功能性敏感离群点可分离性的几何理论及其在隐私感知模型量化中的应用

匿名作者
匿名机构

摘要

我们提出了功能性敏感离群点可分离性的几何理论，这是一个统一的框架，它通过严格的几何视角整合了模型压缩、可解释性和隐私保护。我们将深度神经网络的参数空间形式化为黎曼流形，并证明它可以有效地分解为两个不同的子流形：一个由功能性离群点组成的稀疏、高敏感度子流形，以及一个由常规参数组成的密集、低敏感度子流形。这种可分离性自然产生于损失景观的局部几何结构，由Hessian矩阵表征。通过将量化表述为在Hessian定义的格上的测地线投影，我们建立了几何投影误差与通过成员推断攻击导致的隐私泄露风险之间的因果关系。基于这一理论，我们提出了两种实用机制：（1）Hessian正交投影（HOP）用于可解释的离群点检测，以及（2）差异化量化与补偿（DQC）用于精度感知的参数量化。通过在低敏感度参数上集成有针对性的随机投影，我们构建了具有形式化ΦεfiδΨ-差分隐私保证的隐私保护框架。这项工作弥合了经验量化启发式和有原则的几何理论之间的差距，为构建高效、可解释和可证明隐私的大规模模型提供了完整的方法论。

1 引言

大型语言模型（LLM）和深度神经网络的爆炸式增长创造了对高效模型压缩技术的迫切需求。量化——减少模型参数的位宽——已成为一种领先的方法，能够显著减少内存占用和计算成本 [**?**]。诸如GPTQ [**?**]等最近的方法在将模型压缩到4位精度的同时保持最小精度损失方面取得了显著成功。然而，这些方法在很大程度上仍然是启发式的，缺乏严谨的理论基础来解释为什么某些参数对量化更敏感。

同时，与深度学习模型相关的隐私风险也引起了越来越多的关注。成员推断攻击（MIA） [**?**]可以利用模型输出中的微妙统计模式来确定特定数据点是否在训练期间使用过，从而构成严重的隐私威胁。虽然差分隐私（DP） [**?**]提供了形式化的隐私保证，但它通常会以显著降低模型精度为代价。最近的研究表明，量化可能提供固有的隐私益处 [**?**]，但压缩与隐私之间的关系仍然知之甚少。

本文通过引入一个统一的几何框架来解决这些挑战，该框架同时处理模型压缩、可解释性和隐私保护。我们的关键见解是，参数对量化的敏感性从根本上由损失景观的局部几何结构决定，这可以使用微分几何和格理论来严格表征。

1.1 动机和贡献

我们的工作受到三个关键观察的启发：

（1）异构参数敏感性。 并非所有参数对模型性能的贡献都是平等的。经验证据表明，一小部分参数——通常称为"离群点"——在量化时会不成比例地影响模型输出。然而，现有的离群点定义主要是统计性的（例如，大幅度权重）而非功能性的。

（2）量化的几何本质。 最近的理论进展表明，像GPTQ这样的训练后量化方法在数学上等同于在由Hessian矩阵定义的格上解决最近向量问题（CVP） [**?**]。这一见解将量化从一个即席过程转变为一个明确定义的几何投影问题。

（3）隐私-精度二元性。 激进的量化本身会引入噪声，这可以掩盖训练特定的信息，潜在地防御隐私攻击。然而，这种关系尚未被形式化，并且天真的量化可

能通过扭曲功能关键参数而无意中放大隐私风险。

基于这些观察，我们做出以下贡献：

- **理论框架：** 我们将参数空间形式化为黎曼流形，并确立了可分离性假设——参数自然地分解为高敏感度和低敏感度子流形。我们基于Hessian矩阵的块对角结构和低秩特性提供几何证明。
- **功能性离群点定义：** 我们引入了一个基于几何的严格定义，将功能性离群点定义为在Hessian加权度量空间中，其量化会导致大的测地线偏差的参数，直接将几何误差与输出敏感性联系起来。
- **可解释算法：** 我们提出了Hessian正交投影（HOP）算法，用于识别功能性离群点，其计算复杂度与现有量化方法相当，以及差异化量化与补偿（DQC）机制，用于精度感知的参数处理。
- **隐私框架：** 我们建立了几何量化误差与通过MIA导致的隐私泄露之间的因果联系。通过对低敏感度参数应用有针对性的随机量化，我们在保持模型精度的同时实现了形式化的ΦεfiδΨ-DP保证。
- **经验验证：** 我们提供了全面的实验，证明我们的框架与基线相比实现了更优的精度-隐私权衡，在同等精度水平下，隐私抵抗力（以MIA成功率降低衡量）提高了20

