LibOrtho v2.0: 基于结构化曲率与动态正交性的大语言模型泛化与记忆几何分离

匿名作者
匿名机构

摘要

大语言模型（LLM）面临泛化与记忆之间的根本性张力。我们提出了LibOrtho v2.0框架，在LLM的参数空间中几何地分离通用知识与私有记忆。与依赖对角 Hessian近似的方法不同，我们采用Kronecker因子分解近似曲率（K-FAC）来捕捉Transformer架构中的参数交互。我们引入动态正交梯度下降（OGD）在微调过程中强制正交性，并利用基于SVD分解的任务向量算术来区分低秩能力更新与高秩记忆。我们的方法在TOFU基准测试中实现了95%+的记忆移除有效性，同时在 MMLU和GSM8K上保持了98%+的泛化能力。我们提供了正交性的理论保证，并证明了该方法可扩展到7B+参数模型。

1 引言

在隐私敏感应用中部署大语言模型（LLM）需要机制来防止训练数据泄露，同时保持通用推理能力。这一挑战在高维参数空间中表现为几何问题：通用知识应驻留在低维流形上，而私有记忆则表现为正交补空间中的高曲率扰动。

先前关于机器遗忘和隐私保护机器学习的工作一直受到机制性纠缠这一根本问题的困扰：在标准梯度下降中，模型倾向于复用现有特征，而不是为新信息创建正交子空间。这导致纠缠比为0.1–0.3，意味着记忆移除不可避免地会损害通用能力。

我们提出的LibOrtho v2.0解决了先前几何方法的三个关键局限性：

1. **对角Hessian陷阱**：先前方法将Hessian近似为对角矩阵，忽略了 Transformer架构中的参数交互。我们使用K-FAC来捕捉层间协方差结构。
2. **静态分离**：事后投影无法撤销训练过程中发生的纠缠。我们通过在微调过程中使用OGD动态强制正交性。
3. **曲率二元论误区**：“平坦即泛化，尖锐即记忆”的假设在复杂推理任务中失效。我们使用基于秩的分离，通过任务向量的SVD来区分能力与记忆。

我们的贡献包括：(1) 结合K-FAC、OGD和任务向量算术的理论框架；(2) 在7B+ 参数模型上的实验验证，显示95%+的记忆移除且能力损失最小；(3) 对现代LLM中泛化与记忆几何结构的分析。

2 背景与相关工作

2.1 泛化的几何视角

流形假设 [**?**]认为高维数据位于低维流形上。对于LLM，这表明通用知识占据参数空间的稀疏子空间S_{gen}，内在维度$d \ll D$，其中 D是参数总数。

然而，最近关于顿悟（grokking） [**?**]和功能中心景观 [**?**]的研究挑战了简单的“平坦即泛化”假设。复杂推理能力通常需要尖锐的决策边界，这在参数空间中表现为高曲率方向。

2.2 机制性纠缠

机制性纠缠现象 [**?**]源于梯度下降最小化权重范数变化的趋势。模型倾向于微调现有特征，而不是创建正交子空间，导致纠缠比 $\eta = \frac{|\langle S_{gen}, S_{mem} \rangle|}{|S_{gen}||S_{mem}|}$为 0.1–0.3，远非理想的零值。

2.3 影响函数及其局限性

影响函数 [?]试图通过$H^{-1}\nabla l(z)$识别影响特定预测的训练样本。然而，在LLM中，由于接近零的特征值，逆Hessian在数值上不稳定，而对角近似在像Transformer这样的强耦合系统中引入严重误差。

3 理论框架

3.1 问题表述

设$W \in \mathbb{R}^D$表示LLM的参数。在私有数据$\mathcal{D}_{priv}$上微调后，模型参数变为$W_{ft} = W_{base} + \Delta W$。我们的目标是分解：

$$\Delta W = \Delta W_{gen} + \Delta W_{mem} \tag{1}$$

其中ΔW_{gen}保留通用能力，ΔW_{mem}仅包含私有记忆，且$\langle \Delta W_{gen}, \Delta W_{mem} \rangle = 0$。

3.2 K-FAC结构化曲率

对于权重矩阵$W \in \mathbb{R}^{d_{out} \times d_{in}}$的线性层，K-FAC将Hessian近似为：

$$H_{layer} \approx A \otimes G \tag{2}$$

其中$A = \mathbb{E}[aa^T]$是输入激活协方差，$G = \mathbb{E}[\nabla_{pre}\nabla_{pre}^T]$是预激活梯度协方差。特征分解得到：

$$A = U_A \Sigma_A U_A^T, \quad G = U_G \Sigma_G U_G^T \tag{3}$$

Hessian特征向量为$u_A \otimes u_G$，特征值为$\lambda_{A,i}\lambda_{G,j}$。我们将权重残差$\Delta W$投影到此特征基上：

$$C = U_G^T (\Delta W) U_A \tag{4}$$

其中C_{ij}表示在(i,j)-th特征方向上的分量，曲率为$\lambda_{G,i}\lambda_{A,j}$。

定理 1 (K-FAC分离)**.** *对于具有K-FAC结构的层，如果ΔW_{mem}主要位于高曲率方向（$\lambda_{G,i}\lambda_{A,j} > \tau$），则投影到低曲率方向得到$\Delta W_{gen}$，重构误差有界。*

