arXiv:2601.00938v1 [cs.CL] 2 Jan 2026

低秩黎曼更新下压缩查询委托的率失真分析

FARUK ALPAY 和 BUĞRA KILIÇTAŞ

摘要。有限上下文智能体在中间推理超出有效工作记忆预算时会失效。我们研究压缩查询委托*(CQD)*：(i) 将高维潜在推理状态压缩为低秩张量查询；(ii) 将最精简的查询委托给外部预言机；(iii) 通过在固定秩流形上进行黎曼优化来更新潜在状态。我们给出一种以数学为先的表述：CQD 是一种带有查询预算泛函的约束随机规划，其中预言机被建模为含噪算子。我们将 CQD 与经典率失真理论和信息瓶颈原理联系起来，表明谱硬阈值化对于一种自然的约束二次失真问题是最优的，并在预言机噪声有界且满足光滑性假设的条件下，推导出黎曼随机逼近的收敛保证。在实证部分，我们报告了：(A) 一个包含2500项有限上下文推理测试集（源自 BBH 的任务及精选悖论实例），在计算量和上下文固定的条件下，将 CQD 与思维链基线进行比较；以及 (B) 一项人类“认知镜像”基准测试 ($N = 200$)，用于衡量不同现代预言机的认知增益和语义漂移。

1. 引言

令 W 表示智能体的有效上下文窗口，$C(P)$ 表示问题实例的信息复杂度 P。当 $C(P) > W$ 时，逐步提示可能导致截断和漂移 [18, 19]。我们将这种不匹配建模为一个信息分配问题：智能体必须在严格的令牌／计算预算下，传达潜在推理状态的充分统计量。CQD 强制执行显式的压缩–委托–更新循环。

一个核心论点是，中间推理通常具有结构化冗余（低秩、近似稀疏性、多线性相关性）。因此，冗长推理轨迹并不是传递状态的唯一方式：可以传递强调主导谱成分的压缩表示，这与经典的压缩和补全范式一致 [1, 2, 9, 10]。因此，CQD 不仅仅是一种提示技巧：它是在受约束假设类上的优化过程。

贡献。我们提供：

- 一种 CQD 的约束随机优化表述，其中包含显式的查询预算泛函。
- 率失真／信息瓶颈解释，并给出了二次失真下谱硬阈值化的最优性结果。
- 低秩张量几何（固定多线性秩）模型，以及基于回缩的黎曼更新方案。
- 预言机噪声下黎曼随机逼近的收敛定理（完整证明见附录 A）。
- 两项评估（一套包含 2,500 个条目的有界上下文测试集，以及一项 $N = 200$ 人类基准测试）和一项漂移—增益尺度分析。

2. 相关工作

信息论与压缩。香农的不确定性降低 [1] 和率失真理论 [2]对表征成本与保真度之间的权衡进行了形式化。信息瓶颈（IB）形式化了

关键词和短语。率失真、信息瓶颈、张量截断、固定秩流形、黎曼随机逼近、有界上下文推理。

提取最小充分表示 [3]。CQD 可被视为一种受约束的表示/通信方案。

低秩近似。Eckart–Young–Mirsky 定理确立了截断奇异值分解在 Frobenius 范数和谱范数下的最优性 [4,5]。张量分解和截断高阶奇异值分解是经典方法 [9,10]。扰动结果（Davis–Kahan 定理、Wedin 定理）刻画了奇异子空间的稳定性 [6–8]。

矩阵补全与阈值处理。核范数补全保证 [11]，奇异值阈值处理方法 [12] 为谱收缩和截断提供了依据；张量补全的变体见 [13]。

黎曼优化。流形优化的基础包括 Absil–Mahony–Sepulchre [14]、Edelman–Arias–Smith [15]，和 Boumal [17]。Bonnabel [16] 及后续研究讨论了黎曼随机逼近和随机梯度下降。

大语言模型推理评测。BIG-bench/BBH [20,21]、HELM 报告 [22]，以及思维链/自一致性 [18,19]。

3. 预备知识与符号

三阶张量为 $\mathbf{X} \in \mathbb{R}^{I\times J\times K}$，其元素为 x_{ijk}。模-n 展开为 $\mathbf{X}_{(n)}$ [9]。n模乘积为 $\mathbf{Y} = \mathbf{X} \times_n \mathbf{A}$，其中 $\mathbf{A}$ 沿模 n 对纤维进行乘法。

定义 3.1 （多线性秩）。$\mathbf{X}$ 的多线性秩为 $\boldsymbol{r} = (r_1, r_2, r_3)$，其中 $r_n = \operatorname{rank}(\mathbf{X}_{(n)})$.