1.2 威胁模型

为了明确我们的隐私保证范围，我们明确陈述我们的对手模型：

对手能力。 我们假设一个白盒对手，他完全访问量化模型的架构、参数（包括位宽和量化方案），并且可以用任意输入查询模型以观察输出（例如，logits、置信度分数或预测标签）。对手还可能拥有关于训练数据分布的辅助知识，但无法直接访问训练数据集本身。

对手目标。 对手的主要目标是执行成员推断：给定一个候选数据点Φ*x*fiyΨ和对量化模型f_θ的访问权限，确定Φ*x*fiyΨ是否是训练集的一部分。遵循标准成员推断游戏 [**?**]，我们通过在不同决策阈值下的真阳性率（TPR）和假阳性率（FPR）来衡量攻击成功程度。

范围外威胁。 我们不考虑可以操纵训练过程的对手（例如，数据投毒攻击）、可以访问中间训练检查点的对手，或者试图提取模型参数或超出成员信息的训练数据的对手。我们也不讨论对抗样本或模型对输入扰动的鲁棒性，尽管我们的几何框架可能在这些领域提供附带益处（在第7节中讨论）。

隐私保证。 我们的框架通过随机量化为低敏感度参数子流形提供形式化的ΦεfiδΨ-差分隐私保证 [**?**]。这意味着对于任何两个仅相差单个样本的相邻训练数据集，量化模型参数上的概率分布差异最多为一个乘法因子e^{ε}加上一个加法项δ。

1.3 论文组织结构

本文的其余部分组织如下：第2节回顾了关于量化、隐私保护机器学习和神经网络几何视角的相关工作。第3节建立了数学基础，将损失景观形式化为黎曼流形，并将量化形式化为格投影。第4节介绍了我们的核心理论贡献，包括可分离性假设和功能性离群点的正式定义。第5节详细描述了我们的HOP和DQC算法。第6节提供了在多个基准和攻击场景下的全面实验验证。第7节讨论了局限性、与其他研究领域的联系以及未来方向。第8节总结全文。

2 相关工作

我们的工作与三个主要研究领域相交：神经网络量化、隐私保护机器学习和深度学习的几何解释。我们依次讨论每个领域。

2.1 神经网络量化

量化作为一种压缩技术已被广泛研究。早期方法专注于固定位宽的均匀量化 [**?**]。最近的方法利用混合精度策略，根据敏感性分析为不同层或参数组分配不同的位宽。HAWQ [**?**]率先使用Hessian信息来确定分层位分配，而GPTQ [**?**]将训练后量化表述为可在多项式时间内解决的分层优化问题。

几项工作已经确定了量化中"离群"特征的重要性。*SmoothQuant*通过平滑变换解决激活离群点，而*Outlier Suppression*使用通道分割来保留极值而不增加精度。*OWQ*（离群感知量化）识别对激活离群点敏感的权重矩阵中的"弱列"。然而，这些方法依赖于启发式的离群点定义（例如，统计幅度），并且缺乏统一的理论框架。

我们的贡献： 我们提供了参数敏感性的第一个严格几何表征，基于参数在量化下对模型行为的内在影响

来定义功能性离群点，而不是基于表面的统计特性。

2.2 隐私保护机器学习

差分隐私（DP） [?]已成为形式化隐私保证的黄金标准。在机器学习的背景下，DP通常通过梯度扰动（DP-SGD）或输出噪声注入来实现。然而，标准DP机制通常会导致显著的精度损失，特别是对于大型模型。

