GeoIB：基于统计流形压缩的几何感知信息瓶颈

Weiqi Wang 悉尼科技大学 澳大利亚新南威尔士州悉尼

Zhiyi Tian 悉尼科技大学 澳大利亚新南威尔士州悉尼

Chenhan Zhang 悉尼科技大学 澳大利亚新南威尔士州悉尼

Shui Yu 悉尼科技大学 澳大利亚新南威尔士州悉尼

arXiv:2602.03906v1 [cs.LG] 3 Feb 2026

摘要

信息瓶颈（IB）得到了广泛应用，但在深度学习中，它通常通过易于处理的替代方法实现，例如变分界或神经互信息（MI）估计器，而不是直接控制互信息 $I(X;Z)$ 本身。界的松弛性以及依赖估计器的偏差，可能导致 IB 的“压缩”只能得到间接控制，且优化过程不稳定。

我们从信息几何的视角重新审视信息瓶颈问题，并提出一种无需估计互信息（MI）的**几何信息**瓶颈（**GeoIB**）。我们证明，$I(X;Z)$ 和 $I(Z;Y)$ 可精确表示为从联合分布到各自独立流形的最小 Kullback–Leibler（KL）散度投影形式。基于这一视角，GeoIB 通过两个互补项控制信息压缩：（i）分布层面的费舍尔–拉奥（FR）差异，它与 KL 散度在二阶上相匹配，且具有重参数化不变性；（ii）几何层面的雅可比–弗罗贝尼乌斯（JF）项，它通过惩罚编码器的拉回体积扩张，为 $I_\phi(Z;X)$ 提供局部容量型上界。我们进一步推导出一种与 FR 度量一致的自然梯度优化器，并证明标准的加法式自然梯度步与测地线更新在一阶上等价。我们开展了大量实验，结果表明，在常用数据集上，与主流信息瓶颈基线相比，GeoIB 在信息平面中实现了更好的预测准确率与压缩比之间的权衡。GeoIB 通过在单一瓶颈乘数下统一分布正则化与几何正则化，提升了不变性和优化稳定性。GeoIB 的源代码已发布于 https://anonymous.4open.science/r/G-IB- 0569。

CCS 概念

• **计算方法 → 机器学习**。

关键词

信息瓶颈、流形表示、压缩

1 引言

信息瓶颈（信息瓶颈）原理 [32] 将表征学习刻画为从 X 中提取表示 Z，且该表示仅保留对预测 Y 有用的信息。具体而言，人们寻求一个编码器 $q_\phi(z \mid x)$，使得 Z 尽可能多地携带关于 Y 的信息，同时相对于 X 保持最大程度的压缩，其可表述为：

$$\max_{\phi} I_\phi(Z;Y) \quad \text{s.t.} \quad I_\phi(X;Z) \le R, \tag{1}$$

其中，$I(\cdot;\cdot)$ 表示互信息，ϕ 是编码器的参数，而 R 设定了压缩预算。这里，$I_\phi(\cdot;\cdot)$ 是根据数据分布 $p(x,\ y)$ 和编码器 $q_\phi(z \mid x)$ 计算得出的。为解决这一约束优化问题，信息瓶颈方法（[2, 32]）引入一个正的拉格朗日乘子 β，将问题转化为

$$\min_{\phi} -I_\phi(Z;Y) + \beta I_\phi(X;Z), \tag{2}$$

其中，$\beta \ge 0$（即“瓶颈”参数）用于平衡预测充分性与压缩。

信息瓶颈原理之所以具有吸引力，是因为它通过压缩与预测充分性之间的根本平衡，形式化地界定了何为有用的表示 [2, 32, 35]。信息瓶颈的多种变体 [33, 39–41] 已应用于广泛的领域，包括图像分割 [38],、领域泛化[20], 、语义通信 [34, 36], 和隐私压缩[11, 27]。此外，先前研究 [31] 表明，信息瓶颈为理解深度神经网络的某些训练动态提供了有原则的视角，并揭示了可用于解释视觉Transformer的普遍衰减现象 [14]。

研究空白。 尽管取得了这些成功，大多数实际的信息瓶颈实现仍在优化互信息（MI）的欧几里得替代量，例如 VIB 中的变分界 [2]或 MINE 等神经互信息估计器 [7]。这些替代量忽略了编码器或后验族的统计流形几何，因此无法对 $I(X;\ Z)$ 提供明确的几何保证；即使在互信息较高时，也会削弱学习信号，并降低效用和压缩效果。近期的扩展方法（如结构信息瓶颈 [15, 39], ）尝试从输入中提取结构信息。然而，它们仍依赖欧几里得互信息代理量，在强压缩状态下往往会大幅降低准确率，并使结果对压缩参数 β 高度敏感。这促使我们提出以下研究问题。

研究问题。在高维表示学习中，如何设计一种信息瓶颈，使其在强压缩下仍保持稳定、鲁棒且可控?

我们的工作。本文提出几何信息瓶颈（GeoIB），从统计流形几何的视角重新诠释信息瓶颈。我们首先确立精确的投影刻画：$I(X;Z)$ 和 $I(Z;Y)$ 均可表示为相应联合分布到其独立子流形的最小

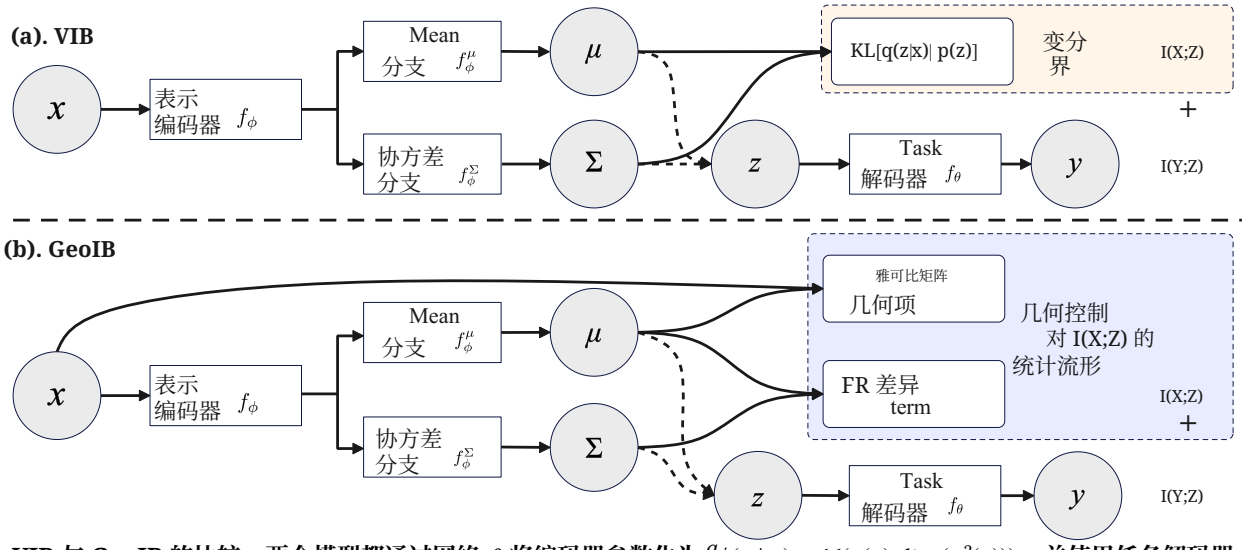


图1：VIB 与 GeoIB 的比较。两个模型都通过网络 f_ϕ将编码器参数化为 $q_\phi(z \mid x) = \mathcal{N}(\mu(x), \mathrm{diag}(\sigma^2(x)))$，并使用任务解码器 $p_\theta(y \mid z)$通过网络 f_θ增大 $I(Z;Y)$。(a) VIB：通过变分上界实现压缩。(b) GeoIB：以两个在统计流形上确定性计算的几何感知惩罚项取代显式互信息估计：费舍尔–拉奥二次代理项 $\mathcal{L}_{\mathrm{FR}}$ 和雅可比-弗罗贝尼乌斯项 $\mathcal{L}_{\mathrm{JF}}$。实线箭头表示确定性映射；虚线箭头表示重参数化采样 $z = \mu + \sigma \odot \varepsilon$。

Kullback–Leibler（KL）散度。随后，我们设计了 GeoIB 方法，通过两个互补组件调节压缩：（i）分布层面的费舍尔–拉奥（FR）差异，该差异在二阶近似下与 KL散度一致，且在对 z进行光滑重参数化时保持不变；（ii）几何层面的雅可比–弗罗贝尼乌斯（JF）惩罚项，它通过抑制编码器的拉回体积扩张，为 $I_\phi(Z;X)$ 提供一个局部容量型上界。这两个组件使压缩在几何上保持一致、易于稳定优化，并忠实体现预期的信息去除概念（分布收缩加容量控制）。最后，我们针对费舍尔–拉奥（FR）度量推导出一个优化器，其更新方向为自然梯度，并证明它与测地线更新在一阶上等价。

总而言之，我们作出以下贡献：

- **信息瓶颈的几何重构**。我们表明，$I(Z;X)$ 和 $I(Z;Y)$ 均可精确表示为从联合分布到各自独立流形的最小 KL散度距离的投影形式，从而阐明了信息瓶颈原理背后的几何结构。

- **一种 GeoIB 解法**。我们提出 GeoIB，通过两个互补项来控制压缩：(i) 分布层面的费舍尔–拉奥（FR）差异，以及 (ii) 几何层面的雅可比矩阵–弗罗贝尼乌斯（JF）惩罚。

- **GeoIB 的一种优化方法**。我们推导出一种与 FR 度量一致的自然梯度优化器，并证明：标准的加性自然梯度步与测地线（指数映射）更新在一阶上等价。

- **实证验证**。我们进行了广泛的实验，与具有代表性的基准方法进行比较。相较于其他方法，GeoIB 在信息平面上实现了良好的准确率–压缩权衡相较于最先进的信息瓶颈基线，在强压缩条件下表现出更强的鲁棒性。

2 相关工作

2.1 信息瓶颈

信息瓶颈（IB）拉格朗日函数 [32]已在表示学习 [1, 28] 和实用训练技术 [21, 37]中得到广泛研究。一种实用的深度实现是变分信息瓶颈（VIB） [2], , 它引入变分编码器$q_\phi(z \mid x)$、先验 $r(z)$和解码器/分类器 $q_\theta(y \mid z)$，并使用以下界：

$$\begin{aligned} I(Z;X) &= \mathbb{E}_{p(x)}\big[\mathrm{KL}\big(q_\phi(z \mid x) \,\|\, q_\phi(z)\big)\big] \\ &\le \mathbb{E}_{p(x)}\big[\mathrm{KL}\big(q_\phi(z \mid x) \,\|\, r(z)\big)\big] \\ &= \mathbb{E}_{p(x)q_\phi(z|x)} \log q_\phi(z \mid x) \;-\; \mathbb{E}_{q_\phi(z)} \log r(z), \end{aligned} \tag{3}$$

$$I(Z;Y) = H(Y) - H(Y \mid Z) \;\ge\; H(Y) + \mathbb{E}_{p(x,y)\,q_\phi(z|x)} \log q_\theta(y \mid z). \tag{4}$$