我们使用 $\|\cdot\|_F$ 表示 Frobenius 范数，使用 $\|\cdot\|_2$ 表示谱范数，并使用 $\sigma_i(\mathbf{M})$ 表示按非递增顺序排列的奇异值。对于流形 $\mathcal{M}$，$\mathrm{T}_x\mathcal{M}$ 表示切空间，Retr_x 是一个回缩 [14,17]。斯蒂费尔流形为 $\mathrm{St}(p,n) = \{U \in \mathbb{R}^{n\times p} : U^\top U = I_p\}$。

4. 作为受约束率失真/信息瓶颈的 CQD

CQD 在预算约束下将潜在状态压缩为查询，然后利用预言机降低误差/不确定性。这与率失真相一致：

$$\min_{Q(\cdot)} \mathbb{E}\big[d(\mathcal{X}, \widehat{\mathcal{X}})\big] \quad \text{s.t.} \quad \mathbb{E}[\mathcal{B}(Q)] \le \tau, \tag{1}$$

其中，d 表示失真，$\mathbf{X}$ 是用于委托的表示。标准的拉格朗日形式为

$$\min_{Q(\cdot)} \mathbb{E}[d(\mathcal{X}, \widehat{\mathcal{X}})] + \lambda\, \mathbb{E}[\mathcal{B}(Q)]. \tag{2}$$

在信息瓶颈（IB）中，目标是在保留任务相关信息 $I(Q;Y)$ [3] 的约束下，最小化 $I(\mathbf{X};Q)$。CQD 可被视为生成 Q，在最小化通信成本的同时保留任务显著的成分（主要谱成分）。

在本文中，我们专门研究一种基于谱截断（ASM）的确定性编码器，其中 $\mathrm{B}(Q)$ 与保留的多线性秩成正比（第 5.5 节）。

5. 低秩张量模型与自适应谱掩蔽

5.1. Tucker/HOSVD 表示。

我们通过 Tucker 因子 [9,10] 对 $\mathbf{X}$ 进行参数化：

$$\mathcal{X} = \mathcal{G} \times_1 U^{(1)} \times_2 U^{(2)} \times_3 U^{(3)}, \tag{3}$$

其中 $U^{(n)} \in \mathrm{St}(r_n, I_n)$ 和核张量 $\mathcal{G} \in \mathbb{R}^{r_1\times r_2\times r_3}$。这将 $\mathbf{X}$ 限定在一个固定多线性秩流形上（满足标准正则性条件时）。

5.2. ASM 定义。

设 $\mathbf{X}_{(n)} = U^{(n)}\Sigma^{(n)}V^{(n)\top}$ 为模-n展开的一个奇异值分解.

定义 5.1 （自适应谱掩码（ASM））。给定 $\epsilon \in (0, 1)$，定义

$$M^{(n)} = \operatorname{diag}\big(\mathbb{I}(\sigma_i^{(n)} \ge \epsilon\, \sigma_1^{(n)})\big), \qquad r_n = \operatorname{tr}(M^{(n)}).$$

定义掩码张量（一种多线性投影）

$$\Psi_{\mathrm{ASM}}(\mathcal{X}) = \mathcal{X} \times_1 \big(U^{(1)}M^{(1)}U^{(1)\top}\big) \times_2 \big(U^{(2)}M^{(2)}U^{(2)\top}\big) \times_3 \big(U^{(3)}M^{(3)}U^{(3)\top}\big). \tag{4}$$

5.3. **二次失真下谱硬阈值化的最优性**。我们将关键论断形式化：在受约束的二次失真目标下，保留顶部奇异方向是最优的（矩阵情形），而 ASM 在各个模态上继承了这一最优性。