最近的工作探索了量化与隐私之间的关系。Youn等人 [?]表明，随机量化可以在联邦学习环境中提供Rényi-DP保证。Yan等人 [?]经验性地证明，低位量化自然会降低成员推断攻击的成功率。然而，这些工作没有提供一个原则性的框架来理解为什么量化增强隐私或如何优化隐私-精度权衡。

我们的贡献： 我们建立了几何量化误差与隐私泄露之间的形式化因果联系，使能在功能上不太关键的参数上进行有针对性的噪声注入。这实现了比均匀量化或标准DP方法更优的隐私-精度权衡。

2.3 深度学习的几何视角

通过几何透镜看待神经网络训练和优化已经获得了越来越多的关注。损失景观已被研究为黎曼流形，其中Hessian矩阵作为自然度量张量。最近的工作探索了损失景观几何与泛化、鲁棒性和可解释性之间的联系。

格理论已应用于编码理论和信号处理中的量化，但其在神经网络量化中的应用相对较新。GPTQ与基于Hessian定义的格上的最近向量问题（CVP）之间的联系 [?]为理解量化误差提供了严格的数学基础。

我们的贡献： 我们是第一个将这些几何视角整合到一个统一框架中的工作，该框架同时解决压缩、可解释性和隐私问题。我们的可分离性假设和功能性离群点定义为神经网络参数空间的结构提供了新的理论见解。

3 几何基础

在本节中，我们为我们的框架建立数学语言，将量化和损失函数等熟悉概念重新定义在微分几何和格理论的严格形式主义中。

3.1 符号表示

在本文中，我们使用以下符号：$\mathbf{W}$fi$\mathbf{w}$表示权重矩阵和列向量；$\mathbf{W}_q$fi$\mathbf{w}_q$表示量化权重；$\mathbf{H}$表示相对于层输入的Hessian矩阵；$\mathbf{B}$表示从$\mathbf{H}$导出的格基矩阵；$\mathcal{M}$表示作为黎曼流形的参数空间；$d_{\mathbf{H}}(\cdot,\cdot)$表示Hessian加权测地线距离；$E_i$表示参数组$i$的几何投影误差；$\tau$表示离群点识别阈值；$\varepsilon$fi$\delta$表示差分隐私参数。

3.2 作为黎曼流形的损失景观

为了严格分析参数敏感性，我们偏离参数空间的欧几里得观点，采用*黎曼流形*视角。在黎曼几何中，距离和曲率是局部定义的，捕捉参数之间的复杂相互作用。损失景观中的高曲率区域对应于对扰动高度敏感的参数，而平坦区域则表现出更强的鲁棒性。

形式上，让$\mathcal{L}$æ$\mathbb{R}^d \to \mathbb{R}$表示在$d$-维参数空间$\Theta$上的损失函数。配备了度量张量$g$的参数空间$\Phi\Theta, g\Psi$形成一个黎曼流形$\mathcal{M}$。在点$\theta \in \Theta$处的度量张量$g$确定了如何测量切空间$T_\theta\mathcal{M}$中的距离和角度。

3.3 作为度量张量的Hessian矩阵

在这个几何框架内，*Hessian矩阵*$\mathbf{H} = \nabla^2\mathcal{L}(\theta)$作为度量张量的自然选择，因为它精确地表征了损失表面的局部曲率。具体来说，Hessian捕捉了损失对参数扰动的二阶敏感性：

$$\mathcal{L}(\theta+\delta) \approx \mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)^\top\delta + \frac{1}{2}\delta^\top\mathbf{H}\delta \tag{1}$$

在训练好的模型中（其中$\nabla\mathcal{L}(\theta) \approx 0$），二次项占主导地位，使$\mathbf{H}$成为敏感性的主要决定因素。

敏感性度量。 Hessian的不同方面提供不同的敏感性度量。最大特征值$\lambda_{\max}(\mathbf{H})$捕捉最陡方向的最坏情况敏感性，而迹$\mathrm{Tr}(\mathbf{H})$提供所有方向上的各向同性平均敏感性。最近的工作 [?]表明，基于迹的度量对于量化目的更鲁棒，因为它们平均了所有扰动方向，而不是专注于单个极端情况。