这里 $q_\phi(z) = \int q_\phi(z \mid x)p(x)\,dx$是编码器边缘分布。式 (3) 使用 $\mathrm{KL}\big(q_\phi(z) \,\|\, r(z)\big) \ge 0$（当 $r(z) = q_\phi(z)$ 时取等号）。式 (4) 源于 $\mathrm{KL}\big(p(y \mid z) \,\|\, q_\theta(y \mid z)\big)$的非负性（交叉熵界）。

除 VIB 外，另一条互补的研究路线是用神经互信息估计器 [8, 12, 17] 取代先验–KL 近似项。这些方法训练判别器，以优化互信息及相关量的变分界，从而能够更直接地优化信息瓶颈目标。代表性方法包括 MINE [7], , 它最大化 Donsker–Varadhan下界；NWJ 及相关的 f 散度界 [19]；以及 CLUB，它提供了一个上界，可用于惩罚 $I(Z;X)$ [9]。这些估计器无需选择先验 $r(z)$，但也带来了实际权衡

（偏差/方差、互信息较大时的饱和，以及对负样本和判别器容量的依赖），McAllester 和 Stratos [25] 对此进行了分析。

尽管效果显著，上述大多数方法都依赖于高维空间中的显式或变分互信息估计，而这类估计可能较为脆弱。我们利用统计流形的几何来实现信息瓶颈。具体而言，在信息几何中，费舍尔–拉奥度量赋予分布一种黎曼结构，在该结构下，KL散度在局部等于测地距离的平方[6]；因此，费舍尔–拉奥（FR）差异 $d^2_{\mathrm{FR}}\big(q_\phi(z \mid x)\ r(z)\big)$ 可为 $I(Z;\ X)$压缩项提供一种重参数化不变替代项。另一方面，将编码器均值映射μ_ϕ: X →Z视为诱导拉回度量 $J_\mu^\top J_\mu$，这表明可以对局部体积失真进行惩罚，从而与收缩正则化及基于雅可比矩阵的正则化 [29]建立联系。我们将在方法论部分详细介绍几何信息瓶颈。

2.2 信息几何优化

信息几何优化（IGO）将学习和优化视为在概率分布的统计流形上移动，该流形配备了以 Fisher 信息为黎曼度量 [4–6, 26]。在这一设定下，自然梯度 [3] 是按照分布族的内在几何度量的最速下降方向，而不是环境欧几里得参数化下的最速下降方向。具体而言，对于参数化分布族 p_θ和目标函数 L(θ)，自然梯度更新形式为

$$\theta \leftarrow \theta - \eta\, F(\theta)^{-1} \nabla_\theta \mathcal{L}(\theta), \tag{5}$$

其中，$F(\theta)$ 是 Fisher信息矩阵。该更新对于 θ 的光滑重参数化具有不变性，并可解释为沿着局部 KL 散度一致的方向前进；与欧几里得梯度 [3, 6] 相比，这通常能带来更好的条件数和稳定性。

主要区别。 我们的方法受到相同信息几何基础的启发，但在目标和几何的使用方式上都不同于 IGO。首先，IGO 主要是一种优化原则：它规定如何更新分布参数，使优化进展与参数化方式无关，并遵循所选分布族的内在几何 [26]。相比之下，我们的方法是一种用于表征学习的信息瓶颈方法：其核心贡献是重新设计压缩机制，使其与几何保持一致。其次，优化对象不同：IGO 通常在黑箱设定 [26], 下优化候选解上的搜索分布，而我们优化的是编码器后验族 $q_\phi(z \mid x)$，并将其与任务预测器结合，明确旨在调节信息瓶颈的权衡。第三，所强调的不变性也相应不同：IGO 所定义的不变性针对分布上的、由目标驱动的黑箱优化 [26], ，而我们利用信息几何，为学习到的表征实现对重参数化稳健的压缩和容量控制。

3 从几何视角出发的问题陈述

令 P 为 $(X,\ Z)$ 上所有联合分布构成的统计流形，并令独立流形为 I $_{XZ}$ ={ $q(x)r(z)$ }。在指数（e-）坐标中，I$_{XZ}$是一个 e-平坦子流形 [5, 6]。对于任意 q

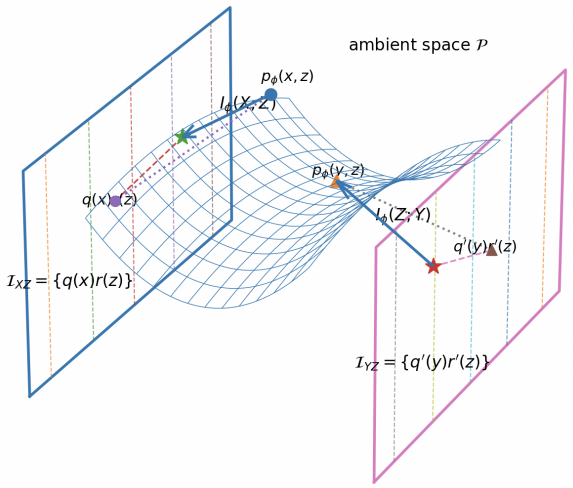


图2：信息瓶颈目标函数的信息几何视角。环境统计流形 P 包含联合分布 $p_\phi(x,z)$ 和 $p_\phi(y,z)$。蓝色和平面表示 e-平坦独立流形 $\mathbf{I}_{XZ}$和 $\mathbf{I}_{YZ}$。箭头表示投影到这些乘积流形上的 KL I-投影，其最小化解为 $p(x)p_\phi(z)$ 和$p(y)p_\phi(z)$；根据信息几何毕达哥拉斯恒等式，相应的 KL 距离等于 $I_\phi(Z;\ X)$ 和 $I_\phi(Z;\ Y)$。因此，最小化 LIB(ϕ) $= \beta I_\phi(Z;\ X) - I_\phi(Z;\ Y)$ 会推动$p_\phi(x,z)$ 趋向 $\mathbf{I}_{XZ}$，同时拉动 $p_\phi(y,z)$ 远离 $\mathbf{I}_{YZ}$。

并且 r，我们有信息几何毕达哥拉斯恒等式 [6]

$$\begin{aligned}\mathrm{KL}\big(p_\phi(x,z) \,\|\, q(x)r(z)\big) = &\underbrace{\mathrm{KL}\big(p_\phi(x,z) \,\|\, p(x)p_\phi(z)\big)}_{=I_\phi(X;Z)} \\ &+ \underbrace{\mathrm{KL}\big(p(x)p_\phi(z) \,\|\, q(x)r(z)\big)}_{\geq 0},\end{aligned} \tag{6}$$

由此 $I_\phi(Z;\ X)$ =min$_{q,r}$KL散度($p_\phi(x,z)$ ‖ $q(x)r(z)$)，其最小化元为 (q,r) =($p(x)\,p_\phi(z)$)。对 (Y,Z) 成立相同的关系，只需将 x替换为 y，将 q替换为 $q_\prime$，并将 r替换为 $r_\prime$：

$$\mathrm{KL}\big(p_\phi(y,z) \,\|\, q'(y)r'(z)\big) = I_\phi(Z;Y) + \mathrm{KL}\big(p(y)p_\phi(z) \,\|\, q'(y)r'(z)\big). \tag{7}$$

我们假设绝对连续性，从而所有 KL 项均为有限，且 Fubini 定理 [16] 适用；特别地，对于候选乘积测度，有 $p_\phi(x,z)$ ≪ $q(x)r(z)$ 以及 $p_\phi(y,z)$ ≪ $q_\prime(y)r_\prime(z)$。我们在附录 A 中给出式 (6) 的详细证明，式 (7) 也可以用同样的方式证明。因此，信息瓶颈损失 [32], L信息瓶颈(ϕ) $= \beta I_\phi(Z;\ X) - I_\phi(Z;\ Y)$ 可以精确地写为

$$\mathcal{L}_{\mathrm{IB}}(\phi) = \beta \min_{q,r} \mathrm{KL}\big(p_\phi(x,z) \,\|\, q(x)r(z)\big) - \min_{q',r'} \mathrm{KL}\big(p_\phi(y,z) \,\|\, q'(y)r'(z)\big), \tag{8}$$

其中，内部最小值分别在 (q,r) =($p(x)\,p_\phi(z)$) 和 $(q_\prime r_\prime)$ =($p(y)$ $p_\phi(z)$) 处取得。因此，对 ϕ 优化式 (8) 等价于：

- *push* $p_\phi(x,z)$ *toward the independence manifold* $\mathcal{I}_{XZ}$ by minimizing $\beta\,\mathrm{KL}\big(p_\phi(x,z) \,\|\, p(x)p_\phi(z)\big)$;

• 通过最大化 $\mathrm{KL}(p_\phi(y,z)\,\|\,p(y)p_\phi(z))$，使 (y,z) 偏离其独立流形 $\mathcal{I}_{YZ}=\{q_r(y)r_r(z)\}$。

我们在图2中展示了一个几何信息瓶颈目标的示例。这种几何化重述为信息瓶颈权衡提供了具体的推拉式解释，并通过将互信息表示为到乘积分布流形的KL投影，将内层极小化解的几何结构与对ϕ进行的外层优化区分开来，从而阐明了优化的对象以及最优解所在的位置。此外，由于独立流形是e平坦的，该表述能够直接运用KL投影和毕达哥拉斯分解等标准信息几何工具，从而简化分析并为构建有理论依据的近似方法提供启发。

4 几何信息瓶颈方法 thod

在第3节中，我们通过信息几何重新表述信息瓶颈目标函数：互信息 $I_\phi(X;Z)$ 和 $I_\phi(Z;Y)$ 分别是联合分布 $p_\phi(x,z)$ 和 $p_\phi(y,z)$ 到各自独立流形的最小KL散度，因此信息瓶颈拉格朗日函数可转化为投影距离之差。基于此，我们提出几何信息瓶颈（GeoIB），通过两个互补项控制压缩：(i) 分布层面 的费舍尔–拉奥（FR）差异；以及 (ii) 几何层面 的雅可比–弗罗贝尼乌斯（JF）项。在训练过程中，我们提出一种自然梯度下降方法，将几何信息结合起来，以获得更好的优化效果。

4.1 通过费舍尔–拉奥（FR）二次型构造分布代理

我们用局部二阶 Fisher–Rao 度量近似定义压缩项的条件分布–边缘分布散度：

$$I_\phi(Z;X)=\mathbb{E}_{p(x)}D_{\mathrm{KL}}\big(q_\phi(z|x)\,\|\,p_\phi(z)\big)\approx\tfrac{1}{2}\,\mathbb{E}_{p(x)}d_{\mathrm{FR}}\big(q_\phi(z|x),r(z)\big)^2, \tag{9}$$