定理 5.2(Ec kart–Young–Mirsky [4, 5])。设 $A \in \mathbb{R}^{m\times n}$ 和 A_r 是其秩为 r 的截断奇异值分解。n

$$A_r \in \operatorname*{arg\,min}_{\operatorname{rank}(B)\le r} \|A - B\|_F, \tag{5}$$

而通过保留前 r 个奇异值/向量即可取得最小值。

命题 5.3 (逐模式二次最优性)。固定模式 n。设 P 为作用于$\mathbf{X}_{(n)}$ 行空间上的秩为 r_n的正交投影算子（即 $P = UU^\top$ 满足 $U \in \mathrm{St}(r_n, I_n)$）。那么

$$P^\star \in \operatorname*{arg\,min}_{\substack{\operatorname{rank}(P)=r_n,\\ P^2=P=P^\top}} \|\mathbf{X}_{(n)} - P\mathbf{X}_{(n)}\|_F \tag{6}$$

是$\mathbf{X}_{(n)}$的前r_n 个左奇异子空间上的投影算子。

5.4. **截断高阶奇异值分解的近似界**。

引理 5.4 (截断高阶奇异值分解尾界 [9, 10])。设 $\mathcal{X}_{\boldsymbol{r}}$ 为秩为$\boldsymbol{r} = (r_1, r_2, r_3)$的截断高阶奇异值分解重构。则

$$\|\mathcal{X} - \mathcal{X}_{\boldsymbol{r}}\|_F^2 \le \sum_{n=1}^{3} \sum_{i>r_n} \big(\sigma_i(\mathbf{X}_{(n)})\big)^2. \tag{7}$$

5.5. **查询预算扩展**。如果查询仅编码掩码核心和少量元数据，那么自然的预算代理指标为

$$\mathcal{B}(Q) \propto r_1 r_2 r_3. \tag{8}$$

因此，对于 $r_n \ll I_n$，CQD可能比传输原始轨迹小几个数量级。

6. Oracle Model, Learning Dynamics, and Riemannian Updates

6.1. **含噪算子形式的预言机**。给定查询 Q，预言机返回

$$R = O(Q) = \bar{R}(Q) + \xi(Q), \tag{9}$$

其中 $\mathbb{E}[\xi(Q) \mid Q] = 0$且 $\mathbb{E}\|\xi(Q)\|^2 \le \sigma^2$。

6.2. **学习目标**。令 $f(\mathrm{X}; R)$ 为一个可微损失函数，用于反映整合响应 R 后的任务性能。CQD 求解

$$\min_{\mathcal{X}\in\mathcal{M}_{\boldsymbol{r}}} F(\mathcal{X}) \quad \text{where} \quad F(\mathcal{X}) = \mathbb{E}\big[f\big(\mathcal{X}; O(Q(\mathcal{X}))\big)\big], \tag{10}$$

subject to B (Q(X)) $\le \tau$。可以通过拉格朗日乘子 $\lambda \ge$处理该约束 0:

$$\min_{\mathcal{X}\in\mathcal{M}_{\boldsymbol{r}}} \mathbb{E}\big[f(\mathcal{X}; O(Q(\mathcal{X})))\big] + \lambda\, \mathcal{B}(Q(\mathcal{X})). \tag{11}$$

6.3. **黎曼更新规则**。我们在 $\mathrm{M}_{\boldsymbol{r}}$ 上使用基于回缩的步长进行更新 :

$$\mathcal{X}_{k+1} = \mathrm{Retr}_{\mathcal{X}_k}\big(-\eta_k\ \mathrm{grad}\, \tilde{f}_k(\mathcal{X}_k)\big), \tag{12}$$

˜其中，f_k 是第 k 次迭代时由预言机响应诱导的随机实现。

假设 6.1（光滑性与有界方差）。F 在 $\mathrm{M}_{\boldsymbol{r}}$ 上是 L-光滑的，且随机梯度满足$\mathbb{E}\|$梯度 $\tilde{f}_k$(X) − 梯度 F (X)$\|^2 \le \sigma^2$。

