D\left(h_{3}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right) - h_3(\mathbf{X})\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\right] \\
+\sum_{r=1}^{R}\left[2\left[D\left(h_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right]\mathrm{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right) + 2h_{4,r}(\mathbf{X})\mathrm{tr}\left(\mathbf{A}_r\mathbf{V}\mathbf{H}_r\mathbf{V}\right)\right] \\
-\sum_{s=1}^{S}\begin{bmatrix} 2\Big[D\left(h_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\Big]\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \\ -h_{5,s}(\mathbf{X})\begin{bmatrix} 4\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \\ +2\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \end{bmatrix} \end{bmatrix} \\
+\sum_{m=1}^{M}\begin{bmatrix} \Big[D\left(h_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\Big]\left[\mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right) - \mathrm{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right)\right] \\ +2h_{6,m}(\mathbf{X})\begin{bmatrix} -\mathrm{tr}\left(\mathbf{P}_m\mathbf{V}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) \\ +\mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \end{bmatrix} \end{bmatrix}
\end{bmatrix}$$

$$= \begin{bmatrix}
-\sum_{p=1}^{P}\left[\Big[D\left(h_{1,p}\right)_{\mathbf{X}}(\mathbf{V})\Big]T_{11p} - 2h_{1,p}(\mathbf{X})T_{12p}\right] + \sum_{q=1}^{Q}\left[\Big[D\left(h_{2,q}\right)_{\mathbf{X}}(\mathbf{V})\Big]T_{21q}\right] \\
+\left[\Big[D\left(h_{3}\right)_{\mathbf{X}}(\mathbf{V})\Big]T_{31} - h_3(\mathbf{X})T_{32}\right] + \sum_{r=1}^{R}\left[2\left[D\left(h_{4,r}\right)_{\mathbf{X}}(\mathbf{V})\right]T_{41r} + 2h_{4,r}(\mathbf{X})T_{42r}\right] \\
-\sum_{s=1}^{S}\left[2\Big[D\left(h_{5,s}\right)_{\mathbf{X}}(\mathbf{V})\Big]T_{51s} - h_{5,s}(\mathbf{X})\left[4T_{52s} + 2T_{53s}\right]\right] \\
+\sum_{m=1}^{M}\left[\Big[D\left(h_{6,m}\right)_{\mathbf{X}}(\mathbf{V})\Big]\left[T_{61m} - T_{62m}\right] + 2h_{6,m}(\mathbf{X})\left[-T_{63m} + T_{64m}\right]\right]
\end{bmatrix} \tag{224}$$

接下来我们计算

$$\begin{aligned}\operatorname{tr}\left(\mathbf{V}\operatorname{grad} f(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = &-\sum_{p=1}^{P}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\sum_{q=1}^{Q}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\\ &+\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\sum_{r=1}^{R}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\\ &-\sum_{s=1}^{S}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\sum_{m=1}^{M}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\end{aligned}\tag{225}$$

其各个分量均可简化为

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = h_{1,p}(\mathbf{X})\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\tag{226}$$

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = h_{2,q}(\mathbf{X})\operatorname{tr}\left(\mathbf{D}_q\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\tag{227}$$

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = h_{3}(\mathbf{X})\operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right)\tag{228}$$

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = 2h_{4,r}(\mathbf{X})\operatorname{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\tag{229}$$

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = 2h_{5,s}(\mathbf{X})\Big[\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\Big]\tag{230}$$

$$\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) = 2h_{6,m}(\mathbf{X})\begin{bmatrix}\operatorname{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\\ -\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\end{bmatrix}\tag{231}$$