其中 $r(z)$ 是参考边缘分布（例如标准正态分布或学习得到的先验），$d_{\mathrm{FR}}(\cdot\,,\cdot)$ 表示分布之间的费舍尔–拉奥测地距离，即由费舍尔信息度量诱导的黎曼距离。对于光滑参数族 $\{p_\theta\}$，局部等价关系表明，当$\theta_r=\theta+\Delta$ 满足 $\|\Delta\|$ 较小时：

$$D_{\mathrm{KL}}\big(p_{\theta'}\,\|\,p_\theta\big)=\tfrac{1}{2}\,\Delta^\top F(\theta)\,\Delta+o(\|\Delta\|^2)=\tfrac{1}{2}\,d_{\mathrm{FR}}\big(p_{\theta'},p_\theta\big)^2+o(\|\Delta\|^2), \tag{10}$$

其中 $F(\theta)$ 是 Fisher 信息。将式（10）中的 $p_{\theta'}$ 替换为 $q_\phi(z|x)$，并将 p_θ 替换为 $r(z)$，即可得到式（9）。证明见附录B。**对角高斯示例**。若$q_\phi(z|x)=\mathrm{N}(\mu_\phi(x),\ \mathrm{diag}(\sigma_\phi^2(x)))$且 $r(z)=\mathrm{N}(0,\ I)$，则精确的 KL 散度为

$$D_{\mathrm{KL}}\big(q_\phi(z|x)\,\|\,\mathcal{N}(0,I)\big)=\tfrac{1}{2}\sum_{j=1}^{d_z}\Big(\mu_j(x)^2+\sigma_j(x)^2-\log\sigma_j(x)^2-1\Big). \tag{11}$$

将式 (11) 在 $\mu=0,\ \sigma^2=1$ 处展开，得到

$$D_{\mathrm{KL}}\big(q_\phi\|\mathcal{N}(0,I)\big)=\tfrac{1}{2}\|\mu\|_2^2+\tfrac{1}{4}\|\log\sigma^2\|_2^2+O(\|\log\sigma^2\|_2^3), \tag{12}$$

这等于$\frac{1}{2}\,d_{\mathrm{FR}}\big(q_\phi,\ \mathrm{N}(0,\ I)\big)^2$ 直至二阶。在实践中，当 $r=\mathrm{N}(0,\ I)$时，我们可以优化式(11)中的闭式KL散度：

$$\text{LKL散度}(\phi)=\widehat{\mathbb{E}}_x\,D_{\text{KL散度}}\big(q_\phi(z|x)\,\|\,r(z)\big).$$

如果 r是学习得到的/复杂的（例如，VampPrior/流），我们可以优化FR代理

$$\mathcal{L}_{\mathrm{FR}}(\phi)=\tfrac{1}{2}\,\widehat{\mathbb{E}}_x\,d_{\mathrm{FR}}\big(q_\phi(z|x),r(z)\big)^2, \tag{13}$$

还可以选择监测 $\widehat{I}_{\mathrm{KL}_{XZ}}=\mathbb{E}_x\,D_{\mathrm{KL}}\big(q_\phi(z|x)\,\|\,\widehat{p}_\phi(z)\big)$，以衡量其紧致程度。

4.2 通过雅可比-弗罗贝尼乌斯项得到的几何界

设重参数化编码器为 $z=f_\phi(x)+\varepsilon,$，其中 $\varepsilon\sim\mathrm{N}(0,\Sigma(x)),$，并将雅可比矩阵记为 $J_f(x)=\partial f_\phi(x)/\partial x$。输入流形上的拉回度量为 $g_x=J_f(x)^\top\Sigma(x)^{-1}J_f(x)$。

局部容量型上界。 在 f_ϕ 附近进行线性化，并令 C_x 为局部输入协方差，高斯信道上界给出

$$I_\phi(Z;X)\le\tfrac{1}{2}\,\mathbb{E}_{p(x)}\Big[\log\det\Big(I+\Sigma(x)^{-\frac{1}{2}}J_f(x)\,C_x\,J_f(x)^\top\Sigma(x)^{-\frac{1}{2}}\Big)\Big]. \tag{14}$$

在单位局部能量约束$C_x\preceq I$（勒夫纳序）以及半正定锥 $\mathbb{S}_+^d$上 log det 单调性的条件下，我们得到逐点界

$$\begin{aligned}I_\phi(Z;X)&\le\tfrac{1}{2}\,\mathbb{E}_{p(x)}\Big[\log\det\Big(I+\Sigma(x)^{-\frac{1}{2}}J_f(x)J_f(x)^\top\Sigma(x)^{-\frac{1}{2}}\Big)\Big]\\&=\tfrac{1}{2}\,\mathbb{E}_{p(x)}\Big[\log\det\Big(I+J_f(x)^\top\Sigma(x)^{-1}J_f(x)\Big)\Big],\\&\le\tfrac{1}{2}\,\mathbb{E}_{p(x)}\ \mathrm{Tr}\big(\Sigma(x)^{-1}J_f(x)J_f(x)^\top\big)\\&=\tfrac{1}{2}\,\mathbb{E}_{p(x)}\big\|\Sigma(x)^{-\frac{1}{2}}J_f(x)\big\|_F^2\\&=:\tfrac{1}{2}\,\mathrm{JF}(\phi).\end{aligned} \tag{15}$$

证明概要。 由于 Z仅通过 $(f_\phi(X)\ \Sigma(X))$ 依赖于 X，因此有马尔可夫链 $X\to(f_\phi(X)\ \Sigma(X))\to Z$，从而 $I(X;Z)=I((f_\phi(X)\ \Sigma(X));Z)$。在给定 x 的条件下，将 f_ϕ 在 x 处线性化，同时在局部固定 $\Sigma(x)$，可得到增益为 $J_f(x)$、噪声协方差为 $\Sigma(x)$ 的（局部）线性高斯信道。在对输入协方差 $C_x\preceq I$ 施加单位局部能量约束时，对于给定协方差，高斯分布的熵最大，因此得到式 (14) 中的对数行列式界。利用 $\det(I+AB)=\det(I+BA)$，并对 $A\succeq 0$ 应用 $\log\det(I+A)\le\mathrm{Tr}(A)$，即可得到式 (15)。□

无偏Hutchinson估计器。 式 (15) 中的迹无需显式构造雅可比矩阵即可估计。对于任意 $v\sim\mathrm{N}(0,I_{d_x})$，

$$\mathbb{E}_v\big[\|\Sigma(x)^{-1/2}J_f(x)v\|_2^2\big]=\mathrm{Tr}\big(\Sigma(x)^{-1}J_f(x)J_f(x)^\top\big). \tag{16}$$

使用 S独立同分布的探测向量 $\{v_s\}_{s=1}^S$，定义逐样本估计器

$$\widehat{\mathrm{JF}}(x):=\frac{1}{S}\sum_{s=1}^{S}\big\|\Sigma(x)^{-1/2}\,J_f(x)\,v_s\big\|_2^2, \tag{17}$$

使得

$$\mathbb{E}_v[\widehat{\mathrm{JF}}(x)]=\mathrm{Tr}\big(\Sigma^{-1}J_f(x)J_f(x)^\top\big). \tag{18}$$

训练目标为批次平均值

$$\mathcal{L}_{\mathrm{JF}}(\phi):=\widehat{\mathbb{E}}_{x\sim\mathrm{batch}}\big[\widehat{\mathrm{JF}}(x)\big]. \tag{19}$$

在自动微分框架中，通过雅可比矩阵–向量积（JVP）计算 $J_f(x)v_s$，每个 x 需要进行 $O(S)$ 次前向模式调用；通常 $S=1$ 取 2 就足够了。

各向同性/对角噪声。若 $\Sigma(x) = \sigma(x)^2 I$，则

$$I_\phi(Z;X) \;\le\; \tfrac{1}{2}\,\mathbb{E}_{p(x)} \log\det\!\Big(I + \tfrac{1}{\sigma(x)^2} J_f J_f^\top\Big) \;\le\; \tfrac{1}{2}\,\widehat{\mathbb{E}}_x\, \frac{\|J_f(x)\|_F^2}{\sigma(x)^2}. \tag{20}$$

设置标量下限 $\sigma^2_{\text{最小值}} > 0$ 后，可得到一个简单且稳定的替代方案。

几何意义。由于 $g_x = J_f^\top \Sigma^{-1} J_f$，我们有 $\mathrm{Tr}(g_x) = \|\Sigma(x)^{-\frac{1}{2}} J_f(x)\|_F^2$，即方向平均局部伸缩（$\Sigma^{-1}$ 度量下的狄利克雷能量密度）。因此，最小化雅可比–弗罗贝尼乌斯项可控制平均测地线长度失真，为 $I_\phi(Z;X)$ 提供一种有理论依据的压缩代理目标。

4.3 GeoIB 的自然梯度优化

基于上述内容，我们将 GeoIB 目标函数表述为

$$\begin{aligned}\mathcal{L}_{\text{GeoIB}}(\phi,\theta) \;=\; & \underbrace{\mathbb{E}_{p(x,y)}\mathbb{E}_{q_\phi(z|x)}\big[-\log p_\theta(y\mid z)\big]}_{\uparrow\, I_\phi(Z;Y)} \\ & + \underbrace{\beta\Big(\widehat{\mathcal{L}}_{\text{FR}}(\phi) + \widehat{\mathcal{L}}_{\text{JF}}(\phi)\Big)}_{\downarrow\, I_\phi(Z;X)},\end{aligned} \tag{21}$$

其中，$\beta \ge 0$ 是瓶颈乘数。最小化 $\mathbb{E}_{p(x,y)}\mathbb{E}_{q_\phi(z|x)}\big[-\log\, p_\theta(y\mid z)\big]$ 会降低 $H(Y|Z)$ 并提高 $I_\phi(Z;Y)$（因为 $H(Y)$ 是固定的）；FR 项和雅可比–弗罗贝尼乌斯项共同惩罚 $I_\phi(Z;X)$。

编码器上的自然梯度。将 $\{q_\phi(z \mid x)\}_\phi$ 视为赋予 Fisher–Rao 度量的统计流形，任意标量目标 $\mathrm{J}(\phi)$ 的自然梯度为

$$\begin{aligned}\widetilde{\nabla}_\phi \mathcal{J} \;&=\; F_\phi^{-1}\,\nabla_\phi \mathcal{J}, \\ F_\phi \;&:=\; \mathbb{E}_{p(x)}\mathbb{E}_{q_\phi(z|x)}\big[\,\nabla_\phi \log q_\phi(z\mid x)\,\nabla_\phi \log q_\phi(z\mid x)^\top\,\big],\end{aligned} \tag{22}$$