3.3 动态正交梯度下降

为了在训练过程中强制正交性，我们修改私有数据的梯度更新：

$$g_{priv} = \nabla L_{priv}(W) \tag{5}$$

$$g_{update} = g_{priv} - P_{S_{gen}}(g_{priv}) \tag{6}$$

其中$P_{S_{gen}}$投影到通用知识子空间S_{gen}，定义为具有最小特征值（最平坦方向）的top-k K-FAC特征向量的张成空间。

定理 2 (正交性保证)**.** *如果g_{update}在每一步都被约束到$S_{gen}^{\perp}$，则通过构造$\Delta W_{mem} \perp S_{gen}$，纠缠比$\eta = 0$。*

3.4 基于SVD的任务向量算术

我们通过SVD分解权重增量$\tau = W_{ft} - W_{base}$：

$$\tau = U \Sigma V^T = \sum_{i=1}^{r} \sigma_i u_i v_i^T \tag{7}$$

关键洞察是通用能力更新是*低秩*的（可在多个样本上压缩），而记忆是*高秩*的（样本特定）。我们分离：

$$\tau_{gen} = \sum_{i=1}^{k} \sigma_i u_i v_i^T, \quad \tau_{mem} = \sum_{i=k+1}^{r} \sigma_i u_i v_i^T \tag{8}$$

其中k选择为捕获90%+的谱能量，同时排除高秩噪声。

4 方法论

4.1 算法概述

我们的LibOrtho v2.0流程包括三个阶段：

1. **子空间识别**：使用K-FAC在公共数据上从基础模型计算S_{gen}。
2. **约束微调**：对私有数据应用OGD，确保更新保持在$S_{gen}^{\perp}$中。
3. **事后精炼**：使用基于SVD的任务向量算术进一步分离低秩能力与高秩记忆。

4.2 实现细节

对于K-FAC计算，我们为每层维护A和G矩阵，在公共数据样本的前向-反向传播过程中更新它们。我们使用衰减率为0.95的指数移动平均。

对于OGD，我们通过存储top-k特征向量（$k = 200$–800，取决于层大小）并在每次更新步骤前投影梯度来计算$P_{S_{gen}}$。

对于SVD，我们计算每层τ的分解，选择k使得$\sum_{i=1}^{k}\sigma_i^2/\sum_{i=1}^{r}\sigma_i^2 \geq 0.9$。

4.3 计算复杂度

K-FAC每层需要$O(d_{in}^2 + d_{out}^2)$存储（vs. 完整Hessian的$O(d_{in}d_{out})$），使其适用于7B+模型。OGD为每次更新添加$O(k \cdot d_{in}d_{out})$用于投影。SVD在训练后计算一次，每层复杂度为$O(\min(d_{in}, d_{out})^3)$。

5 实验评估

5.1 实验设置

我们在LLaMA-7B和GPT-2（1.5B）模型上评估。私有数据包括来自TOFU [**?**] 基准测试的1000个金丝雀样本。公共数据是C4的10K子集，用于计算S_{gen}。

5.2 评估指标

- **记忆移除**：金丝雀提取率（目标：$< 1\%$）
- **泛化保持**：MMLU（57个任务）和GSM8K（数学推理）准确率
- **几何验证**：线性模式连通性（LMC）、纠缠比η、Hessian谱密度

5.3 结果

方法	金丝雀率	MMLU	GSM8K
基线（无移除）	98.5%	45.2	32.1
朴素剪枝	12.3%	38.1	25.4
LibOrtho v1.0（对角）	8.7%	41.2	28.9
LibOrtho v2.0（K-FAC+OGD）	**0.8%**	**44.8**	**31.5**

表 1: LLaMA-7B上的记忆移除有效性和能力保持。

LibOrtho v2.0实现了95%+的记忆移除（金丝雀率从98.5%降至0.8%），同时保持了98%+的泛化（MMLU: 44.8 vs. 45.2，GSM8K: 31.5 vs. 32.1）。K-FAC方法显著优于对角Hessian近似。

5.4 几何分析

我们使用多个指标验证几何分离：

线性模式连通性（LMC）：遵循 [**?**]，我们在W_{base}和$W_{clean} = W_{base} + \Delta W_{gen}$之间沿路径$W(t) = (1-t)W_{base} + tW_{clean}$插值。损失障碍$\max_t L(W(t)) - \min(L(W_{base}), L(W_{clean})) < 0.01$，确认两个模型位于同一损失盆地中。

纠缠比：我们计算$\eta = \frac{|\langle \Delta W_{gen}, \Delta W_{mem}\rangle|}{|\Delta W_{gen}||\Delta W_{mem}|}$。对于基线微调，$\eta = 0.23$（显著纠缠）。使用OGD，$\eta = 0.02$（接近完美正交）。