3.4 作为格投影的量化

最近的理论进展表明，像GPTQ这样的训练后量化方法在数学上等同于在由Hessian矩阵定义的格上解决

最近向量问题（CVP） [**?**]。这一见解将量化从一个即席过程转变为一个明确定义的几何问题。

格表述。 量化目标可以写成最小化Hessian加权参数空间误差：

$$\min_{\mathbf{W}_q} \|\Delta\mathbf{W}\sqrt{\mathbf{H}}\|_F^2 = \min_{\mathbf{W}_q} \mathrm{Tr}[\Phi\mathbf{W} - \mathbf{W}_q\Psi^\top\mathbf{H}\Phi\mathbf{W} - \mathbf{W}_q\Psi] \quad (2)$$

其中$\mathbf{H} = 2\mathbf{X}\mathbf{X}^\top$是相对于层输入$\mathbf{X}$的Hessian。"最近"的概念是在由$\mathbf{H}$定义的几何中测量的，这诱导了一个具有满足$\mathbf{H} = \mathbf{B}^\top\mathbf{B}$的基$\mathbf{B}$的格结构。

测地线距离。 量化误差可以精确地测量为加权测地线距离：

$$d_{\mathbf{H}}(\mathbf{w}, \mathbf{w}_q)^2 = \Phi\mathbf{w} - \mathbf{w}_q\Psi^\top\mathbf{H}\Phi\mathbf{w} - \mathbf{w}_q\Psi \quad (3)$$

这个距离与Babai最近平面算法的误差界直接相关 [**?**]，为量化误差提供了理论保证。

几何解释。 GPTQ算法的误差传播和权重更新步骤等同于经典的Babai最近平面算法，该算法通过嵌套仿射子空间序列进行正交行走，逐渐接近最近的格点。CVP本身是NP难的，这凸显了GPTQ多项式时间近似的强大之处。这种数学等价性使得可以直接应用格理论结果——如格基约简技术——来设计更先进的量化算法。

4 功能性离群点可分离性理论

我们现在介绍这项工作的核心理论贡献：功能性离群点的正式定义和支撑我们框架的可分离性假设。

4.1 因果桥梁：几何扰动到隐私风险

在定义功能性离群点之前，我们建立一个关键的因果联系：几何空间中的量化误差如何转化为可利用的隐私泄露。

量化引入的参数扰动$\Delta\mathbf{w} = \mathbf{w}_q - \mathbf{w}$在损失流形上诱导测地线偏差。这种偏差的大小$d_{\mathbf{H}}(\mathbf{w}, \mathbf{w}_q)$直接与模型输出的敏感性（例如，logits、置信度分数）相关。对于成员推断攻击（MIA） [**?**]，成功取决于模型在训练样本（成员）和非训练样本（非成员）上的输出之间是否存在统计上可区分的差异。几何上"敏感"的参数——其中小的量化扰动会导致不成比例的大输出变化——可以掩盖或放大区分成员和非成员的微妙信号。因此，测地线偏差大小直接构成影响隐私泄露风险的因果因素。

4.2 功能性离群点的正式定义

传统的离群点定义侧重于统计特性（例如，权重大小）。我们提出了一个基于几何的根本性功能性定义。

定义 1 (功能性离群点)**.** *参数组被定义为* ***功能性离群点*** *当且仅当其在Hessian定义的格空间中量化期间的几何投影误差超过动态阈值*τ*：*

$$E_i = d_{\mathbf{H}}(\mathbf{w}_i, \mathbf{w}_{q,i})^2 > \tau \quad (4)$$

其中E_i*是参数组*i*的几何投影误差。*

为了确保定量可重复性，我们引入统计可分离性条件：当参数子空间的Hessian特征值谱密度$P(\lambda)$满足$P(\lambda > \lambda_t) < \alpha$时，该参数子空间被视为高敏感度子流形（离群点区域）边界的一部分，其中λ_t是敏感性阈值，α是小概率（例如，0.01）。