其中，公式 (22) 中的矩阵 F_ϕ 是 Fisher–Rao 度量张量。我们通过对完整目标的欧几里得梯度进行预条件化来更新编码器：

$$\begin{aligned}\phi_{t+1} \;=\; \phi_t - \eta_\phi\, F_{\phi_t}^{-1}\Big(&\nabla_\phi\, \mathbb{E}_{x,y,z}[-\log p_\theta(y\mid z)] \\ &+\beta\big[\nabla_\phi \mathcal{L}_{\text{FR}} + \nabla_\phi \mathcal{L}_{\text{JF}}\big]\Big),\end{aligned} \tag{23}$$

该更新在 ϕ 的光滑重参数化下具有一阶不变性，并通过单个预条件器 F_ϕ 耦合分布层面的 FR 信号与几何层面的 JF 信号。我们还可以证明，加性的自然梯度步 $\phi^+ = \phi - \eta F_\phi^{-1}\nabla_\phi \mathrm{J}$ 是几何中测地线更新的一阶近似。

命题1（自然梯度等于黎曼梯度）。设 $\mathcal{M} = \{p_\phi : \phi \in \Theta \subset \mathbb{R}^d\}$ 是赋予 *Fisher–Rao* 度量 g_ϕ（u, v）：$= u^\top F_\phi v$的正则统计流形，其中 $F_\phi = \mathbb{E}_{p}(x)\ \mathbb{E}_{q_\phi}(z|x)\ [\nabla_\phi \log\, q_\phi\, \nabla_\phi \log\, q_\phi^\top]$。对于标量目标 $\mathcal{J}$：$\mathcal{M} \to \mathbb{R}$，其在 ϕ 处的黎曼梯度满足

$$\mathrm{grad}\,\mathcal{J}(\phi) \;=\; F_\phi^{-1}\,\nabla_\phi \mathcal{J}, \tag{24}$$

也就是说，自然梯度 $\widetilde{\nabla}_\phi \mathcal{J} := F_\phi^{-1}\nabla_\phi \mathcal{J}$ 正是$(\mathcal{M}, g)$上的黎曼梯度。

证明概要。根据定义，黎曼梯度 $\mathrm{grad}\,\mathrm{J}(\phi) \in T_\phi\mathcal{M} \simeq \mathbb{R}^d$ 是满足以下条件的唯一向量场：对于所有切方向 $v \in T_\phi\mathcal{M}$，

$$\langle \mathrm{grad}\,\mathcal{J}(\phi),\, v\rangle_{g_\phi} \;=\; \mathrm{D}\mathcal{J}_\phi[v].$$

在 Fisher–Rao 度量下，内积为 $u, v_{g_\phi} = u^\top F_\phi v$。在坐标表示中，微分等于与通常梯度的欧几里得配对：$\mathrm{dJ}_\phi[v] = v^\top \nabla_\phi \mathrm{J}$。因此，对于所有 v，

$$v^\top F_\phi\, \mathrm{grad}\,\mathcal{J}(\phi) \;=\; v^\top \nabla_\phi \mathcal{J}.$$

由于 F_ϕ 在正则点处正定，我们得出 $F_\phi \mathrm{grad}\,\mathrm{J}(\phi) = \nabla_\phi \mathrm{J}$，即 $\mathrm{grad}\,\mathrm{J}(\phi) = F_\phi^{-1}\nabla_\phi \mathrm{J}$。

命题 2（Fisher–Rao 度量下的最速下降）。设$(\mathcal{M}, u\ v = u^\top F_\phi v$)配备 *Fisher–Rao* 度量(,)，且$\mathcal{J}$为光滑函数。单位 *FR*长度下的最速下降方向满足

$$\min_{\|v\|_{g_\phi} \le 1} \mathrm{D}\mathcal{J}(\phi)[v], \tag{25}$$

且（单位范数）优化器为 $v_\star = -$梯度 $\mathcal{J}(\phi)$/梯度 $\mathcal{J}(\phi)\ \|_{g_\phi}$，约定当 $v_\star = 0$ *if* 梯度 $\mathcal{J}(\phi) = 0$。特别地，根据命题*1*，其方向与负自然梯度一致：$-F_\phi^{-1}\nabla_\phi$

梯度 $\mathcal{J}(\phi)\mathcal{J}$。

证明概要。根据

[6], $\phi\ [v]\,\mathrm{DJ}(\,)\,\mathrm{grad}\,\mathrm{J}(\phi)$ $= v g_\phi$ 中的黎曼梯度定义。柯西-施瓦茨不等式给出 $\mathrm{grad}\,\mathrm{J}\ v g_\phi \ge -\|\mathrm{grad}\,\mathrm{J}\ \|_{g_\phi}\ \|v\|_{g_\phi} \ge -\|\mathrm{grad}\,\mathrm{J}\ \|_{g_\phi}$，等号成立当且仅当 v 与 $-\mathrm{grad}\,\mathrm{J}$ 共线且 $\|v\|_{g_\phi} = 1$。□

根据命题2，每单位 FR长度的最速下降方向为 $-$梯度 $\mathrm{J}(\phi)$。因此，我们沿该方向使用指数映射进行更新。

定理 1（通过指数映射进行测地线更新）。令$\mathrm{Exp}_\phi\colon T_\phi \mathcal{M} \to \mathcal{M}$为 *FR* 度量的黎曼指数映射。离散更新

$$\phi^+ \;=\; \mathrm{Exp}_\phi\big(-\eta\, \mathrm{grad}\,\mathcal{J}(\phi)\big)$$

位于唯一的 *FR*测地线 γ上，该测地线以 ϕ为起点，初速度为 $\dot\gamma(0) = -\eta$梯度 $\mathcal{J}(\phi)$；即，$\phi^+ = \gamma(1)$。

证明见附录 C。

推论 1（与加性更新的一阶等价性）。设$\mathcal{R}_\phi$是$\mathcal{M}$上的任意收缩映射，满足$\mathcal{R}_\phi(0) = \phi$且$\mathrm{D}\mathcal{R}_\phi(0) = \mathrm{Id}$（指数映射是规范的收缩映射）。在局部坐标中，

$$\mathrm{Exp}_\phi\big(-\eta\, F_\phi^{-1}\nabla_\phi \mathcal{J}\big) \;=\; \phi - \eta\, F_\phi^{-1}\nabla_\phi \mathcal{J} \;+\; O(\eta^2).$$

根据命题*1*，梯度 $\mathcal{J} = F_\phi^{-1}\nabla_\phi \mathcal{J}$。因此，常用的加性自然梯度步 $\phi^+ = \phi - \eta F_\phi^{-1}\nabla_\phi \mathcal{J}$是测地线更新的一阶近似。

解码器上的自然梯度。对于解码器，我们对 $\{p_\theta(y\mid z)\}_\theta$ 使用自然梯度：

$$\theta_{t+1} \;=\; \theta_t - \eta_\theta\, F_\theta^{-1}\,\nabla_\theta \mathbb{E}_{x,y,z}[-\log p_\theta(y\mid z)], \tag{26}$$

其中 $F_\theta \;=\; \mathbb{E}_{q_\phi(z)}\mathbb{E}_{p_\theta(y|z)}\big[\nabla_\theta \log\, p_\theta(y\mid z)\nabla_\theta \log\, p_\theta(y\mid z)^\top\big]$。实践中，我们使用可扩展的近似方法和求解器（如 K-FAC [23, 24]）来计算 $F_\theta^{-1} g$。受篇幅限制，我们在附录 D 中给出了完整的 GeoIB 算法。

表1：在图像数据集 MNIST、CIFAR10 和 CelebA 上的总体评估结果。粗体表示最佳结果；斜体表示次佳结果。

方法	MNIST, $\beta = 0.0001$, $K = 128$				CIFAR10, $\beta = 0.0001$, $K = 128$				CelebA, $\beta = 0.0001$, $K = 128$			
	准确率	$I(X;Z)$	MSE	RT	准确率	$I(X;Z)$	MSE	RT	准确率	$I(X;Z)$	MSE	RT
VIB [2]	98.72%	1.81	0.034	**199.84**	*82.65%*	0.87	*0.887*	**277.27**	95.85%	0.55	0.054	**264.18**
SIB [39]	*99.08%*	1.86	0.037	399.31	75.81%	**0.63**	0.051	524.49	**97.25%**	**0.38**	**0.072**	456.92
MINE [7]	98.85%	1.76	0.032	*234.04*	82.64%	0.88	0.806	*362.07*	96.29%	0.51	0.048	*314.26*
AIB [41]	99.01%	*1.69*	*0.043*	293.56	82.63%	*0.84*	0.812	509.56	96.05%	0.52	0.053	454.80
GeoIB（我们的方法）	**99.28%**	**1.69**	**0.043**	442.32	**85.54%**	1.01	**0.899**	556.08	*97.01%*	*0.47*	*0.066*	464.91

5 实验

本节通过实验回答有关 GeoIB 的以下研究问题（RQ）：

- **RQ1**：与最先进的信息瓶颈解决方案相比，所提出的 GeoIB 在信息压缩和预测准确率方面表现如何？（见第 5.2 节和第 5.3 节）
- **RQ2**：不同的超参数（如拉格朗日乘子 β 和表示维度 K）会如何影响 GeoIB？（见第 5.4 节和第 5.5 节）

5.1 实验设置

数据集。 我们在三个广泛采用的公开数据集上开展了实验：MNIST [10],CIFAR10 [18],和 CelebA [22], ，这些数据集涵盖了多种目标类别，学习复杂性各不相同。我们在附录 E 中详细介绍了所有数据集的统计信息及其使用方式。

模型。 我们在实验中选取了三种不同规模的模型架构：一个 7 层卷积神经网络（CNN）、一个 5 层多层感知机（MLP）和 ResNet18。对于 MNIST 数据集，我们使用两个 MLP 构成 GeoIB 模型（一个作为压缩编码器，另一个作为任务解码器）。对于 CIFAR10 和 CelebA，我们使用 ResNet18 作为编码器，并分别使用 MLP 和 CNN 作为 CIFAR10 和 CelebA 的解码器。

评估指标。 遵循具有代表性的信息瓶颈研究 [2, 39, 41], ，我们通过模型在留出测试集上的 top-1 分类**准确率** 来量化模型效用。我们使用 MINE [7]估算学习得到的表示与输入之间的互信息$I(Z;X)$，以评估信息压缩。为探究 Z中包含的信息泄漏，我们实施两种标准的表示级攻击：(i) 模型反演攻击，根据 z [13]重建 x，并使用均方误差（**均方误差**）进行评估；均方误差越低，表示信息泄漏越严重；(ii) 成员推断攻击 [30], ，通过成员推断准确率（**成员推断准确率**）进行评估，数值越高表示信息泄漏越严重。我们还通过记录方法的运行时间（**运行时间**）来评估其效率。