结合式 (226)–(231)，可得

$$\begin{aligned}\operatorname{tr}\left(\mathbf{V}\operatorname{grad} f(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right) &= \begin{bmatrix}-\sum_{p=1}^{P}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{1,p}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\sum_{q=1}^{Q}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{2,q}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\\ +\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{3}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\sum_{r=1}^{R}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{4,r}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\\ -\sum_{s=1}^{S}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{5,s}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)+\frac{1}{2}\sum_{m=1}^{M}\operatorname{tr}\left(\mathbf{V}\mathbf{F}_{6,m}(\mathbf{X})\mathbf{V}\mathbf{X}^{-1}\right)\end{bmatrix}\\ &= \begin{bmatrix}-\sum_{p=1}^{P}\left[h_{1,p}(\mathbf{X})\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\right]\\ +\sum_{q=1}^{Q}\left[h_{2,q}(\mathbf{X})\operatorname{tr}\left(\mathbf{D}_q\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\right]+\left[h_3(\mathbf{X})\operatorname{tr}\left(\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right)\right]\\ +\sum_{r=1}^{R}\left[2h_{4,r}(\mathbf{X})\operatorname{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\right]\\ -\sum_{s=1}^{S}\left[2h_{5,s}(\mathbf{X})\Big[\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\Big]\right]\\ +\sum_{m=1}^{M}h_{6,m}(\mathbf{X})\begin{bmatrix}\operatorname{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\\ -\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right)\end{bmatrix}\end{bmatrix}\\ &= \begin{bmatrix}-\sum_{p=1}^{P}\left[h_{1,p}(\mathbf{X})T_{12p}\right]+\sum_{q=1}^{Q}\left[h_{2,q}(\mathbf{X})T_{22q}\right]+\left[h_3(\mathbf{X})T_{32}\right]\\ +\sum_{r=1}^{R}\left[2h_{4,r}(\mathbf{X})T_{43r}\right]-\sum_{s=1}^{S}\left[2h_{5,s}(\mathbf{X})T_{52s}\right]\\ +\sum_{m=1}^{M}h_{6,m}(\mathbf{X})\left[T_{65m}-T_{64m}\right]\end{bmatrix}\end{aligned}\tag{232}$$

最后，我们进行计算。

$$\begin{aligned}\operatorname{tr}\left(\mathbf{F}(\mathbf{X})\mathbf{V}\right) &= \operatorname{tr}\left(\operatorname{grad} f(\mathbf{X})\mathbf{V}\right)\\ &= \begin{bmatrix}-\sum_{p=1}^{P}h_{1,p}(\mathbf{X})\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right)+\sum_{q=1}^{Q}h_{2,q}(\mathbf{X})\operatorname{tr}\left(\mathbf{D}_q\mathbf{V}\right)+h_3(\mathbf{X})\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right)\\ +\sum_{r=1}^{R}2h_{4,r}(\mathbf{X})\operatorname{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right)-\sum_{s=1}^{S}2h_{5,s}(\mathbf{X})\operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right)\\ +\sum_{m=1}^{M}h_{6,m}(\mathbf{X})\left(\ \operatorname{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\right)-\operatorname{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right)\ \right)\end{bmatrix}\end{aligned}$$

$$= \begin{bmatrix} -\sum_{p=1}^{P} h_{1,p}(\mathbf{X}) T_{11p} + \sum_{q=1}^{Q} h_{2,q}(\mathbf{X}) T_{21q} + h_3(\mathbf{X}) T_{31} + \sum_{r=1}^{R} 2h_{4,r}(\mathbf{X}) T_{41r} \\ -\sum_{s=1}^{S} 2h_{5,s}(\mathbf{X}) T_{51s} + \sum_{m=1}^{M} h_{6,m}(\mathbf{X}) \left[T_{61m} - T_{62m}\right] \end{bmatrix} \tag{233}$$

当$\mathbf{V}$ 为基向量时，所有上述计算都可以进一步简化，如下所示。我们将其分为两种情况:

情形 $i \neq j$。

$$\mathbf{V} = \frac{1}{\sqrt{2}} \mathbf{B} \left(\mathbf{e}_i \mathbf{e}_j^\mathsf{T} + \mathbf{e}_j \mathbf{e}_i^\mathsf{T} \right) \mathbf{B}^\mathsf{T} \tag{234}$$

$$T_{11p} = \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\right) = \sqrt{2}\Big[\mathbf{M}_{1,p}(\mathbf{X})\Big]_{ij} \tag{235}$$

$$T_{12p} = \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{C}_p\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = \frac{1}{2}\Big(\left[\mathbf{M}_{1,p}(\mathbf{X})\right]_{ii} + \left[\mathbf{M}_{1,p}(\mathbf{X})\right]_{jj} \Big) \tag{236}$$

$$T_{21q} = \operatorname{tr}\left(\mathbf{D}_q\mathbf{V}\right) = \sqrt{2}\Big[\mathbf{M}_{2,q}(\mathbf{X})\Big]_{ij} \tag{237}$$