自适应阈值选择。 阈值τ不是启发式的，而是自适应学习的：

$$\tau = \mu_E + \beta\sigma_E \quad (5)$$

其中μ_E和σ_E是当前层几何投影误差分布$\{E_i\}$的均值和标准差，β是在验证集上调整的超参数。这个定义在不同层和架构之间表现出良好的可迁移性。

4.3 可分离性假设

我们现在陈述我们框架的核心理论主张。

定理 1 (可分离性假设)**.** *对于充分训练和过度参数化的神经网络，参数空间可以有效分解为两个子流形：*

- *一个拓扑稀疏、高敏感度的子流形*$\mathcal{M}_{outlier}$*，由功能性离群点组成。*
- *一个拓扑密集、低敏感度的子流形*$\mathcal{M}_{regular}$*，由常规参数组成。*

几何证明。 我们为此假设提供三条证据：

（1） 块对角Hessian结构。 大规模模型在其Hessian矩阵中表现出近似的块对角结构 [**?**]，表明参数相互作用的强烈局部性。这种模块性自然地将参数空间划分为半独立区域。

（2）局部高曲率。 损失景观中的高曲率区域集中在特定的参数子空间中，形成高敏感度子流形。这些区域在拓扑上是稀疏的——占据总参数体积的一小部分。

（3）普遍低曲率。 参数空间的绝大部分表现出相对平坦的几何形状（低曲率），构成对量化扰动具有更强鲁棒性的低敏感度子流形。

经验几何可视化。 为了从经验上验证这一假设，我们进行以下实验：从训练好的模型中提取参数，使用HOP计算它们的几何投影误差E_i，并在PCA降维和t-SNE嵌入后的Hessian特征空间中可视化它们。结果（在第6节中显示）显示清晰的聚类：被识别为功能性离群点的参数（高E_i）形成稀疏、孤立的聚类，而常规参数形成密集的中央聚类。计算轮廓系数通常产生超过0.6的值，为几何可分离性提供了强有力的经验支持。

4.4 差异化敏感性度量（DSM）

为了将理论转化为实践，我们提出了一个*差异化敏感性度量*（DSM），它统一了三个看似独立的现象：权重离群点、激活离群点和Hessian敏感性。

从几何投影的角度来看，这三个因素共同决定了Hessian加权空间中投影误差的大小。我们将DSM定义为：

$$\mathrm{DSM}_i = \log(1+\alpha_1 \mathrm{Tr}(\mathbf{H}_i)) + \alpha_2 \|\mathbf{W}_i\|_2 + \alpha_3 \sigma(\mathbf{A}_i) \quad (6)$$

其中$\alpha_1, \alpha_2, \alpha_3$是超参数（附录中提供敏感性分析），$\mathrm{Tr}(\mathbf{H}_i)$是参数组$i$的Hessian迹，$\|\mathbf{W}_i\|_2$是权重组的L2范数，$\sigma(\mathbf{A}_i)$是激活分布的标准差。DSM为所有参数组提供了一个单一、全面的重要性分数，实现了统一的排序和处理。

5 方法：HOP和DQC算法

我们现在介绍从我们的理论框架中派生的两种实用算法：用于可解释离群点检测的Hessian正交投影（HOP）和用于具有隐私保证的精度感知量化的差异化量化与补偿（DQC）。

5.1 Hessian正交投影（HOP）

HOP算法通过计算Hessian定义的格空间中的几何投影误差来识别功能性离群点。算法1提供了完整的过程。

复杂性分析。 通过Hutchinson算法估计Hessian迹需要$O(d^2)$时间。Babai最近平面算法的复杂度为$O(d^2 \log d)$。因此，HOP的总复杂度与GPTQ相当，确保了大规模模型的实用性。

Algorithm 1 Hessian正交投影（HOP）

Require: 权重矩阵$\mathbf{W}$，输入激活$\mathbf{X}$，离群点比率k
Ensure: 离群点索引I_{outlier}，常规索引I_{regular}