对比信息瓶颈基准方法。我们将 GeoIB 与四种具有代表性的信息瓶颈（IB）变体进行比较：（1）标准的变分信息瓶颈（**VIB**）[2]；（2）使用**MINE** [7]估计互信息项的信息瓶颈变体；（3）专注于结构感知特征学习的先进结构化信息瓶颈（**SIB**）[39]；以及（4）将对抗正则化融入瓶颈的对抗信息瓶颈（**AIB**） [41]。为确保公平，所有方法均采用相同的骨干网络、数据预处理和训练计划；在适用时，按照原论文的方法在验证集上调整超参数。

5.2 GeoIB整体评估

实验设置。我们在 MNIST、CIFAR10 和 CelebA 上，使用相同的骨干网络和训练计划，将 GeoIB 与四种具有代表性的信息瓶颈变体进行比较。为进行公平比较，我们固定拉格朗日乘子 $\beta = 10^{-4}$ 和表示维度 $K = 128$。表1报告了 top-1 准确率（越高越好）、通过 MINE 估计的互信息 I（X； Z）（I（X； Z）值越低表示压缩能力越强），以及通过从 z 重建 x 所得的模型反演均方误差（MSE）（值越高表示信息泄漏越少）。

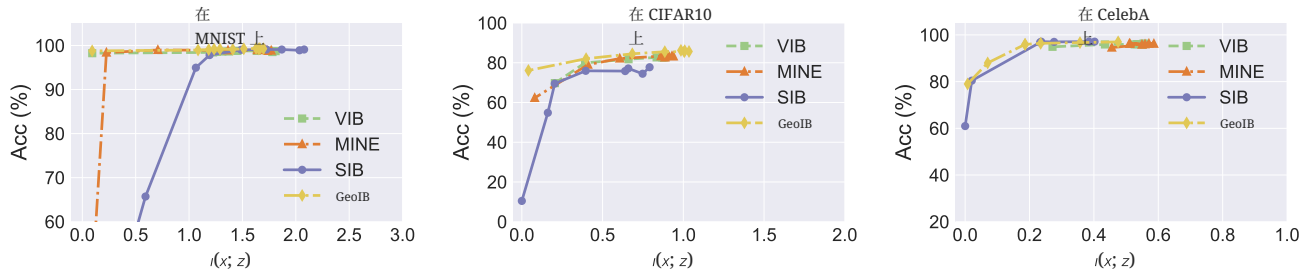
结果。 在这三个数据集上，GeoIB在所有评估指标中均取得最佳或次佳结果。在MNIST上，GeoIB的准确率最高，且 I（X； Z）并列最低，同时反演均方误差也位居榜首。这表明它能够实现强效压缩，同时不牺牲效用。在CIFAR10上，GeoIB的准确率最佳，反演均方误差最高（信息泄漏最少）。尽管SIB的I（X； Z）最低，GeoIB整体上提供了更好的准确率与隐私权衡。在CelebA上，SIB的准确率和互信息略胜一筹。GeoIB排名第二，在准确率和隐私保护方面均表现出色（由均方误差体现），证实了其在更细粒度、结构化数据集上的稳健性。

从效率角度来看，VIB在所有数据集上均展现出最佳效率表现，而采用MINE的IB排名第二。它们在互信息估计时忽略几何信息，因此训练效率较高。GeoIB的运行时间始终更长，但与考虑结构信息的SIB运行时间相近。

5.3 信息平面评估：压缩与准确率

实验设置。 我们通过从 10^{-6} 到 10_1 随机选择拉格朗日乘子 β，追踪每种方法在 MNIST、CIFAR10 和 CelebA 上的信息平面。在本实验中，我们固定表示大小 K=128。对于每个 β，我们训练至收敛，并记录测试准确率（越高越好）以及通过 MINE 估计的互信息 $I(X;Z)$，如图3所示。

结果。 在图3中，跨不同数据集来看，GeoIB 的曲线相较于基线方法始终向左上方偏移：在相同的 $I(X;Z)$ 下准确率更高，或在准确率相同的情况下 $I(X;Z)$ 更小（帕累托改进）。压缩较弱时，即 $I(X;Z)$ 较大时，所有方法的准确率都达到较高的平台。随着压缩增强，即 $I(X;Z)$ 变小，VIB、MINE 和 SIB 的准确率明显下降，而 GeoIB 在更大的低-$I(X;Z)$ 范围内仍能保持准确率，之后才开始下降。在 CIFAR10 上，这一差距最为明显：GeoIB 在较低的 $I(X;Z)$ 下仍能保持准确率。在 CelebA 上，各条曲线聚集在顶部附近，但 GeoIB 只需更少的 Z 信息，就能达到相当的准确率。



Fi图3：从信息平面对压缩比和预测准确率进行评估。

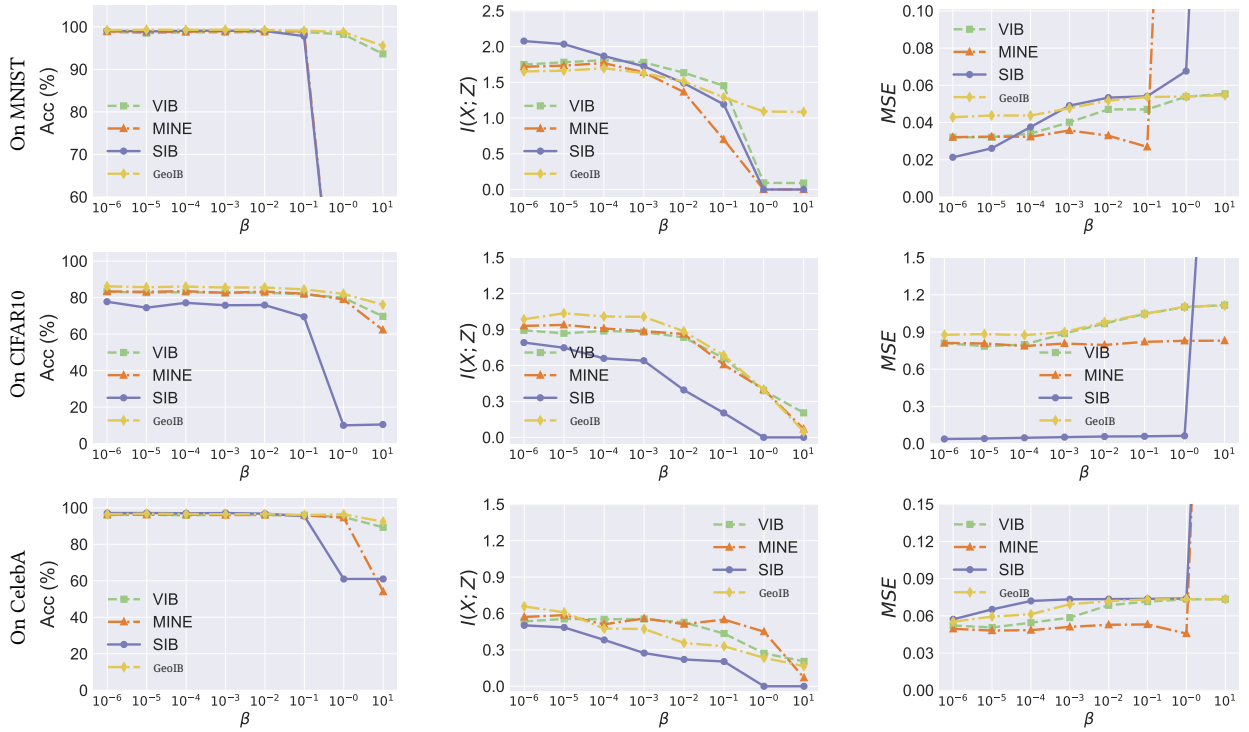


图4：评估瓶颈乘数 β 的影响。

5.4 消融实验：瓶颈乘数 β 的影响

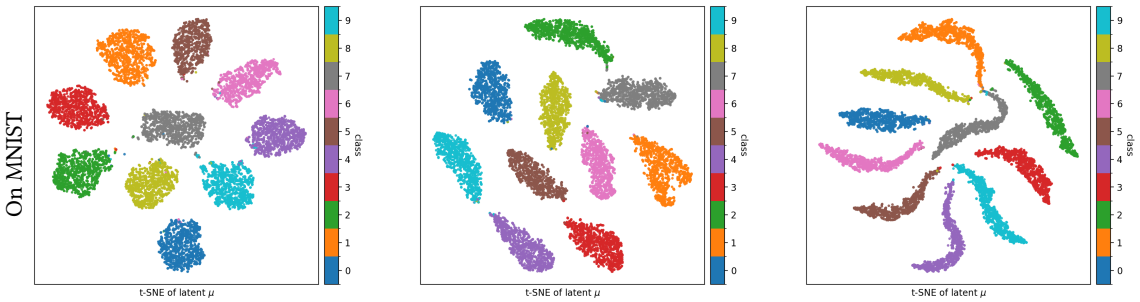
设置。 我们通过在对数网格上从 10^{-6} 到 10_1 扫描 β，研究瓶颈乘数的影响，并将表示大小固定为 $K = 128$。对于每个 β，我们训练一个模型，并报告测试准确率（越高越好）、估计的互信息 $I(X; Z)$（越低表示压缩越强），以及通过从 z 重建 x 得到的模型反演均方误差（越高表示信息泄漏越少）。MNIST、CIFAR10 和 CelebA 的结果见图4。

结果。 （左列）准确率与 β。在所有数据集上，GeoIB 在较宽的 β 范围内始终保持最高或接近最高的准确率。基线方法，尤其是 MINE 和 SIB，在 $\beta \geq 10^{-1}$ 时准确率会急剧下降；在 CIFAR10 和 CelebA 上，SIB 在 β 较大时会崩溃。

（中）$I(X; Z)$ 与 β。随着 β 增加，所有方法的 $I(X; Z)$ 均呈现预期的单调下降。SIB实现了最小的 $I(X; Z)$（压缩最强），但代价是准确率显著下降；而GeoIB在保持准确率的同时，实现了具有竞争力的压缩效果。

（右列）均方误差与 β。在大多数 β 上，GeoIB 的均方误差表现最佳或次佳，表明其抵御反演攻击的能力更强。在极端 β（10^0–10^1）下，部分基线方法的均方误差会因表征崩溃而飙升。重建结果实际上变得随机，这会推高均方误差，但同时也意味着效用较差。

定性可视化。为了考察瓶颈强度如何塑造表示，我们在三个不同的 β 取值下，对 10,000 张 MNIST 测试图像的表示 $\mu(x) = \mathbb{E}[Z \mid x]$ 进行二维 t-SNE 可视化（图5）；CIFAR10 的结果见附录 F 中的图7。当瓶颈乘数较小时（$\beta = 10^{-4}$），各簇分离良好，且类内分布相对较大，这表明 Z 保留了细粒度输入细节。当 $\beta = 10^0$ 时，表示簇向类别原型收缩，同时



(a) $\beta = 10^{-4}$，准确率 = 99.28%　　**(b)** $\beta = 10^{0}$，准确率 = 98.77%　　**(c)** $\beta = 10^{1}$，准确率 = 95.53%