$$T_{22q} = \operatorname{tr}\left(\mathbf{D}_q\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = \frac{1}{2}\Big(\left[\mathbf{M}_{2,q}(\mathbf{X})\right]_{ii} + \left[\mathbf{M}_{2,q}(\mathbf{X})\right]_{jj} \Big) \tag{238}$$

$$T_{31} = \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{V}\right) = 0 \tag{239}$$

$$T_{32} = \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = 1 \tag{240}$$

$$T_{41r} = \operatorname{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\right) = \frac{1}{\sqrt{2}}\Big(\Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{j:} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{:i} + \Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{i:} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{:j} \Big) \tag{241}$$

$$\begin{aligned} T_{42r} &= \operatorname{tr}\left(\mathbf{A}_r\mathbf{V}\mathbf{H}_r\mathbf{V}\right) \\ &= \frac{1}{2} \begin{pmatrix} 2\Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{ij} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{ij} + \Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{ii} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{jj} \\ + \Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{jj} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{ii} \end{pmatrix} \end{aligned} \tag{242}$$

$$T_{43r} = \operatorname{tr}\left(\mathbf{A}_r\mathbf{X}\mathbf{H}_r\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = \frac{1}{2}\Big(\Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{i:} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{:i} + \Big[\mathbf{M}_{4,1,r}(\mathbf{X})\Big]_{j:} \Big[\mathbf{M}_{4,2,r}(\mathbf{X})\Big]_{:j} \Big) \tag{243}$$

$$\begin{aligned} T_{51s} &= \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \\ &= \frac{1}{\sqrt{2}}\Big(\Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{i:} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{:j} + \Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{j:} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{:i} \Big) \end{aligned} \tag{244}$$

$$\begin{aligned} T_{52s} &= \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) \\ &= \frac{1}{2}\Big(\Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{i:} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{:i} + \Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{j:} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{:j} \Big) \end{aligned} \tag{245}$$

$$\begin{aligned} T_{53s} &= \operatorname{tr}\left(\mathbf{X}^{-1}\mathbf{F}_s\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{G}_s\mathbf{X}^{-1}\mathbf{V}\right) \\ &= \frac{1}{2} \begin{pmatrix} 2\Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{ij} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{ij} + \Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{ii} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{jj} \\ + \Big[\mathbf{M}_{5,1,s}(\mathbf{X})\Big]_{jj} \Big[\mathbf{M}_{5,2,s}(\mathbf{X})\Big]_{ii} \end{pmatrix} \end{aligned} \tag{246}$$

$T_{61m} =$ 土耳其语 $(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V})$

$$= \frac{1}{\sqrt{2}} \left(\left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{:j} \right]^{\mathsf{T}} [\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{:i} + \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{:i} \right]^{\mathsf{T}} [\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{:j} \right) \quad (247)$$

$$T_{62m} = \mathrm{tr}\left(\mathbf{P}_m \mathbf{X} \mathbf{Q}_m \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1}\right)$$

$$= \frac{1}{\sqrt{2}} \left([\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{j:} \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{i:} \right]^{\mathsf{T}} + [\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{i:} \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{j:} \right]^{\mathsf{T}} \right) \quad (248)$$

$$T_{63m} = \mathrm{tr}\left(\mathbf{P}_m \mathbf{V} \mathbf{Q}_m \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1}\right)$$

$$= \frac{1}{2} \mathrm{tr} \left(\begin{array}{l} [\mathbf{M}_{6,1,m}(\mathbf{X})]_{ji} [\mathbf{M}_{6,2,m}(\mathbf{X})]_{ij} + [\mathbf{M}_{6,1,m}(\mathbf{X})]_{ii} [\mathbf{M}_{6,2,m}(\mathbf{X})]_{jj} \\ + [\mathbf{M}_{6,1,m}(\mathbf{X})]_{jj} [\mathbf{M}_{6,2,m}(\mathbf{X})]_{ii} + [\mathbf{M}_{6,1,m}(\mathbf{X})]_{ij} [\mathbf{M}_{6,2,m}(\mathbf{X})]_{ji} \end{array} \right) \quad (249)$$

$$T_{64m} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{P}_m \mathbf{X} \mathbf{Q}_m \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right)$$

$$= \frac{1}{2} \left([\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{i:} \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{i:} \right]^{\mathsf{T}} + [\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{j:} \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{j:} \right]^{\mathsf{T}} \right) \quad (250)$$

$$T_{65m} = \mathrm{tr}\left(\mathbf{Q}_m \mathbf{X}^{-1} \mathbf{P}_m \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right)$$

$$= \frac{1}{2} \left(\left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{:i} \right]^{\mathsf{T}} \left[[\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{:i} \right] + \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{:j} \right]^{\mathsf{T}} \left[[\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{:j} \right] \right) \quad (251)$$

