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

Yunhao Wang
Lenovo Research

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(ε, δ)-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(ε, δ)-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(x, y)和对量化模型f_θ的访问权限，确定(x, y)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(ε, δ)-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：**W**fi**w**表示权重矩阵和列向量；$\mathbf{W}_q$fi$\mathbf{w}_q$表示量化权重；**H**表示相对于层输入的Hessian矩阵；**B**表示从**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$(\Theta, g)$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**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：

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

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：

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

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 从几何扰动到信息泄露— 形式化桥梁

本节建立了几何扰动（由量化决定）与信息差异（KL / 可辨别性 / MIA 指标）之间的定量桥梁，为后续 HOP 与 DQC 的设计提供了理论支柱：HOP 用以识别高d_H的"功能性离群点"，而对低d_H子流形的随机化能以可解释的方式"购买"差分隐私。

4.1 记号与假设

令$L(\theta)$为参数空间上的训练损失，$\hat{\theta}$为训练完毕的参数点。设 Hessian $H := \nabla^2 L(\hat{\theta})$存在且正定（或在感兴趣子空间上正定）。对小参数扰动$\Delta\theta$（例如由量化产生），我们采用局部二阶近似：

$$\Delta L \approx \frac{1}{2}\Delta\theta^\top H \Delta\theta. \quad (4)$$

此外假设模型的输出分布$p(y \mid x,\theta)$关于θ在$\hat{\theta}$处光滑，且 Fisher 信息可由H或其近似替代。

4.2 几何扰动定义

定义参数扰动在 Hessian 度量下的局部平方测地线能量：

$$E(\theta,\theta+\Delta\theta) := d_H(\theta,\theta+\Delta\theta)^2 = \Delta\theta^\top H \Delta\theta. \quad (5)$$

4.3 定理（局部 KL 上界）

在上述假设下，对于任意输入x与小扰动$\Delta\theta$，有

$$\mathrm{KL}\big(p(\cdot \mid x,\theta)\,\|\,p(\cdot \mid x,\theta+\Delta\theta)\big) \le \frac{1}{2}\Delta\theta^\top H \Delta\theta + o(|\Delta\theta|^2). \quad (6)$$

由此可得：预测分布的可辨别度（KL）由局部几何能量上界，从而将几何扰动直接映射到信息差异量度。

定理 1 (局部 KL 上界)**.** *对于充分训练的参数$\hat{\theta}$，Hessian $H=\nabla^2 L(\hat{\theta})$正定，以及任意输入$x$和小扰动$\Delta\theta$，预测分布之间的 KL 散度满足：*

$$\mathrm{KL}\big(p(\cdot \mid x,\hat{\theta})\,\|\,p(\cdot \mid x,\hat{\theta}+\Delta\theta)\big) \le \frac{1}{2}\Delta\theta^\top H \Delta\theta + o(|\Delta\theta|^2). \quad (7)$$

证明概要。 通过 Taylor 展开和 Fisher 信息矩阵的性质，KL 散度的二阶项由$\Delta\theta^\top H\Delta\theta$控制，其中$H$在局部近似 Fisher 信息矩阵。

4.4 推论（对 MIA 的约束）

若 MIA 的成功率与预测分布间的区分量（如 KL 或 logit 差异）相关，则样本/参数组级的d_H^2提供了该组对 MIA 贡献的上界估计。换言之，通过控制d_H（例如对低敏感度组随机化）可获得可解释的隐私改善。

引理 1 (对 MIA 的约束)**.** *如果成员推断攻击的成功率与预测分布之间的 KL 散度或 logit 差异相关，那么对于参数组i，其几何能量$E_i = d_H(\theta_i,\theta_{q,i})^2$提供了该组对 MIA 贡献的上界估计。*