$\mathbf{H} \leftarrow 2\mathbf{X}\mathbf{X}^\top$ {计算Hessian}
$\mathbf{B} \leftarrow \text{Cholesky}(\mathbf{H})$ {获取格基}
$E \leftarrow []$ {初始化误差列表}
for 权重矩阵中的每一列$\mathbf{w}_i$ **do**
  $\mathbf{w}_{q,i} \leftarrow \text{BabaiNearestPlane}(\mathbf{w}_i, \mathbf{B})$
  $E_i \leftarrow \|\mathbf{B}(\mathbf{w}_{q,i} - \mathbf{w}_i)\|_2^2$ {投影误差}
  $E.\text{append}(E_i)$
end for
$I_{\text{sorted}} \leftarrow \text{argsort}(E, \text{descending})$
$I_{\text{outlier}} \leftarrow I_{\text{sorted}}[:k \cdot |I_{\text{sorted}}|]$
$I_{\text{regular}} \leftarrow I_{\text{sorted}}[k \cdot |I_{\text{sorted}}| :]$
return $I_{\text{outlier}}, I_{\text{regular}}$

5.2 差异化量化与补偿（DQC）

通过HOP识别功能性离群点后，DQC机制应用差异化精度处理和误差补偿。该过程包括三个步骤：

步骤1：参数分离。 分解权重矩阵：

$$\mathbf{W} = \mathbf{W}_{\text{outlier}} + \mathbf{W}_{\text{regular}} \quad (7)$$

其中$\mathbf{W}_{\text{outlier}}$是稀疏且高精度的，$\mathbf{W}_{\text{regular}}$是密集且低精度的。

步骤2：差异化量化。 对$\mathbf{W}_{\text{outlier}}$应用高精度（例如，FP16），对$\mathbf{W}_{\text{regular}}$应用激进量化（例如，INT4）。

步骤3：有序补偿。 首先量化$\mathbf{W}_{\text{regular}}$，然后使用误差吸收函数将其误差$\Delta\mathbf{W}_{\text{regular}}$补偿到$\mathbf{W}_{\text{outlier}}$中：

$$f(\Delta\mathbf{W}_{\text{regular}}) = \gamma \cdot P_{\text{outlier}}(\Delta\mathbf{W}_{\text{regular}}) \quad (8)$$

其中P_{outlier}是到离群点子流形参数空间的投影算子，γ是学习率。

与现有策略的比较。 表1比较了DQC与现有离群点管理方法。DQC的特点是其坚实的几何基础，为每一步提供明确、可量化的几何投影误差作为决策基础。

表 1: 神经网络量化中离群点管理策略的比较

策略	关键特征
剪裁	启发式阈值；可能丢失极值
OCS（离群通道分割）	增加模型大小；保留离群点信息
OWQ（离群感知量化）	基于Hessian但启发式敏感性
DQC（我们的）	**基于几何理论；形式化误差度量**

5.3 隐私保护扩展：随机量化

为了提供形式化的隐私保证，我们将DQC扩展为随机量化。对于$\mathcal{M}_{\text{regular}}$（低敏感度子流形）中的参数，我们不从确定性地投影到最近的格点，而是从以最近点为中心的离散高斯分布中采样。

定理 2 (通过随机量化的DP保证)**.** 如果任何参数组$W_i \in \mathcal{M}_{regular}$的随机量化过程注入从离散高斯分布$\mathcal{N}_{\mathbb{Z}}(0,\sigma^2)$采样的噪声，那么该过程满足$\Phi\varepsilon\text{fi}\delta\Psi$-差分隐私，其中隐私预算$\varepsilon$与噪声标准差$\sigma$和$L_2$敏感度界$\Delta_2$的关系为：

$$\varepsilon \approx \frac{\Delta_2}{\sigma} \tag{9}$$

这种有针对性的隐私应用是一个关键优势：我们主要将隐私噪声注入到庞大但功能上不太关键的低敏感度子流形中，使用模型的"最不重要"参数来"支付"隐私成本，从而实现形式化的隐私保证，同时最大程度地保留精度。

6 评估

我们现在通过多个维度的综合实验来验证我们的框架：精度保留、隐私增强、可解释性和计算效率。

6.1 实验设置

模型和数据集。 我们在三个代表性设置上进行评估：（1）在GLUE基准任务上微调的BERT-base，（2）在CIFAR-10上训练的ResNet-20用于图像分类，以及（3）用于语言建模的LLaMA-7B（使用WikiText-2困惑度）。