情形 $i = j_\circ$

$$\mathbf{V} = \mathbf{B} \mathbf{e}_i \mathbf{e}_i^{\mathsf{T}} \mathbf{B}^{\mathsf{T}} \quad (252)$$

$$T_{11p} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{C}_p \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{1,p}(\mathbf{X}) \right]_{ii} \quad (253)$$

$$T_{12p} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{C}_p \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{1,p}(\mathbf{X}) \right]_{ii} \quad (254)$$

$$T_{21q} = \mathrm{tr}\left(\mathbf{D}_q \mathbf{V}\right) = \left[\mathbf{M}_{2,q}(\mathbf{X}) \right]_{ii} \quad (255)$$

$$T_{22q} = \mathrm{tr}\left(\mathbf{D}_q \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{2,q}(\mathbf{X}) \right]_{ii} \quad (256)$$

<escape>$T_{31} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{V}\right) = 1$</escape> (257)

$$T_{32} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right) = 1 \quad (258)$$

$$T_{41r} = \mathrm{tr}\left(\mathbf{A}_r \mathbf{X} \mathbf{H}_r \mathbf{V}\right) = \left[\mathbf{M}_{4,1,r}(\mathbf{X}) \right]_{i:} \left[\mathbf{M}_{4,2,r}(\mathbf{X}) \right]_{:i} \quad (259)$$

$$T_{42r} = \mathrm{tr}\left(\mathbf{A}_r \mathbf{V} \mathbf{H}_r \mathbf{V}\right) = \left[\mathbf{M}_{4,1,r}(\mathbf{X}) \right]_{ii} \left[\mathbf{M}_{4,2,r}(\mathbf{X}) \right]_{ii} \quad (260)$$

$$T_{43r} = \mathrm{tr}\left(\mathbf{A}_r \mathbf{X} \mathbf{H}_r \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{4,1,r}(\mathbf{X}) \right]_{i:} \left[\mathbf{M}_{4,2,r}(\mathbf{X}) \right]_{:i} \quad (261)$$

$$T_{51s} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{F}_s \mathbf{X}^{-1} \mathbf{G}_s \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{5,1,s}(\mathbf{X}) \right]_{i:} \left[\mathbf{M}_{5,2,s}(\mathbf{X}) \right]_{:i} \quad (262)$$

$$T_{52s} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{F}_s \mathbf{X}^{-1} \mathbf{G}_s \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{5,1,s}(\mathbf{X}) \right]_{i:} \left[\mathbf{M}_{5,2,s}(\mathbf{X}) \right]_{:i} \quad (263)$$

$$T_{53s} = \mathrm{tr}\left(\mathbf{X}^{-1} \mathbf{F}_s \mathbf{X}^{-1} \mathbf{V} \mathbf{X}^{-1} \mathbf{G}_s \mathbf{X}^{-1} \mathbf{V}\right) = \left[\mathbf{M}_{5,1,s}(\mathbf{X}) \right]_{ii} \left[\mathbf{M}_{5,2,s}(\mathbf{X}) \right]_{ii} \quad (264)$$

$$T_{61m} = \mathrm{tr}\left(\mathbf{Q}_m \mathbf{X}^{-1} \mathbf{P}_m \mathbf{V}\right) = \left[[\mathbf{M}_{6,2,m}(\mathbf{X}_t)]_{:i} \right]^{\mathsf{T}} [\mathbf{M}_{6,1,m}(\mathbf{X}_t)]_{:i} \quad (265)$$

$$T_{62m} = \mathrm{tr}\left(\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) = \left[\mathbf{M}_{6,1,m}(\mathbf{X}_t)\right]_{i:}\left[\left[\mathbf{M}_{6,2,m}(\mathbf{X}_t)\right]_{i:}\right]^{\mathsf{T}} \tag{266}$$