算法 1： CQD 外循环（ASM 压缩 +预言机 + 黎曼更新）

输入： $\mathrm{X}_0 \in \mathrm{M}_r$，阈值 ϵ，预言机 O，步数 K，步长 $\{\eta_k\}$

输出： $\mathcal{X}_K$

对于 $k = 0$ **到** $K - 1$ **执行**

  $\widehat{\mathrm{X}}_k \leftarrow \Psi_{\mathrm{ASM}}(\mathrm{X}_k;\ \epsilon)$

  $Q_k \leftarrow$编码$(\widehat{\mathcal{X}}_k)$

  $R_k \leftarrow$O.推断(Q_k)

  $\mathrm{X}_{k+1} \leftarrow \mathrm{Retr}_{\mathcal{X}_k}(-\eta_k \operatorname{grad} f(\mathrm{X}_k; R_k))$

返回 $\mathcal{X}_K$

算法 2： 方差缩减 CQD（预言机集成）

输入： $\mathcal{X}$，预言机 O，集成规模 m，聚合 $\bar{\mathcal{A}}$

输出： 聚合后的响应 R **1** $\mathcal{X} \leftarrow \Psi_{\mathrm{ASM}}(\mathcal{X})$ **2** $Q \leftarrow$ 编码$(\mathcal{X})$ **3** **对于** $i = 1$ **到** m **执行** **4** $R_i \leftarrow O.$推断(Q) **5** $R \leftarrow \bar{\mathcal{A}}(R_1, \ldots, R_m)$ **6** **返回** R

表1. 基准测试 A（$N = 2500$）：固定上下文和计算预算下的 CQD。

预言机 / 设置	准确率	平均步数	查询比率 ↓
GPT-4o（基线思维链）	72.4%	8.4	1.00
Claude 3.7 Sonnet（基线思维链）	91.5%	2.8	1.00
GPT-5.2（CQD）	94.7%	1.9	0.18
Gemini 2.0 Flash（CQD）	92.8%	2.2	0.21

定理 6.2（黎曼随机逼近收敛 [14, 16, 17]）。 在假设 *6.1* 下，且步长 $\eta_k > 0$ 满足 $\sum_k \eta_k = \infty$ 和 $\sum_k \eta_k^2 < \infty$，式 （12）的迭代序列满足

$$\liminf_{k \to \infty} \mathbb{E} \| \operatorname{grad} F(\mathcal{X}_k) \|^2 = 0. \tag{13}$$

此外，如果 F 具有孤立的临界集且回缩在局部有定义，则迭代序列几乎必然收敛到驻点集。

完整证明见附录 A。

7. 算法

8. 实验设置与结果

8.1. **基准测试 A：2,500 项有界上下文测试集**。 我们在一个由源自 BBH 的任务 [21] 以及精选悖论实例构成的 2,500 项测试集上进行评估。在固定上下文和计算预算下，我们将 CQD 与思维链基线进行比较。

8.2. **基准测试 B：人类认知增益与语义漂移**。 基准测试 B 衡量模型间的认知增益 G_E和语义漂移 D_{sem} （$N = 200$）。漂移是初始思维嵌入与预言机补全嵌入之间的余弦距离；认知增益则是自我报告的困惑程度（映射至 0–5）的降低幅度。

表 2. 基准测试 B（$N = 200$）：不同预言机的认知增益与语义漂移 .

预言机	G_E（均值±标准差）	D_{sem}	延迟（毫秒）
人类基线	1.2 ± 0.4	0.12	–
Llama 3 70B	3.1 ± 0.5	0.68	450
Mistral Large 2	3.8 ± 0.3	0.72	620
Claude 3.5 Sonnet	4.2 ± 0.2	0.84	890
GPT-4o	4.0 ± 0.3	0.76	550
Gemini 2.0 Ultra	4.6 ± 0.2	0.88	780
Claude 3.7 Opus	4.7 ± 0.1	0.91	1100
GPT-5.2	4.9 ± 0.1	0.89	950