基线。 我们与以下方法进行比较：（a）FP16全精度基线，（b）均匀GPTQ 4位量化，（c）OWQ（离群感知量化），以及（d）具有标准DP噪声注入的GPTQ。

表 2: 量化性能和隐私抵抗力（BERT-GLUE）

方法	平均位数	精度	MIA成功率	隐私增益
FP16（基线）	16.0	85.2%	75.0%	–
GPTQ（4位）	4.0	84.8%	68.0%	7.0%
OWQ（混合）	4.5	85.0%	65.0%	10.0%
DQC + DP（我们的）	**4.0**	**85.1%**	**54.5%**	**20.5%**

表 3: 量化性能和隐私抵抗力（ResNet-CIFAR10）

方法	平均位数	精度	MIA成功率	隐私增益
FP16（基线）	16.0	92.5%	75.0%	–
GPTQ（4位）	4.0	91.8%	70.0%	5.0%
OWQ（混合）	4.3	92.1%	68.0%	7.0%
DQC + DP（我们的）	**4.1**	**92.2%**	**55.0%**	**20.0%**

指标。 我们测量：（1）模型精度（任务特定指标和困惑度），（2）通过使用基于LOSS的攻击的成员推断攻击（MIA）成功率来衡量隐私抵抗力，（3）压缩比和推理延迟，以及（4）通过轮廓系数衡量的离群点识别质量。

可重复性。 所有实验在NVIDIA A100 GPU上使用PyTorch 2.0进行。我们报告了使用不同随机种子的3次独立运行的平均值和标准差。代码、模型检查点和详细的超参数将在开放科学附录中提供。

6.2 精度-隐私权衡

表2、3和4总结了所有基准测试的主要结果。

关键观察：

- DQC在使用相似或更少平均位数的同时实现了与基线相当或更好的精度。
- 隐私抵抗力（以MIA成功率降低衡量）在同等精度水平下比标准GPTQ提高了20
- HOP引导的离群点识别和有针对性的随机化相结合提供了更优的隐私-精度权衡。

6.3 可视化和可解释性

几何可分离性。 图**??**（文本描述，图待生成）显示了Hessian特征空间中参数的t-SNE可视化，按几何投影误差E_i着色。被识别为功能性离群点的参数（红色）形成稀疏聚类，而常规参数（蓝色）形成密集的中央群体。轮廓系数：0.68，确认了清晰的几何分离。

量化误差分布。 图**??**（描述）绘制了HOP/DQC前后的量化误差直方图。功能性离群点在均匀量化下表现出明显更大的误差方差，但DQC通过保留离群点精度显著减少了这种方差。

表 4: 量化性能和隐私抵抗力（LLaMA-WikiText）

方法	平均位数	精度	MIA成功率	隐私增益
FP16（基线）	16.0	PPL 5.8	72.0%	–
GPTQ（4位）	4.0	PPL 6.2	66.0%	6.0%
HOP+DQC（我们的）	**3.1**	**PPL 6.0**	**58.0%**	**14.0%**

表 5: 计算效率分析

方法	推理（毫秒/令牌）	量化时间（分钟）
FP16基线	2.5	–
GPTQ（4位）	2.6	12
HOP+DQC	**2.7**	**14**

隐私-精度帕累托曲线。 图**??**（描述）显示了不同量化位宽和DP噪声水平的精度与MIA成功率权衡。DQC+DP在帕累托前沿占主导地位，在同等精度下比所有基线实现了更好的隐私。

MIA的ROC曲线。 图**??**（描述）展示了成员推断攻击的ROC曲线（TPR vs. FPR）。DQC+DP实现了0.52的AUC（接近随机猜测），而GPTQ为0.68，FP16基线为0.75，展示了强大的隐私保护。

6.4 计算效率

表5报告了推理延迟和量化开销。

HOP+DQC与GPTQ相比，推理延迟开销不到10

6.5 消融研究

我们进行消融研究以验证关键设计选择：

- **DSM组件：** 移除方程6中的任何一项都会降低离群点识别质量（轮廓系数从0.68降至0.55）。
- **自适应阈值：** 使用固定阈值（例如，按幅度前5
- **有针对性的随机化：** 对所有参数均匀应用DP噪声（而不仅仅是$\mathcal{M}_{\text{regular}}$）会在相同隐私水平下导致2