$$T_{63m} = \mathrm{tr}\left(\mathbf{P}_m\mathbf{V}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\right) = \left[\mathbf{M}_{6,2,m}(\mathbf{X})\right]_{ii}\left[\mathbf{M}_{6,1,m}(\mathbf{X})\right]_{ii} \tag{267}$$

$$T_{64m} = \mathrm{tr}\left(\mathbf{X}^{-1}\mathbf{P}_m\mathbf{X}\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = \left[\mathbf{M}_{6,1,m}(\mathbf{X})\right]_{i:}\left[\left[\mathbf{M}_{6,2,m}(\mathbf{X})\right]_{i:}\right]^{\mathsf{T}} \tag{268}$$

$$T_{65m} = \mathrm{tr}\left(\mathbf{Q}_m\mathbf{X}^{-1}\mathbf{P}_m\mathbf{V}\mathbf{X}^{-1}\mathbf{V}\right) = \left[\left[\mathbf{M}_{6,2,m}(\mathbf{X})\right]_{:i}\right]^{\mathsf{T}}\left[\mathbf{M}_{6,1,m}(\mathbf{X})\right]_{:i} \tag{269}$$

附录 C. 两个对称正定矩阵的矩阵几何平均

其解　　矩阵几何平均问题 $\min_{\mathbf{X}\in\mathbb{P}^n}\sum_{i=1}^{2}\left\|\log\left(\mathbf{X}^{-1/2}\mathbf{W}_i\mathbf{X}^{-1/2}\right)\right\|_F^2$ 为

$$\mathbf{X}_g^\star = \mathbf{W}_1^{\frac{1}{2}}\left(\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right)^{\frac{1}{2}}\mathbf{W}_1^{\frac{1}{2}} \tag{270}$$

还要注意,

$$\mathbf{W}_2 = \mathbf{W}_1^{\frac{1}{2}}\left(\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right)\mathbf{W}_1^{\frac{1}{2}} \tag{271}$$

函数 $f(\mathbf{X}) = \sum_{i=1}^{2}\mathrm{tr}\left(\mathbf{W}_i\mathbf{X}^{-1} + \mathbf{W}_i^{-1}\mathbf{X}\right)$ 对于 $\mathbf{W}_i \succ 0$ 是测地线强凸的，其梯度为

$$\mathrm{grad} f(\mathbf{X}) = \sum_{i=1}^{2}\left(-\mathbf{X}^{-1}\mathbf{W}_i\mathbf{X}^{-1} + \mathbf{W}_i^{-1}\right) \tag{272}$$

在　　最小化元 $\mathbf{X}^*$ 的 $f(\mathbf{X})$，$\mathrm{grad} f(\mathbf{X}^*) = 0$。相反，我们来计算在 $\mathbf{X}$ 处的 $\mathrm{grad} f(\mathbf{X})$　$_g^\star$

$$\mathrm{grad} f(\mathbf{X}_g^\star) = \sum_{i=1}^{2}\left(-(\mathbf{X}_g^\star)^{-1}\mathbf{W}_i(\mathbf{X}_g^\star)^{-1} + \mathbf{W}_i^{-1}\right) \tag{273}$$

$$= -(\mathbf{X}_g^\star)^{-1}\mathbf{W}_1(\mathbf{X}_g^\star)^{-1} + \mathbf{W}_1^{-1} - (\mathbf{X}_g^\star)^{-1}\mathbf{W}_2(\mathbf{X}_g^\star)^{-1} + \mathbf{W}_2^{-1} \tag{274}$$

$$= -\mathbf{W}_1^{-\frac{1}{2}}\left[\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right]^{-\frac{1}{2}}\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_1\mathbf{W}_1^{-\frac{1}{2}}\left[\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right]^{-\frac{1}{2}}\mathbf{W}_1^{-\frac{1}{2}} + \mathbf{W}_1^{-1}$$
$$-\begin{bmatrix}\mathbf{W}_1^{-\frac{1}{2}}\left(\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right)^{-\frac{1}{2}}\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_1^{\frac{1}{2}}\left(\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right)\mathbf{W}_1^{\frac{1}{2}} \\ \times\mathbf{W}_1^{-\frac{1}{2}}\left(\mathbf{W}_1^{-\frac{1}{2}}\mathbf{W}_2\mathbf{W}_1^{-\frac{1}{2}}\right)^{-\frac{1}{2}}\mathbf{W}_1^{-\frac{1}{2}}\end{bmatrix} + \mathbf{W}_2^{-1} \tag{275}$$

$$= -\mathbf{W}_2^{-1} + \mathbf{W}_1^{-1} - \mathbf{W}_1^{-1} + \mathbf{W}_2^{-1} = 0 \tag{276}$$