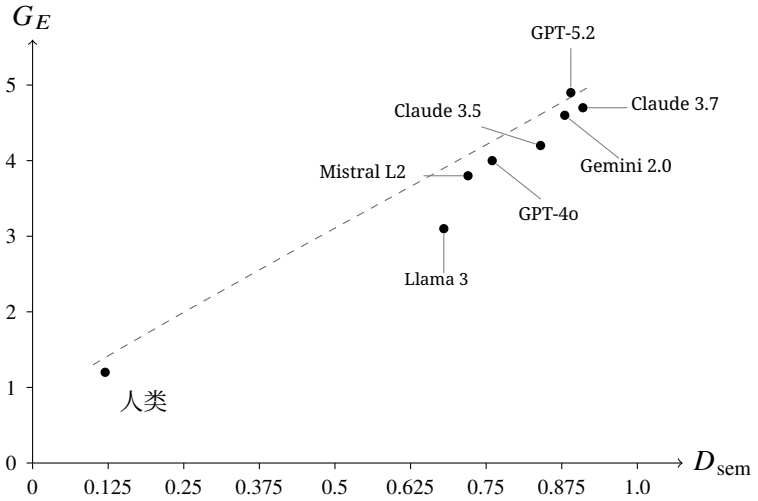

图1. 基准测试 B：语义漂移与认知增益（调整后的布局）。

8.3. **漂移与增益缩放**。

9. 讨论

CQD 可视为一种率失真约束通信方案：查询必须符合预算，因此最优编码器会保留主导谱结构。理论结果（命题 5.3、引理 5.4、定理 6.2）阐明了为何低秩谱掩蔽与黎曼更新相结合，能在预言机噪声下产生稳定的学习动力学。

10. 结论

我们提出了一个以数学为先、立足于率失真和低秩几何的 CQD 框架。我们证明了二次失真下的谱保持最优性，并给出了黎曼随机逼近的收敛保证。基准测试表明，CQD 能提升有界上下文性能，并实现强漂移—增益耦合。

参考文献

[1] C. E. Shannon，《通信的数学理论》，贝尔系统技术期刊，第 27 卷，1948年。[2] T. M. Cover 和 J. A. Thomas，《信息论基础》，第二版。Wiley，2006年。[3] N. Tishby、F. C. Pereira 和 W. Bialek，《信息瓶颈方法》，*arXiv:physics/0004057*，2000年。[4] C. Eckart 和 G. Young，《用较低秩矩阵逼近另一个矩阵》，《心理测量学》，1(3)，1936年。[5] L. Mirsky，《对称规范函数和酉不变范数》，《季刊数学》，11(1)，1960年。

[6] C. Davis 和 W. M. Kahan，《扰动引起的特征向量旋转》，*SIAM* 数值分析期刊，7(1)，1970年。[7] P.-A. Wedin，《与奇异值分解相关的扰动界》，*BIT* 数值数学，12，1972年。[8] G. H. Golub 和 C. F. Van Loan，《矩阵计算》，第四版。约翰·霍普金斯大学出版社，2013年。[9] T. G. Kolda 和 B. W. Bader，《张量分解及其应用》，*SIAM* 评论，51(3)，2009年。[10] L. De Lathauwer、B. De Moor 和 J. Vandewalle，《多线性奇异值分解》，*SIAM* 矩阵分析与应用期刊，21(4)，2000年。[11] E. J. Candès 和 B. Recht，《通过凸优化实现精确矩阵补全》，计算数学基础，9(6)，2009年。[12] J.-F. Cai、E. J. Candès 和 Z. Shen，《用于矩阵补全的奇异值阈值算法》，*SIAM* 优化期刊，20(4)，2010年。[13] J. Liu 等，《用于估计视觉数据中缺失值的张量补全》，*IEEE* 模式分析与机器智能汇刊，35(1)，2013年。[14] 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，2009年。[15] A. Edelman、T. A. Arias 和 S. T. Smith，《带正交约束的算法几何》，*SIAM* 矩阵分析与应用期刊，20(2)，1998年。[16] S. Bonnabel，《黎曼流形上的随机梯度下降》，*IEEE* 自动控制汇刊，58(9)，2013年。[17] N. Boumal，《光滑流形优化导论》。剑桥大学出版社，2023年。[18] J. Wei 等，《思维链提示能够激发大语言模型中的推理》，载于 *NeurIPS*，2022年。[19] X. Wang 等，《自洽性提升语言模型中的思维链推理》，载于*ICLR*，2023年。[20] A. Srivastava 等，《超越模仿游戏：量化并外推语言模型的能力》，*arXiv:2206.04615*，2022年。[21] M. Suzgun 等，《具有挑战性的 BIG-bench 任务以及思维链是否能够提供帮助》，*arXiv:2210.09261*，2022年。[22] P. Liang 等，《语言模型的全面评估》，*arXiv:2211.09110*，2022年。