图5：在MNIST（$K = 128$）上将10,000张测试图像的后验均值 $\mu(x)$ 的表征嵌入可视化为二维。颜色表示真实标签。从左到右：$\beta = 10^{-4}$、100和101；相应的测试准确率显示在各面板下方。随着 β 增大，类内离散度缩小，簇向类别原型靠拢，表明压缩更强；准确率随之下降。

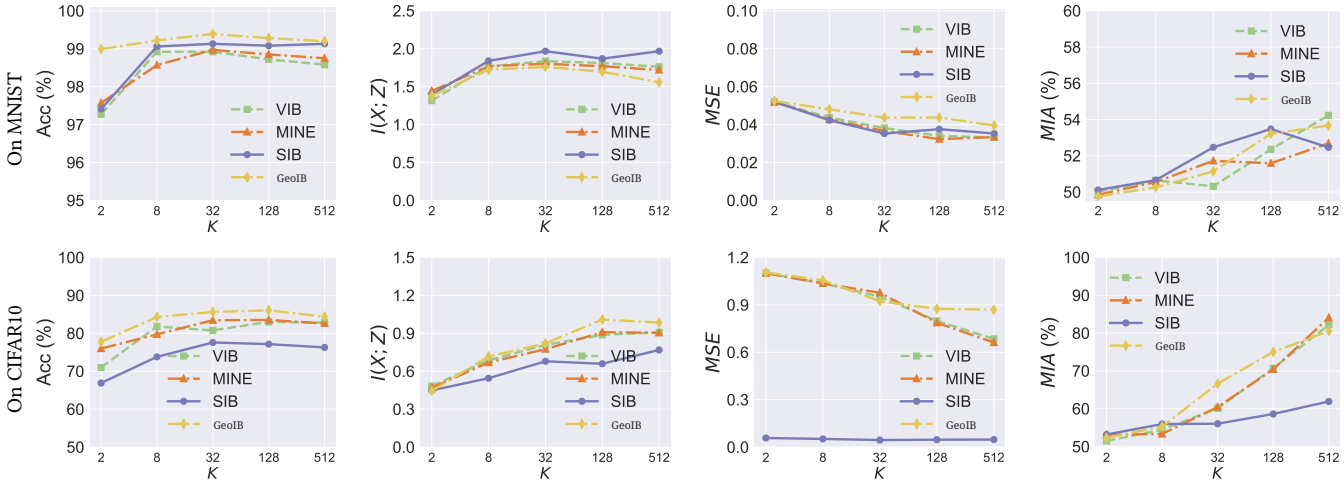


图6：评估表示维度的影响 K。

仍然可分。在 $\beta = 10^{1}$ 处，表征嵌入集中在类中心附近的狭窄弧线上，这与更强的压缩以及图4中观察到的准确率下降相一致。

5.5 消融实验：表示维度的影响 K

实验设置。我们将表示维度 K 从 2_1 调整至 2_9，并将拉格朗日乘子固定为 $\beta = 10^{-4}$。对于每个 K，我们报告不同信息瓶颈方法的准确率、互信息 $I(X;Z)$、均方误差和成员推断准确率。MNIST 和 CIFAR10 的结果见图6。

结果。 在图 6 的左列中，准确率随着 K提高，随后趋于饱和（MNIST：增益在 $K \geq 32$左右达到平台；CIFAR10：在 $K \geq 128$左右）。在 K范围内，GeoIB 的准确率最高或接近最高，尤其是在低至中等范围内。在中左列中，正如预期，$I(X;Z)$ 会随着 K增加而增大，这表明潜空间越宽，压缩越弱。MNIST 在 K 非常大时出现的尾部非单调性影响很小，这可能是因为对于 MNIST 数据集而言，$K = 32$ 的表示维度已经足够大。中右列显示，随着 K 增大，均方误差通常会降低（重建变得更容易），这反映了高维 Z带来的信息泄漏增加。GeoIB 在较小和中等的 K下仍能保持有竞争力的较高均方误差（即更能抵抗反演）。右列中，在两个数据集上，成员推断准确率都会随着 K 提高而上升，这印证了更大的 Z 携带更多成员信号。在广泛的 K范围内，GeoIB 始终具有竞争力；当 K 较小至中等时，它能够实现良好的效用-隐私平衡。

6 总结与未来工作

我们提出 GeoIB，从信息几何的角度解决信息瓶颈问题。其核心要素包括：(i) 分布层面的费舍尔–拉奥（FR）差异度，它在局部二阶匹配 KL 散度，并且在潜变量的光滑重参数化下保持不变；(ii) 几何层面的雅可比–弗罗贝尼乌斯（JF）惩罚项，它通过抑制拉回体积扩张，为 $I_\phi(Z;X)$ 提供局部容量型上界。

值得探索的方向包括：将基于 FR 的代理指标拓展到局部（二阶）范围之外，并以更精确的谱近似取代迹松弛；提高 GeoIB 训练的效率；以及将 GeoIB 拓展到联邦和隐私保护设置，在这些场景中，FR 和 JF 控制或可带来可验证的机器遗忘保证。

致谢

作者谨向匿名评审致谢，感谢他们提出的宝贵意见和评审建议。

参考文献

[1] Alessandro Achille 和 Stefano Soatto。2018。信息丢弃：通过噪声计算学习最优表示。IEEE模式分析与机器智能汇刊 40，12（2018），2897–2905。[2] Alexander A Alemi、Ian Fischer、Joshua V Dillon 和 Kevin Murphy。2016。深度变分信息瓶颈。arXiv 预印本 arXiv:1612.00410（2016）。[3] Shun-Ichi Amari。1998。自然梯度在学习中高效发挥作用。神经计算 10，2（1998），251–276。[4] S-I Amari。2002。概率分布层次结构上的信息几何。IEEE信息论汇刊 47，5（2002），1701–1711。[5] Shun-ichi Amari。2016。信息几何及其应用。第194卷。施普林格。[6] Shun-ichi Amari 和 Hiroshi Nagaoka。2000。信息几何方法。第191卷。美国数学学会。[7] Mohamed Ishmael Belghazi、Aristide Baratin、Sai Rajeshwar、Sherjil Ozair、Yoshua Bengio、Aaron Courville 和 Devon Hjelm。2018。互信息神经估计。载于：国际机器学习会议。PMLR，531–540。[8] Ivan Butakov、Aleksandr Tolmachev、Sofia Malanchuk、Anna Neopryatnaya 和 Alexey Frolov。2024。通过归一化流进行互信息估计。神经信息处理系统进展 37（2024），3027–3057。[9] Pengyu Cheng、Weituo Hao、Shuyang Dai、Jiachang Liu、Zhe Gan 和 Lawrence Carin。2020。CLUB：互信息的对比对数比上界。载于：国际机器学习会议。PMLR，1779–1788。[10] Li Deng。2012。用于机器学习研究的手写数字图像MNIST数据库 [网络精选]。IEEE信号处理杂志 29，6（2012），141–142。[11] Yann Dubois、Benjamin Bloem-Reddy、Karen Ullrich 和 Chris J Maddison。2021。用于无损预测的有损压缩。神经信息处理系统进展 34（2021），14014–14028。[12] Giulio Franzese、Mustapha Bounoua 和 Pietro Michiardi。2023。MINDE：互信息神经扩散估计。arXiv 预印本 arXiv:2310.09031（2023）。[13] Matt Fredrikson、Somesh Jha 和 Thomas Ristenpart。2015。利用置信度信息的模型反演攻击及基本对策。载于：第22届ACM SIGSAC计算机与通信安全会议论文集。1322–1333。[14] Jung-Ho Hong、Ho-Joong Kim、Kyu-Sung Jeon 和 Seong-Whan Lee。2025。用于揭示通用归因以解释视觉Transformer的综合信息瓶颈。载于：计算机视觉与模式识别会议论文集。25166–25175。[15] Dou Hu、Lingwei Wei、Yaxin Liu、Wei Zhou 和 Songlin Hu。2024。结构化概率编码。载于：AAAI人工智能会议论文集，第38卷。12491–12501。[16] Olav Kallenberg。1997。现代概率论基础。施普林格。[17] Artemy Kolchinsky、Brendan D Tracey 和 David H Wolpert。2019。非线性信息瓶颈。熵 21，12（2019），1181。[18] Alex Krizhevsky、Geoffrey Hinton 等。2009。从微型图像中学习多层特征。（2009）。[19] Nunzio Alexandro Letizia、Nicola Novello 和 Andrea M Tonello。2024。通过 f-散度和数据错排进行互信息估计。神经信息处理系统进展 37（2024），105114–105150。[20] Bo Li、Yifei Shen、Yezhen Wang、Wenzhen Zhu、Dongsheng Li、Kurt Keutzer 和 Han Zhao。2022。用于领域泛化的不变信息瓶颈。载于：AAAI人工智能会议论文集，第36卷。7399–7407。[21] Jin Li、Yaoming Wang、Xiaopeng Zhang、Dongsheng Jiang、Wenrui Dai、Chenglin Li、Hongkai Xiong 和 Qi Tian。2025。通过变分信息瓶颈进行对比学习。IEEE模式分析与机器智能汇刊（2025）。[22] Ziwei Liu、Ping Luo、Xiaogang Wang 和 Xiaoou Tang。2018。大规模CelebFaces属性（CelebA）数据集。检索于2018年8月15日（2018），11。[23] James Martens、Jimmy Ba 和 Matt Johnson。2018。循环神经网络的Kronecker因子化曲率近似。载于：国际学习表征会议。[24] James Martens 和 Roger Grosse。2015。使用Kronecker因子化近似曲率优化神经网络。载于：国际机器学习会议。PMLR，2408–2417。[25] David McAllester 和 Karl Stratos。2020。互信息测量的形式限制。载于：国际人工智能与统计会议。PMLR，875–884。[26] Yann Ollivier、Ludovic Arnold、Anne Auger 和 Nikolaus Hansen。2017。信息几何优化算法：通过不变性原理形成的统一图景。机器学习研究杂志 18，18（2017），1–65。[27] Behrooz Razeghi、Flavio P Calmon、Deniz Gunduz 和 Slava Voloshynovskiy。2023。Bottlenecks CLUB：统一复杂度、信息泄漏与效用之间的信息论权衡。*IEEE* 信息取证与安全汇刊 18 (2023), 2060–2075。[28] Domenic Rosati、Jan Wehner、Kai Williams、Lukasz Bartoszcze、Robie Gonzales、Subhabrata Majumdar、Hassan Sajjad、Frank Rudzicz 等。2024。表示加噪：一种抵御有害微调的防御机制。神经信息处理系统进展 37 (2024), 12636–12676。[29] Andrew Ross 和 Finale Doshi-Velez。2018。通过对输入梯度进行正则化，提高深度神经网络的对抗鲁棒性和可解释性。载于*AAAI*人工智能会议论文集，第 32 卷。[30] Reza Shokri、Marco Stronati、Congzheng Song 和 Vitaly Shmatikov。2017。针对机器学习模型的成员推理攻击。载于*2017 年 IEEE 安全与隐私研讨会（SP）*。IEEE，3–18。[31] Ravid Shwartz-Ziv 和 Naftali Tishby。2017。通过信息揭开深度神经网络的黑箱。*arXiv* 预印本 *arXiv:1703.00810* (2017)。[32] Naftali Tishby、Fernando C Pereira 和 William Bialek。2000。信息瓶颈方法。*arXiv* 预印本 *physics/0004057* (2000)。[33] Zhibin Wan、Changqing Zhang、Pengfei Zhu 和 Qinghua Hu。2021。多视图信息瓶颈表征学习。载于*AAAI*人工智能会议论文集，第 35 卷。10085–10092。[34] Weiqi Wang、Zhiyi Tian、Chenhan Zhang 和 Shui Yu。2024。Scu：一种用于深度学习赋能语义通信的高效机器遗忘方案。*IEEE* 信息取证与安全汇刊 (2024)。[35] Tailin Wu、Ian Fischer、Isaac L Chuang 和 Max Tegmark。2020。信息瓶颈的可学习性。载于人工智能中的不确定性。PMLR，1050–1060。[36] Songjie Xie、Shuai Ma、Ming Ding、Yuanming Shi、Mingjian Tang 和 Youlong Wu。2023。采用数字调制的面向任务通信的鲁棒信息瓶颈。*IEEE* 通信领域精选期刊 41, 8 (2023), 2577–2591。[37] Mengting Xu、Tao Zhang、Zhongnian Li 和 Daoqiang Zhang。2022。InfoAT：利用信息瓶颈原理改进对抗训练。*IEEE* 神经网络与学习系统汇刊 35, 1 (2022), 1255–1264。[38] Zhengze Xu、Dongyue Wu、Changqian Yu、Xiangxiang Chu、Nong Sang 和 Changxin Gao。2024。Sctnet：融合 Transformer 语义信息的单分支卷积神经网络，用于实时分割。载于*AAAI*人工智能会议论文集，第 38 卷。6378–6386。[39] Hanzhe Yang、Youlong Wu、Dingzhu Wen、Yong Zhou 和 Yuanming Shi。2025。结构化信息瓶颈：通过结构化特征学习改进信息瓶颈。载于*AAAI*人工智能会议论文集，第 39 卷。21922–21928。[40] Shujian Yu、Xi Yu、Sigurd Løkse、Robert Jenssen 和 Jose C Principe。2024。用于回归的柯西-施瓦茨散度信息瓶颈。载于第十二届国际学习表征会议。[41] Penglong Zhai 和 Shihua Zhang。2022。对抗信息瓶颈。*IEEE* 神经网络与学习系统汇刊 35, 1 (2022), 221–230。