因此，由于 $f(\mathbf{X})$ 具有测地线强凸性，两个对称正定（SPD）矩阵的矩阵几何平均是函数 $f(\mathbf{X})$ 的唯一最小化元。

参考文献

P.A. Absil、R. Mahony 和 R. Sepulchre。矩阵流形上的优化算法。普林斯顿大学出版社，2008 年。

Kwangjun Ahn 和 Suvrit Sra。从 Nesterov 的估计序列到黎曼加速。载于学习理论会议，第 84–118 页。PMLR，2020年。

Foivos Alimisis、Antonio Orvieto、Gary Becigneul 和 Aurelien Lucchi。动量改进了黎曼流形上的优化。载于 Arindam Banerjee 和 Kenji Fukumizu（编），第 *24* 届国际人工智能与统计学会议论文集，机器学习研究论文集第 130 卷，第 1351–1359 页。PMLR，2021 年 4 月 13–15 日。网址：https://proceedings.mlr.press/v130/alimisis21a. html。

Reza Babanezhad、Issam H. Laradji、Alireza Shafaei 和 Mark Schmidt。MASAGA：一种用于流形优化的线性收敛随机一阶方法。载于 Michele Berlingerio、Francesco Bonchi、Thomas G¨artner、Neil Hurley 和 Georgiana Ifrim（编），数据库中的机器学习与知识发现，第 344–359 页，Cham，2019年。Springer International Publishing。

Gary B´ecigneul 和 Octavian-Eugen Ganea。黎曼自适应优化方法。*arXiv* 预印本 *arXiv:1810.00760*，2018年。

Amir Beck、Edouard Pauwels 和 Shoham Sabach。凸优化问题的循环分块条件梯度法。*SIAM* 优化杂志，25(4)：2024– 2049，2015年。

Hande Y Benson 和 Robert J Vanderbei。使用非线性规划的内点法求解具有半定约束及相关约束的问题。数学规划，95：279–302，2003年。

D.P. Bertsekas。非线性规划。Athena 科学优化与计算丛书。Athena Scientific，2016年。

R. Bhatia。正定矩阵。普林斯顿应用数学丛书。普林斯顿大学出版社，2007年。ISBN 9780691129181。

Rajendra Bhatia。正矩阵的黎曼均值。载于矩阵信息几何，第35–51页。Springer，2013年。

Rajendra Bhatia 和 John Holbrook。黎曼几何与矩阵几何平均。线性代数及其应用，413(2):594 – 618，2006年。ISSN 0024-3795。doi: https: // doi.org/10.1016/j.laa.2005.08.025。URL http://www.sciencedirect.com/science/article/pii/S0024379505004350。国际线性代数协会第 11 届会议特刊，科英布拉，2004 年。