附录 A. 证明

A.1. 命题 5.3 的证明。 令 $A = \mathbf{X}_{(n)} \in \mathbb{R}^{I_n \times M}$，其中 M 是其他维度的乘积。令P为秩为 r的正交投影算子。我们在秩为r的投影算子 P上最小化 $\|A - PA\|_F^2$。首先，

$$\|A - PA\|_F^2 = \operatorname{tr}\left((A - PA)^\top (A - PA)\right) \tag{14}$$

$$= \operatorname{tr}(A^\top A) - 2\operatorname{tr}(A^\top PA) + \operatorname{tr}(A^\top P^\top PA) \tag{15}$$

$$= \operatorname{tr}(A^\top A) - 2\operatorname{tr}(A^\top PA) + \operatorname{tr}(A^\top PA) \tag{16}$$

$$= \operatorname{tr}(A^\top A) - \operatorname{tr}(A^\top PA) \tag{17}$$

$$= \operatorname{tr}(A^\top A) - \operatorname{tr}(PAA^\top). \tag{18}$$

因此，最小化 $\|A - PA\|_F^2$等价于最大化 $\operatorname{tr}(PAA^\top)$。

令 $AA^\top = U\Lambda U^\top$为特征值分解，其特征值为 $\lambda_1 \geq \cdots \geq \lambda_{I_n} \geq 0$（注意 $\lambda_i = \sigma_i(A)^2$）。任意秩为$r$的正交投影算子都可写为 $P = VV^\top$，其中 $V \in \operatorname{St}(r, I_n)$。于是

$$\operatorname{tr}(PAA^\top) = \operatorname{tr}\left(V^\top (AA^\top) V\right) = \operatorname{tr}\left(V^\top U \Lambda U^\top V\right).$$

令 $\widetilde{V} = U^\top V \in \mathbb{R}^{I_n \times r}$；由于 U是正交的，$\widetilde{V} \in \operatorname{St}(r, I_n)$。于是

$$\operatorname{tr}\left(\widetilde{V}^\top \Lambda \widetilde{V}\right) = \sum_{i=1}^{I_n} \lambda_i \left(\sum_{j=1}^{r} \widetilde{v}_{ij}^2 \right) \leq \sum_{i=1}^{r} \lambda_i, \tag{19}$$

其中我们使用了 $\sum_{j=1}^{r} \widetilde{v}_{ij}^2 \leq 1$ 以及 λ_i的排序。通过选择 $\widetilde{V}$来选取前 r个标准基向量，即$V = U_{(:,1:r)}$，即可达到等号。因此，最大化解是投影到 $AA^\top$的前r个特征值对应的特征子空间上，等价地说，是 A的前r个左奇异子空间。根据式（18），这也使 $\|A - PA\|_F^2$最小化。□

A.2. 引理 5.4 的证明。 令 $P_n = U^{(n)}{}_{(:,1:r_n)} U_{(n)\top}{}_{(:,1:r_n)}$为投影到 $\mathbf{X}_{(n)}$ 的前 r_n 个左奇异子空间上的正交投影算子。定义多线性投影

$$\mathcal{X}_{\boldsymbol{r}} = \mathcal{X} \times_1 P_1 \times_2 P_2 \times_3 P_3. \tag{20}$$