A 方程 6 的证明

证明 . 利用对数因子分解恒等式和 Fubini–Tonelli 定理

$$\begin{aligned}\mathrm{KL}\big(p(x,z)\,\|\,q(x)r(z)\big) &= \iint p(x,z)\,\log\frac{p(x,z)}{q(x)r(z)}\,dx\,dz\\ &= \iint p(x,z)\,\log\frac{p(x,z)}{p_X(x)p_Z(z)}\,dx\,dz\\ &\quad+\iint p(x,z)\,\log\frac{p_X(x)}{q(x)}\,dx\,dz+\iint p(x,z)\,\log\frac{p_Z(z)}{r(z)}\,dx\,dz\\ &= \underbrace{\mathrm{KL}\big(p(x,z)\,\|\,p_X(x)p_Z(z)\big)}_{=I(X;Z)}\\ &\quad+\underbrace{\int p_X(x)\,\log\frac{p_X(x)}{q(x)}\,dx}_{=\mathrm{KL}(p_X\|q)}+\underbrace{\int p_Z(z)\,\log\frac{p_Z(z)}{r(z)}\,dz}_{=\mathrm{KL}(p_Z\|r)}.\end{aligned}$$

最后一步分别利用 Fubini 定理对 z 和 x 积分。由于每个 KL 项均非负，因此关于 q,r 的最小值在 $q=p_X,\ r=p_Z$ 处取得，最小值为 $I(X;Z)$。□

备注 1（信息几何毕达哥拉斯关系）。由于 $\mathcal{I}$ 在适当的对偶仿射坐标中是平坦的，上述分解也是信息几何的“毕达哥拉斯定理”:

$$\mathrm{KL}\big(p\,\|\,qr\big)=\mathrm{KL}\big(p\,\|\,p_Xp_Z\big)+\mathrm{KL}\big(p_Xp_Z\,\|\,qr\big),\qquad qr\in\mathcal{I}.$$

因此， p_Xp_Z 是 p 在 $\mathcal{I}$ 上的 e-投影，而 $I(X;Z)$ 恰好是投影距离。

B 局部二阶（FR）的证明。

我们证明，对于正则参数族 $\{p_\theta:\ \theta\in\Theta\subset\mathbb{R}^d\}$，且在 θ 附近有 θ'，即 $\Delta=\theta'-\theta$,

$$\mathrm{KL}\big(p_\theta\,\|\,p_{\theta'}\big)=\tfrac12\,\Delta^\top F(\theta)\,\Delta+o(\|\Delta\|^2)=\tfrac12\,d_{\mathrm{FR}}\big(p_\theta,p_{\theta'}\big)^2+o(\|\Delta\|^2),$$

其中 $F(\theta)$ 是 Fisher 信息， dFR 是 Fisher–Rao 距离。全文假设满足标准正则性条件：共同支撑集、关于 θ 的二阶连续可微性、 $F(\theta)$ 有限，以及期望与微分可交换。

记

$$\mathrm{KL}\big(p_\theta\,\|\,p_{\theta'}\big)=\mathbb{E}_{p_\theta}\big[\log p_\theta(Z)-\log p_{\theta'}(Z)\big].$$

固定 z并在 θ 处对 $\log p_{\theta'}(z)$ 展开:

$$\log p_{\theta'}(z)=\log p_\theta(z)+\Delta^\top\nabla_\theta\log p_\theta(z)+\tfrac12\,\Delta^\top\nabla^2_\theta\log p_\theta(z)\,\Delta+o(\|\Delta\|^2).$$

从 $\log\ p_\theta(z)$ 中减去，取 $\mathbb{E}_{p_\theta}$，并利用 $\mathbb{E}_{p_\theta}\big[\nabla_\theta\log\ p_\theta(Z)\big]=0$（得分函数均值为零），得到

$$\mathrm{KL}\big(p_\theta\,\|\,p_{\theta'}\big)=-\tfrac12\,\Delta^\top\mathbb{E}_{p_\theta}\big[\nabla^2_\theta\log p_\theta(Z)\big]\,\Delta\ +\ o(\|\Delta\|^2).$$

根据信息恒等式,

$$F(\theta)\ =\ \mathbb{E}_{p_\theta}\big[\nabla_\theta\log p_\theta(Z)\,\nabla_\theta\log p_\theta(Z)^\top\big]\ =\ -\mathbb{E}_{p_\theta}\big[\nabla^2_\theta\log p_\theta(Z)\big],$$

hence

$$\mathrm{KL}\big(p_\theta\,\|\,p_{\theta'}\big)=\tfrac12\,\Delta^\top F(\theta)\,\Delta\ +\ o(\|\Delta\|^2).\qquad g_\theta\ u\ s\ (2\underline{7})$$

FR度量是在 Θ 上由 $u^\top F\ \theta\ v$ $u\ v\in\mathbb{R}^d$

(,) ( ) 给出的黎曼度量，适用于切向量 , 。令 $\gamma\colon[0,1]\to\Theta$ 为任意满足 $\gamma(0)=\theta$ 且 $\gamma(1)=\theta'$ 的曲线。其 FR长度为

$$L(\gamma)\ =\ \int_0^1\sqrt{\dot\gamma(t)^\top F(\gamma(t))\,\dot\gamma(t)}\,dt,$$

而 FR距离为 dFR$(p_\theta,\ p_{\theta'})\ =\inf_\gamma L(\gamma)$。对于 $\|\Delta\|\ \to\ 0$，取直线段 $\gamma(t)\ =\theta+t\Delta$，可得

$$L(\gamma)^2=\Big(\int_0^1\sqrt{\Delta^\top F(\theta+t\Delta)\,\Delta}\,dt\Big)^2=\Delta^\top F(\theta)\,\Delta\ +\ o(\|\Delta\|^2),$$

利用 F的连续性以及关于 t的二阶展开。由于测地线长度是最小的,

$$d_{\mathrm{FR}}\big(p_\theta,p_{\theta'}\big)^2=\Delta^\top F(\theta)\,\Delta\ +\ o(\|\Delta\|^2).\tag{28}$$

这可由标准黎曼几何得出：在 θ处的正规坐标中， dFR$(p_\theta,\ p_{\theta'})$ $^2=\Delta^\top F(\theta)\Delta+O(\|\Delta\|^3)$，因此得到方程 28。结合方程 27 和方程 28，可得局部等价

$$\mathrm{KL}\big(p_\theta\,\|\,p_{\theta'}\big)=\tfrac12\,d_{FR}\big(p_\theta,p_{\theta'}\big)^2\ +\ o(\|\Delta\|^2).$$

设$q_\phi(z|x)$和参考边缘分布 $r(z)$属于$\{p_\theta\}$，其参数分别为 $\theta(x)$和 θ_r，并假设对于 $p(x)$–几乎处处的 x， $\|\theta(x)\ -\theta_r\|$很小。将上述逐点结果应用于 $\Delta(x)\ =\theta(x)\ -\theta_r$，得到

$$\mathrm{KL}\big(q_\phi(z|x)\,\|\,r(z)\big)=\tfrac12\,d_{\mathrm{FR}}\big(q_\phi(z|x),r(z)\big)^2\ +\ o\big(\|\Delta(x)\|^2\big).$$

取$\mathbb{E}_{p(x)}$并利用控制收敛定理（由正则性假设保证），可得

$$\begin{aligned}\mathbb{E}_{p(x)}\mathrm{KL}\big(q_\phi(z|x)\,\|\,r(z)\big)\ =&\tfrac12\,\mathbb{E}_{p(x)}d_{\mathrm{FR}}\big(q_\phi(z|x),r(z)\big)^2\\ &+\ o\Big(\mathbb{E}_{p(x)}\|\Delta(x)\|^2\Big).\end{aligned}$$

当 $r(z)=\ p_\phi(z)$（聚合后验）时，这正是压缩项$\mathbb{E}_{p(x)}$KL散度$\big(q_\phi(z|x)\ \|\,p_\phi(z)\big)$的局部二阶近似。□