S. Bonnabel. 黎曼流形上的随机梯度下降。*IEEE* 自动控制汇刊, 58(9):2217–2229, 2013年。

N. Boumal。《光滑流形上的优化导论》。剑桥大学出版社，2023年。ISBN 9781009178716。网址：https://books.google.co.in/books?id=fO2wEAAAQBAJ。Nicolas Boumal。流形上的优化与估计。2014年。M.R. Bridson 和 A. H¨afliger。《非正曲率度量空间》。数学科学基础丛书。Springer Berlin Heidelberg，2011年。ISBN 9783540643241。Elena Celledoni 和 Simone Fiori。齐性流形上的优化下降方法。《数学与计算机模拟》，79(4):1298–1323，2008年。Olivier Fercoq 和 Peter Richtarik。加速、并行和近端坐标下降。《style id='9'>SIAM 优化杂志》，25(4):1997–2023，2015年。Orizon P Ferreira、Mauricio S Louzeiro 和 LF4018420 Prudente。具有曲率下界的黎曼流形上的优化梯度法。《style id='11'>SIAM 优化杂志》，29(4):2517–2541，2019年。Orizon P Ferreira、Maur´ıcio S Louzeiro 和 Leandro F Prudente。黎曼流形优化的一阶方法。载于《非线性几何数据的变分方法手册》，第499–525页。Springer，2020年。Mohammad Hamed Firouzehtarash 和 Reshad Hosseini。用于低多线性秩逼近的黎曼预条件坐标下降。《*arXiv* 预印本 *arXiv:2109.01632*》，2021年。Kimon Fountoulakis 和 Rachael Tappenden。一种灵活的坐标下降方法。《计算优化与应用》，70(2):351–394，2018年。Bin Gao、Xin Liu、Xiaojun Chen 和 Ya-xiang Yuan。一种用于带正交约束优化问题的新型一阶算法框架。《*SIAM* 优化杂志》，28(1):302–332，2018年。Evan S Gawlik。矩阵平方根的 Zolotarev 迭代。《*SIAM* 矩阵分析与应用杂志》，40(2):696–719，2019年。A.K. Gupta 和 D.K. Nagar。《矩阵变量分布》。纯数学与应用数学专著与综述。Taylor & Francis，1999年。ISBN 9781584880462。网址：https://books.google.co.in/books?id=PQOYnT7P1loC。David H Gutman 和 Nam Ho-Nguyen。无坐标的坐标下降：黎曼流形上的切子空间下降。《运筹学数学》，2022年。Nicholas J Higham。矩阵平方根的稳定迭代。《数值算法》，15：227–242，1997年。Reshad Hosseini 和 Suvrit Sra。高斯混合模型的矩阵流形优化。载于 C. Cortes、N. D. Lawrence、D. D. Lee、M. Sugiyama 和 R. Garnett 编，《神经信息处理系统进展 *28*》，第910–918页。Curran Associates, Inc.，2015年。

Reshad Hosseini 和 Suvrit Sra。随机黎曼优化的最新进展。载于非线性几何数据变分方法手册，第 527–554 页。Springer，2020年。Minhui Huang、Shiqian Ma 和 Lifeng Lai。用于计算投影鲁棒 Wasserstein 距离的黎曼分块坐标下降方法。载于 Marina Meila 和 Tong Zhang（编），第 *38* 届国际机器学习会议论文集，第 139 卷，机器学习研究论文集，第 4446–4455 页。PMLR，2021 年 7 月 18–24 日。网址：https://proceedings.mlr.press/v139/huang21e.html。Prateek Jain、Chi Jin、Sham Kakade 和 Praneeth Netrapalli。用于计算矩阵平方根的非凸梯度下降的全局收敛性。载于人工智能与统计学，第 479–488 页。PMLR，2017年。Hamed Karimi、Julie Nutini 和 Mark Schmidt。在 Polyak-Łojasiewicz 条件下梯度方法和近端梯度方法的线性收敛。载于机器学习与数据库知识发现联合欧洲会议，第 795–811 页。Springer，2016年。Hiroyuki Kasai、Hiroyuki Sato 和 Bamdev Mishra。Grassmann 流形上的黎曼随机方差缩减梯度。*arXiv* 预印本 *arXiv:1605.07367*，2016年。Hiroyuki Kasai、Pratik Jawanpuria 和 Bamdev Mishra。矩阵流形上的黎曼自适应随机梯度算法。载于国际机器学习会议，第 3262–3271 页。PMLR，2019年。J.M. Lee。黎曼流形导论。《数学研究生教材》。Springer，2018年。ISBN 9783319917559。Fuxin Li、Yunshan Fu、Yu-Hong Dai、Cristian Sminchisescu 和 Jue Wang。通过无约束优化进行核学习。载于人工智能与统计学，第 328–335 页。PMLR，2009年。Yuanyuan Liu、Fanhua Shang、James Cheng、Hong Cheng 和 Licheng Jiao。黎曼流形上测地凸优化的加速一阶方法。载于 I. Guyon、U. V. Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编），神经信息处理系统进展 *30*，第 4868 – 4877 页。Curran Associates, Inc.，2017年。D.G. Luenberger 和 Y. Ye。线性与非线性规划（第四版）。《运筹学与管理科学国际丛书》。Springer International Publishing，2015年。Maher. Moakher。对称正定矩阵几何平均的微分几何方法。*SIAM* 矩阵分析与应用杂志，26(3)：735–747，2005年。Yu. Nesterov。大规模优化问题中坐标下降方法的效率。*SIAM* 优化杂志，22(2)：341–362，2012年。