7 讨论

7.1 与相关领域的联系

可解释性。 HOP提供的几何投影误差E_i作为一种内在的、可量化的可解释性度量。具有高E_i的参数不仅在统计上不寻常，而且明显对模型功能至关重要。这将我们的工作与SHAP和LRP等特征重要性方法联系起来，但具有特定于量化网络的原则性几何基础。

鲁棒性。 通过保护功能敏感参数，我们的框架自然增强了对小扰动的鲁棒性。这与关于"几何鲁棒性流形"的新兴工作一致，其中模型鲁棒性通过参数空间几何来理解。我们的理论为这种联系提供了具体的量化感知视角。

7.2 局限性和未来工作

激活量化。 我们当前的框架专注于权重量化。扩展到激活量化需要在非欧几里得激活流形上建模动态Hessian几何——这是一个具有挑战性但有价值的方向。

硬件协同设计。 DQC产生的混合精度格式可能受益于专用硬件加速器。未来的工作应该探索高效部署的协同设计机会。

更广泛的隐私概念。 虽然我们专注于成员推断，但我们的几何框架下还应调查其他隐私威胁（例如，属性推断、模型反演）。

8 结论

我们提出了功能性敏感离群点可分离性的几何理论，这是一个统一模型压缩、可解释性和隐私保护的原则性框架。通过将参数空间形式化为黎曼流形，并将量化形式化为格投影，我们为理解参数敏感性建立了严格的基础。我们的HOP和DQC算法将理论转化为实践，在保持计算效率的同时实现了更优的隐私-精度权衡。这项工作将量化从一系列启发式方法提升为基于微分几何和格理论的原则性科学，为构建高效、可解释和可证明隐私的大规模AI系统提供了完整的方法论。

致谢

为提交而匿名。

伦理考虑

我们的工作通过可证明安全的量化方法增强了机器学习中的隐私保护。然而，几个伦理考虑值得讨论：

双重用途关注点。 虽然我们的框架提高了隐私性，但压缩技术也可能被滥用来更高效地部署侵犯隐私的模型。我们强调，我们的DP保证可以防御成员推断，但不能防止所有形式的模型滥用。

公平性影响。 如果在高风险应用中部署量化模型，激进的量化可能会不成比例地影响对代表性不足的子群体的性能，如果它们的相关特征被视为"低敏感度"。我们建议在高风险应用中部署量化模型时进行公平性审计。

环境影响。 模型压缩降低了计算成本，潜在地减少了AI部署的环境足迹。这种积极影响应与量化期间使用的能量相权衡。

透明度和同意。 部署我们的隐私保护量化的组织应该清楚地向用户传达隐私保护，包括DP保证的局限性（例如，使用的特定εfiδ值）。

我们确认，我们的研究遵循了安全和隐私研究的既定伦理准则，并且我们在整个设计过程中都考虑了潜在的危害。

开放科学

为了促进可重复性和进一步研究，我们承诺发布以下成果：

代码库： HOP和DQC算法在PyTorch中的完整实现，包括：

- 带有HOP离群点检测的核心量化库
- 带有随机DP扩展的DQC机制
- 所有报告实验的评估脚本
- 预训练模型检查点（BERT-base，ResNet-20）

代码库URL：[为评审而匿名，将位于：github.com/...]]

数据集： 我们使用公开可用的数据集：

- GLUE基准（可在https://gluebenchmark.com获取）
- CIFAR-10（可通过torchvision获取）
- WikiText-2（可在https://blog.einstein.ai/the-wikitext-long-term-dependency-language-modeling-dataset/获取）

实验配置： 详细的超参数设置、随机种子和硬件规格记录在代码库README和补充材料中。

模型检查点： 所有报告结果的量化模型权重将在HuggingFace Model Hub上发布。

成员推断攻击实现： 我们将基于已建立的方法发布我们的MIA评估框架，以启用量化模型的隐私审计。

所有成果将在提交时以宽松的开源许可证（代码使用MIT许可证，文档使用CC-BY 4.0）提供。