C 定理1的证明

定理1的证明。设 $v:=-\eta\,\mathrm{grad\ J}\,(\phi)\ \in\ T_\phi\mathrm{M}$。根据 FR度量的 Levi–Civita 联络下测地线方程的存在唯一性定理，存在 $\varepsilon>0$，且存在唯一的测地线 $\gamma_v\colon(-\varepsilon,\ \varepsilon)\ \to\mathrm{M}$，使得 $\gamma_v(0)\ =\phi$，且 $\dot\gamma_v(0)\ =v$。

在 ϕ 处的黎曼指数映射定义为

$$\mathrm{Exp}_\phi(w)\ =\ \gamma_w(1)\quad\text{whenever 1 lies in the domain of }\gamma_w,$$

等价地，对于 0 的某个邻域内的 t，有 $\mathrm{Exp}\phi(tw)\ =\gamma_w(t)$。（为确保全局定义良好，可将范围限制在注入半径以下的 $\|w\|$。）将 $w=v$ 代入可得

$$\phi^+\ =\ \mathrm{Exp}_\phi(v)\ =\ \gamma_v(1),$$

也就是说， ϕ^+位于从 ϕ出发、初速度为 $\dot\gamma(0)\ =v=-\eta$ grad J (ϕ) 的唯一 FR测地线上。□

D GeoIB 算法

GeoIB 算法如下所示，见算法1。

算法1在 K-FAC 曲率模式下，使用自然梯度更新解码器 θ和编码器 ϕ，执行一次 GeoIB 训练迭代。（步骤1）给定一个小批量 $\{(x_i,y_i)\}_{i=1}^B$，通过重参数化后验 $z_i\sim q_{\phi_t}(z\mid x_i)$采样潜在编码。（步骤2）无需构造完整的雅可比矩阵，即可估计几何感知瓶颈代理量$\widehat{\ }$：利用 S Hutchinson 探测向量和雅可比–向量积（JVP）计算费舍尔–拉奥代理量 $\mathrm{L_{FR}}$ 以及$\widehat{\ }$雅可比–Frobenius 惩罚项 $\mathrm{L_{JF}}$。（步骤

算法1： GeoIB自然梯度步（每次迭代）

输入： 当前参数（ϕ_t，θ_t）；小批量大小 B；步长（η_ϕ，η_θ）；瓶颈 β；Hutchinson探针数量 S；阻尼 λ；Fisher近似模式 K-FAC

输出： 更新后的参数（ϕ_{t+1}，θ_{t+1}）

过程 GIB_Step$(\phi_t, \theta_t, B, \eta_\phi, \eta_\theta, \beta, S, \lambda,$ 模式$)$**：**

// 1) 采样小批量与潜在编码

抽取小批量 $\{(x_i, y_i)\}_{i=1}^{B}$;;

$z_i \sim q_{\phi_t}(z \mid x_i)$ 通过重参数化；

// 2) 估计FR/JF项（Hutchinson + 雅可比向量积）

估计 $\widehat{\mathrm{L}}_{\mathrm{FR}}$ 与 $\widehat{\mathrm{L}}_{\mathrm{JF}}$ 使用 S 探测向量和雅可比向量积；

// 3) 计算欧几里得梯度

$g_\theta \leftarrow \nabla_\theta \frac{1}{B}\sum_{i=1}^{B}\left[-\log p_{\theta_t}(y_i \mid z_i)\right]$;

$g_\phi \leftarrow \nabla_\phi \left\{\frac{1}{B}\sum_{i=1}^{B}\left[-\log p_{\theta_t}(y_i \mid z_i)\right] + \beta(\widehat{\mathrm{L}}_{\mathrm{FR}} + \widehat{\mathrm{L}}_{\mathrm{JF}})\right\}$;

// 4) 构建费舍尔近似（经验费舍尔信息矩阵或 K-FAC）

如果模式为 *=K-FAC* **则**

为编码器/解码器逐层构建克罗内克因子，以获得 $\widehat{F}_\phi$ 和 $\widehat{F}_\theta$;

$\widehat{F}_\theta^{\lambda} \leftarrow \widehat{F}_\theta + \lambda I$，$\widehat{F}_\phi^{\lambda} \leftarrow \widehat{F}_\phi + \lambda I$;

// 5) 求解自然方向（不显式反演）

通过共轭梯度法（CG）求解满足 v_θ 的 $\widehat{F}_\theta^{\lambda}\, v_\theta = g_\theta$；

通过 CG 求解满足 v_ϕ 的 $\widehat{F}_\phi^{\lambda} v_\phi = g_\phi$；

// 6) 参数更新（自然梯度）

$\theta_{t+1} \leftarrow \theta_t - \eta_\theta\, v_\theta$ （式 (26)）；

$\phi_{t+1} \leftarrow \phi_t - \eta_\phi\, v_\phi$ （式（23））；

返回 $(\phi_{t+1}, \theta_{t+1})$;

3）随后，我们计算解码器负对数似然和完整编码器目标的欧几里得梯度，

$$g_\theta = \nabla_\theta \frac{1}{B}\sum_{i=1}^{B}\left[-\log p_{\theta_t}(y_i \mid z_i)\right],$$

$$g_\phi = \nabla_\phi \left\{\frac{1}{B}\sum_{i=1}^{B}\left[-\log p_{\theta_t}(y_i \mid z_i)\right] + \beta\left(\widehat{\mathcal{L}}_{\mathrm{FR}} + \widehat{\mathcal{L}}_{\mathrm{JF}}\right)\right\}.$$

（步骤 4）为获得自然方向，我们为两个网络构建逐层 Kronecker 分解的费舍尔近似。对于每一层 ℓ 及其权重矩阵 W_ℓ，K-FAC 使用块对角模型 $F_\ell \approx A_\ell \otimes G_\ell$，其中 $A_\ell := \frac{1}{B}\sum_i a_{\ell,i} a_{\ell,i}^{\top}$，$a_{\ell,i}$ 是层输入（激活值）的协方差，而 $G_\ell := \frac{1}{B}\sum_i g_{\ell,i} g_{\ell,i}^{\top}$，$g_{\ell,i}$ 是反向传播的预激活梯度的协方差。蒂霍诺夫阻尼得到 $\widehat{F}_\theta^{\lambda} = \widehat{F}_\theta + \lambda I$ 和 $\widehat{F}_\phi^{\lambda} = \widehat{F}_\phi + \lambda I$。（步骤 5）我们不直接反演这些矩阵，而是通过共轭梯度（CG）求解线性方程组（$\widehat{F}_\theta^{\lambda}$）$v_\theta = g_\theta$ 和（$\widehat{F}_\phi^{\lambda}$）$v_\phi = g_\phi$。每次共轭梯度迭代只需计算费舍尔向量积，K-FAC 可高效地提供这些积：若 V_ℓ 将向量 v 重塑为该层的权重形状，则

$$\mathrm{FVP}_\ell(v) = \mathrm{vec}\Big((G_\ell + \lambda I)\, V_\ell\, (A_\ell + \lambda I)\Big).$$

（步骤 6）最后，沿自然方向更新参数：$\theta_{t+1} = \theta_t - \eta_\theta v_\theta$（参见公式（26））和 $\phi_{t+1} = \phi_t - \eta_\phi v_\phi$（参见公式（23））。超参数（$\eta_\phi$，$\eta_\theta$，$\beta$，$S$，$\lambda$）控制步长、压缩强度、估计量方差和条件数；K-FAC 则以曲率信号的忠实度为代价，实现可扩展且无需反演的自然梯度步进。

表2：数据集统计信息。

数据集	特征维度	#. 类别	#. 样本
MNIST [10]	28×28×1	10	70,000
CIFAR10 [18]	32×32×3	10	60,000
CelebA [22]	178×218×3	2（性别）	202,599

E 数据集

我们实验中使用的所有数据集的统计信息列于表2。MNIST 和 CIFAR10 均用于训练 10 类分类模型。CelebA 实验旨在识别人脸图像中的性别属性。该任务是一个二分类问题，与 MNIST 和 CIFAR10 上的任务不同。这些数据集涵盖了目标类别和学习复杂度各不相同的多种情形。下面将分别介绍这些数据集。

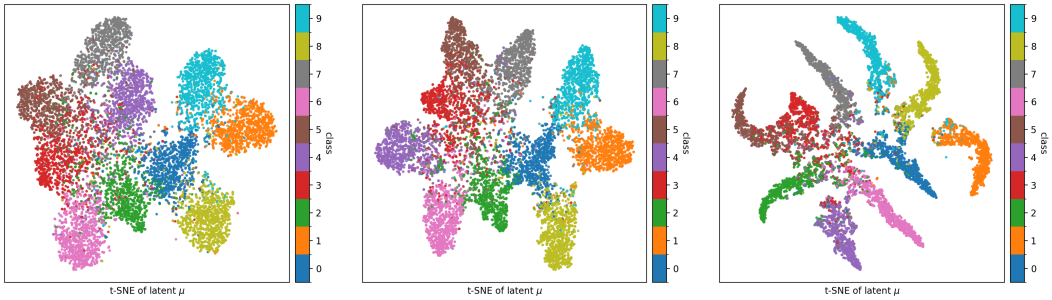
• **MNIST [10]**。 MNIST 包含 60,000 张用于训练的手写数字图像和 10,000 张用于测试的手写数字图像。所有这些黑白数字都经过尺寸归一化，并居中置于一张固定尺寸的图像中，该图像包含 28×28 个像素。

• **CIFAR10 [18]**。 CIFAR10 数据集包含 60,000 张 32x32 的彩色图像，分为 10 个类别，每个类别有 6,000 张图像。其中有 50,000 张训练图像和 10,000 张测试图像。

• **CelebA [22]**。 CelebA是一个大规模人脸属性数据集，包含超过 200,000 张名人图像，每张图像均标注了 40 项属性。

F 附加实验

定性可视化。 为考察瓶颈强度如何塑造表示，我们在三个不同的 β 值下，对 10,000 张 CIFAR10 测试图像的表示 $\mu(x) = \mathbb{E}[Z \mid x]$ 进行二维 t-SNE 可视化（图7）。当瓶颈乘数较小（$\beta = 10^{-4}$）时，各簇彼此分离良好，且类内分布相对较大，表明 Z 保留了输入的细粒度细节。当 $\beta = 10^{0}$ 时，表示簇向类别原型收缩，同时仍保持可分性。当 $\beta = 10^{1}$ 时，表征嵌入集中在类中心附近的狭窄弧线上，这与更强的压缩以及图4中观察到的准确率下降相一致。



(a) $\beta = 10^{-4}$，准确率 = 86.05%　　**(b)** $\beta = 10^{0}$，准确率 = 82.05%　　**(c)** $\beta = 10^{1}$，准确率 = 76.11%

图7：在 CIFAR10 上将 10000 张测试图像的表征嵌入可视化为二维图像。图像根据其真实类别标签着色。我们 β 增大时，会逐渐遗忘更多关于输入的信息，并且每个类别的表征嵌入都会压缩到接近平均值 μ的位置。我们还报告了测试准确率，该准确率会随着 β 增大而下降。