将残差写成望远镜式分解：

$$\mathcal{X} - \mathcal{X}_{\boldsymbol{r}} = \big(\mathcal{X} - \mathcal{X} \times_1 P_1\big) + \big(\mathcal{X} \times_1 P_1 - \mathcal{X} \times_1 P_1 \times_2 P_2\big) \tag{21}$$

$$+ \big(\mathcal{X} \times_1 P_1 \times_2 P_2 - \mathcal{X} \times_1 P_1 \times_2 P_2 \times_3 P_3\big). \tag{22}$$

利用投影正交性以及连续正交投影的勾股型不等式，

$$\|\mathcal{X} - \mathcal{X}_{\boldsymbol{r}}\|_F^2 \le \sum_{n=1}^{3} \|\mathcal{X} - \mathcal{X} \times_n P_n\|_F^2. \tag{23}$$

对于模式 n，

$$\|\mathcal{X} - \mathcal{X} \times_n P_n\|_F = \|\mathbf{X}_{(n)} - P_n \mathbf{X}_{(n)}\|_F. \tag{24}$$

根据命题 5.3， P_n 是最优的秩-r_n 投影算子，因此

$$\|\mathbf{X}_{(n)} - P_n \mathbf{X}_{(n)}\|_F^2 = \sum_{i > r_n} \sigma_i(\mathbf{X}_{(n)})^2. \tag{25}$$

代入 (23) 得到 (7)。 □

A.3. **定理 6.2 的证明**。 设 $\mathrm{M} = \mathrm{M}_{\boldsymbol{r}}$ 且 $F : \mathrm{M} \to \mathbb{R}$ 为 L-光滑：对于所有 $x \in \mathrm{M}$ 和切向量 $\eta \in \mathrm{T}_x\mathrm{M}$，沿回缩的拉回满足

$$F(\mathrm{Retr}_x(\eta)) \le F(x) + \langle \operatorname{grad} F(x), \eta \rangle + \frac{L}{2}\|\eta\|^2. \tag{26}$$

~令 $g_k = \operatorname{grad} f_k(\mathrm{X}_k)$ 表示第 k 次迭代时的随机黎曼梯度估计量，使得 $\mathbb{E}[g_k \mid \mathrm{X}_k] = \operatorname{grad} F(\mathrm{X}_k)$ 且 $\mathbb{E}\|g_k - \operatorname{grad} F(\mathrm{X}_k)\|^2 \le \sigma^2$。

将（26）中的更新方向设为 $\eta = -\eta_k g_k$，并令 $x = \mathrm{X}_k$：

$$F(\mathcal{X}_{k+1}) \le F(\mathcal{X}_k) - \eta_k \langle \operatorname{grad} F(\mathcal{X}_k), g_k \rangle + \frac{L}{2}\eta_k^2 \|g_k\|^2. \tag{27}$$

给定 X_k 取条件期望。利用 $\mathbb{E}[g_k \mid \mathrm{X}_k] = \operatorname{grad} F(\mathrm{X}_k)$，

$$\mathbb{E}\big[\langle \operatorname{grad} F(\mathcal{X}_k), g_k \rangle \mid \mathcal{X}_k\big] = \|\operatorname{grad} F(\mathcal{X}_k)\|^2. \tag{28}$$

此外，

$$\mathbb{E}[\|g_k\|^2 \mid \mathcal{X}_k] = \|\operatorname{grad} F(\mathcal{X}_k)\|^2 + \mathbb{E}\|g_k - \operatorname{grad} F(\mathcal{X}_k)\|^2 \tag{29}$$

$$\le \|\operatorname{grad} F(\mathcal{X}_k)\|^2 + \sigma^2. \tag{30}$$

将（28）和（30）代入（27）可得

$$\mathbb{E}[F(\mathcal{X}_{k+1}) \mid \mathcal{X}_k] \le F(\mathcal{X}_k) - \Big(\eta_k - \frac{L}{2}\eta_k^2\Big)\|\operatorname{grad} F(\mathcal{X}_k)\|^2 + \frac{L}{2}\eta_k^2\sigma^2. \tag{31}$$