4.5 对随机量化的 DP 解释

对低敏感度子流形$\mathcal{M}_{\text{regular}}$使用离散高斯随机量化，令该子流形的$H$-度量下的敏感度为$\Delta_2$，则根据高斯机制分析可得到近似：

$$\varepsilon \approx \frac{\Delta_2^2}{2\sigma^2}, \quad (8)$$

将隐私预算ε与局部几何敏感度联系起来。

定理 2 (通过随机量化的 DP 保证)**.** 对于低敏感度子流形$\mathcal{M}_{\text{regular}}$中的参数，使用离散高斯随机量化（标准差$\sigma$），若该子流形在$H$-度量下的$L_2$敏感度为$\Delta_2$，则量化过程满足$(\varepsilon,\delta)$-差分隐私，其中

$$\varepsilon \approx \frac{\Delta_2^2}{2\sigma^2}. \tag{9}$$

4.6 小结

本节建立了几何扰动（由量化决定）与信息差异（KL / 可辨别性 / MIA 指标）之间的定量桥梁，为后续 HOP 与 DQC 的设计提供了理论支柱：HOP 用以识别高d_H的"功能性离群点"，而对低d_H子流形的随机化能以可解释的方式"购买"差分隐私。

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。

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：

$$\mathbf{W} = \mathbf{W}_{\text{outlier}} + \mathbf{W}_{\text{regular}} \tag{10}$$

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）。

Algorithm 1 Hessian正交投影（HOP）

Require: Weight matrix $\mathbf{W}$, input activations $\mathbf{X}$, outlier ratio k

Ensure: Outlier indices I_{outlier}, regular indices I_{regular}

$\mathbf{H} \leftarrow 2\mathbf{X}\mathbf{X}^\top$ {Compute Hessian}
$\mathbf{B} \leftarrow \text{Cholesky}(\mathbf{H})$ {Get lattice basis}
$E \leftarrow []$ {Initialize error list}
for each column $\mathbf{w}_i$ in weight matrix **do**
  $\mathbf{w}_{q,i} \leftarrow \text{BabaiNearestPlane}(\mathbf{w}_i, \mathbf{B})$
  $E_i \leftarrow \|\mathbf{B}(\mathbf{w}_{q,i} - \mathbf{w}_i)\|_2^2$ {Projection error}
  $E.\text{append}(E_i)$
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 $I_{\text{outlier}}, I_{\text{regular}}$

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**

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：

$$f(\Delta\mathbf{W}_{\text{regular}}) = \gamma \cdot P_{\text{outlier}}(\Delta\mathbf{W}_{\text{regular}}) \tag{11}$$

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。

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 3 (通过随机量化的DP保证（方法部分）)**.** 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$(\varepsilon,\delta)$-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：

$$\varepsilon \approx \frac{\Delta_2}{\sigma} \tag{12}$$

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

[注意：本节所有实验内容均为计划中的实验设计，尚未实际完成。以下数据、图表和分析均为预期结果或占位符。]

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

[未完成：以下为实验设计计划]

模型和数据集。 我们计划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。 我们计划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。

指标。 我们计划测量：（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。 所有实验计划在NVIDIA A100 GPU上使用PyTorch 2.0进行。我们计划报告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。

表 2: 量化性能和隐私抵抗力（BERT-GLUE）[数据待生成]

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）[数据待生成]

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%**

6.2 精度-隐私权衡

[未完成：以下表格中的数据均为预期结果或占位符]

表2、3和4总结了所有基准测试的主要结果（数据待生成）。

关键观察（预期结果）：

- 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（待验证）。
- 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（待验证）。

6.3 可视化和可解释性

[未完成：以下所有图表均为计划生成，尚未实际完成]

几何可分离性。 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（预期值），确认了清晰的几何分离。

量化误差分布。 图**??**（待生成）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。

隐私-精度帕累托曲线。 图**??**（待生成）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。

表 4: 量化性能和隐私抵抗力（LLaMA-WikiText）[数据待生成]

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[数据待测量]

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**

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**??**（待生成）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（预期值，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

[未完成：以下数据为预期结果]

表5报告了推理延迟和量化开销（数据待测量）。HOP+DQC与GPTQ相比，推理延迟开销不到10

6.5 消融研究

[未完成：以下消融研究尚未实际执行]

我们计划进行消融研究以验证关键设计选择：

- **DSM组件：** 移除方程**??**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）（预期结果，待验证）。
- **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

伦理考虑

双重用途关注点。
公平性影响。
环境影响。
透明度和同意。