Harry F Oviedo、Hugo J Lara 和 Oscar S Dalmau。用于计算矩阵平方根的缩放不动点算法。*arXiv* 预印本 *arXiv:2002.08471*，2020年。T. Rapcs´ak。中的光滑非线性优化。《非凸优化及其应用》。Springer US，1997年。ISBN 9780792346807。Peter Richt´arik 和 Martin Tak´ac。用于最小化复合函数的随机分块坐标下降方法的迭代复杂度。数学规划，144 (1-2)：1–38，2014年。Soumava Kumar Roy、Zakaria Mhammedi 和 Mehrtash Harandi。面向几何的深度学习约束优化技术。载于*IEEE* 计算机视觉与模式识别会议论文集，第 4460–4469 页，2018年。Ankan Saha 和 Ambuj Tewari。循环坐标下降方法的非渐近收敛性。*SIAM* 优化杂志，23(1)：576–601，2013年。George A. F. Seber。统计学家矩阵手册。《Wiley 概率与数理统计丛书》。Wiley，2008年。Uri Shalit 和 Gal Chechik。通过 Givens 旋转学习正交矩阵的坐标下降方法。载于国际机器学习会议，第 548–556 页。PMLR，2014年。Suvrit Sra。通过几何优化求矩阵平方根。*arXiv* 预印本 *arXiv:1507.08366*，2015年。Suvrit Sra 和 Reshad Hosseini。椭圆等高分布正定矩阵上的几何优化。神经信息处理系统进展，26，2013年。Suvrit. Sra 和 Reshad. Hosseini。正定矩阵流形上的锥几何优化。*SIAM* 优化杂志，25(1)：713–739，2015年。Nilesh Tripuraneni、Nicolas Flammarion、Francis Bach 和 Michael I Jordan。黎曼流形上的随机梯度下降平均化。载于学习理论会议，第 650–687 页。PMLR，2018年。Melanie Weber 和 Suvrit Sra。通过 Frank-Wolfe 方法进行黎曼优化，2020年。A. Wiesel。测地凸性与协方差估计。*IEEE* 信号处理汇刊，60(12)：6182–6189，2012年。Ami Wiesel、Teng Zhang 等。结构化鲁棒协方差估计。基础与趋势® 信号处理，8(3)：127–216，2015年。Hongyi Zhang 和 Suvrit Sra。测地凸优化的一阶方法。载于 Vitaly Feldman、Alexander Rakhlin 和 Ohad Shamir（编），第 *29* 届年度学习理论会议，第 49 卷，机器学习研究论文集，第 1617 – 1638 页，美国纽约州纽约市哥伦比亚大学，2016 年 6 月 23–26 日。PMLR。

Hongyi Zhang、Sashank J. Reddi 和 Suvrit Sra。黎曼 SVRG：黎曼流形上的快速随机优化。载于 D. D. Lee、M. Sugiyama、U. V. Luxburg、I. Guyon 和 R. Garnett（编），神经信息处理系统进展第*29*卷，第 4592–4600 页。CurranAssociates,Inc.，2016年。

Hongyi Zhang、Suvrit Sra 和 Suvrit Sra。测地凸优化的估计序列。载于 S´ebastien Bubeck、Vianney Perchet 和 Philippe Rigollet 编，第 *31* 届学习理论会议论文集，第 75 卷，机器学习研究论文集，第 1703–1723 页。PMLR，2018 年 7 月 6–9 日。URL https: //proceedings.mlr.press/v75/zhang18a.html。

Jingzhao Zhang、Hongyi Zhang 和 Suvrit Sra。R-spider：一种具有与曲率无关收敛速率的快速黎曼随机优化算法。*arXiv* 预印本 *arXiv:1811.04194*，2018b。

Teng Zhang、Ami Wiesel 和 Maria Sabrina Greco。多元广义高斯分布：凸性与图模型。*IEEE* 信号处理汇刊，61(16):4141–4148，2013年。