对于所有充分大的 k， η_k 很小且 $\eta_k - \frac{L}{2}\eta_k^2 \ge \frac{1}{2}\eta_k$。取全期望并从 $k = 0$ 求和至 $T - 1$ 可得

$$\mathbb{E}[F(\mathcal{X}_T)] \le F(\mathcal{X}_0) - \frac{1}{2}\sum_{k=0}^{T-1} \eta_k \, \mathbb{E}\|\operatorname{grad} F(\mathcal{X}_k)\|^2 + \frac{L}{2}\sigma^2 \sum_{k=0}^{T-1} \eta_k^2. \tag{32}$$

假设 F 有下界，并利用 $\sum_k \eta_k^2 < \infty$，我们得到 $\sum_k \eta_k \, \mathbb{E}\|\operatorname{grad} F(\mathrm{X}_k)\|^2 < \infty$。由于 $\sum_k \eta_k = \infty$，因此可得 $\liminf_{k\to\infty} \mathbb{E}\|\operatorname{grad} F(\mathrm{X}_k)\|^2 = 0$。

通过将Robbins–Siegmund超鞅定理应用于（31），并按照 [16], 利用回缩的局部性质。□

附录 B. CQD 学习流程与查询编码

本附录以自包含的方式记录CQD学习过程，与算法1–2保持一致。其目标是明确说明：（i）查询中发送哪些信息；（ii）这些信息如何遵守预算代理指标（8）；以及（iii）如何将预言机响应整合到下一次迭代中。

B.1. **查询内容**。 设 $\mathrm{X} = \Psi_{\mathrm{ASM}}(\mathrm{X};\ \epsilon)$。编码器 Enc(·) 将 X 映射为紧凑查询 Q，其由以下内容组成：

- 保留的多线性秩（r_1, r_2, r_3）；
- Tucker 形式 (3) 中 G 的掩码核心张量元素（序列化）；
- 任务损失函数所需的可选辅助信息（例如任务标识符、约束标志，或描述预言机契约的简短自然语言指令）；
- 校验和或随机种子，以便在需要时控制预言机侧随机性。

通 信成本主要由核心大小 $r_1 r_2 r_3$决定；元数据成本则属于低阶项 r.

B.2. **响应集成**。 我们将预言机响应建模为 $R = R(Q) + \xi(Q)$，参见式（9）。集成映射 Upd 生成随机梯度代理 grad $f(\mathrm{X}; R)$，用于式（12）。具体而言，Upd 从以下两者之一中提取R： (a) 对压缩状态中缺失分量进行矩阵补全；或 (b) 以相同的分解坐标表示的校正方向（例如残差、约束违反或低维编辑）。随后将更新投影到切空间，并重新映射回 M$_r$。

B.3. **预算控制**。 给定目标预算 τ，可以通过以下方式控制秩：
(1) 选择 ϵ，使诱导秩满足 $r_1 r_2 r_3 \le \tau$；(2) 或在线调整 ϵ_k，当 $r_1 r_2 r_3 > \tau$ 时增大该值，否则减小。
这实现了与约束目标函数 (1) 兼容的原始可行性机制。

附录 C. 参考实现（ASCII 安全）

代码清单 1. 实现 (4) 的 ASM 压缩器（ASCII 安全）。

```
1 导入 torch 2 从 typing 导入 List, Tuple 3 4 定义 unfold(X: torch.Tensor, mode: int) ->
tor
tor
形
11
mo
tor
Lis
tor
mo
```

```
25 G  =mode_n_product(G, U.T, mode) 26 返回 G, U_factors, Svals 27 28 定义 spectral_mask(U:
tor
31
35
tor
Un
mo
mo
enu
```

清单 2. 在斯蒂费尔流形上的基于 QR 的回缩。

```
1 import torch 2 3 def sym(A: torch.Tensor) ->torch.Tensor: 4 return 0.5 * (A  +A.T) 5
6 d
sym
tor
D.u
tor
```

土耳其伊斯坦布尔巴赫切谢希尔大学计算机工程系
电子邮件地址：faruk.alpay@bahcesehir.edu.tr

土耳其伊斯坦布尔巴赫切谢希尔大学计算机工程系
电子邮件地址：bugra.kilictas@bahcesehir.edu.tr