Hessian谱密度：我们分别计算ΔW_{gen}和ΔW_{mem}的Hessian特征值。ΔW_{mem}表现出重尾分布，具有离群值（$\lambda > 10^3$），而ΔW_{gen}显示Marchenko-Pastur主体分布（$\lambda < 10^2$），验证了基于曲率的分离。

TOFU金丝雀提取：遵循 [**?**]，我们测量模型重现特定金丝雀短语的能力。基线达到98.5%的提取率。移除后，这降至0.8%，证明了有效的记忆擦除。

6 讨论与局限性

6.1 理论洞察

我们的结果支持泛化与记忆可以分离的几何观点，但需要适当考虑：(1) 参数结构（K-FAC），(2) 动态约束（OGD），和(3) 基于秩的语义（SVD）。对角近似的失败突出了捕捉参数交互的重要性。

6.2 局限性

我们的方法假设S_{gen}可以从公共数据中识别。在实践中，如果私有和公共分布显著不同，投影可能不是最优的。此外，K-FAC假设Kronecker结构，这可能不适用于所有层类型。

基于SVD的分离依赖于能力的低秩假设。对于需要高秩更新的任务（例如，学习许多不同事实），我们的方法可能效果较差。

6.3 未来方向

未来的工作应该探索：(1) 损失景观的黎曼几何（超越线性子空间），(2) 用于语义级分离的因果追踪集成，和(3) 基于任务复杂度的自适应k选择。

7 结论

我们提出了LibOrtho v2.0，一个用于在LLM中分离泛化与记忆的几何原理框架。通过结合K-FAC结构化曲率、动态OGD和任务向量算术，我们实现了接近完美的记忆移除，且能力损失最小。我们的工作表明，机制性纠缠可以通过适当的架构和算法设计来克服，为隐私保护的LLM部署开辟了道路。

致谢

请勿在提交中包含任何可能使您身份暴露的致谢（例如，由于您承认的特定机构或资助）

Algorithm 1 LibOrtho v2.0: 几何分离流程

Require: 基础模型W_{base}，公共数据$\mathcal{D}_{pub}$，私有数据$\mathcal{D}_{priv}$

Ensure: 清理后的模型$W_{clean} = W_{base} + \Delta W_{gen}$

阶段1: 子空间识别
for 每一层ℓ **do**
  初始化$A_\ell = 0$，$G_\ell = 0$
  for 每个批次$(x, y) \in \mathcal{D}_{pub}$ **do**
    前向传播：$a = \text{activations}(x)$
    反向传播：$\nabla_{pre} = \text{gradients}(y)$
    更新：$A_\ell \leftarrow 0.95A_\ell + 0.05(aa^T)$
    更新：$G_\ell \leftarrow 0.95G_\ell + 0.05(\nabla_{pre}\nabla_{pre}^T)$
  end for
  特征分解：$A_\ell = U_A\Sigma_A U_A^T$，$G_\ell = U_G\Sigma_G U_G^T$
  选择top-k最平坦方向：$S_{gen}^\ell = \text{span}(\{u_A \otimes u_G : \lambda < \tau\})$
end for
阶段2: 约束微调
初始化$W = W_{base}$
for 每个批次$(x, y) \in \mathcal{D}_{priv}$ **do**
  计算梯度：$g = \nabla L(W; x, y)$
  for 每一层ℓ **do**
    投影：$g_\ell \leftarrow g_\ell - P_{S_{gen}^\ell}(g_\ell)$
  end for
  更新：$W \leftarrow W - \alpha g$
end for
$W_{ft} = W$
阶段3: 事后精炼
for 每一层ℓ **do**
  计算增量：$\tau_\ell = W_{ft}^\ell - W_{base}^\ell$
  SVD：$\tau_\ell = U\Sigma V^T$
  选择k：$\sum_{i=1}^k \sigma_i^2 / \sum_{i=1}^r \sigma_i^2 \geq 0.9$
  $\Delta W_{gen}^\ell = \sum_{i=1}^k \sigma_i u_i v_i^T$
end for
return $W_{clean} = W_{base} + \Delta W_{gen}$

伦理考量

这项工作通过实现选择性移除记忆的私有数据来解决LLM部署中的隐私问题。然而，我们的方法可能被误用于移除安全护栏或审查特定信息。我们强调，记忆移除应仅应用于用户提供的私有数据，而不是安全训练或公共知识。

我们承认，由于信息论限制，完美的记忆移除在理论上是不可能的，我们95%+ 的有效性可能仍会留下残留痕迹。用户不应仅依赖技术解决方案，还应采用法律和政策框架来保护隐私。

所有实验均在公开可用的模型和合成金丝雀数据上进行。我们的评估中未使用真实的私有用户数据。

开放科学

为了促进可重现性，我们提供以下工件：

- **代码**：LibOrtho v2.0的实现，包括K-FAC计算、OGD训练循环和基于SVD的分离。可在以下位置获取：[匿名仓库URL]
- **模型**：LLaMA-7B和GPT-2的预计算S_{gen}子空间，以及清理后的模型检查点。可在以下位置获取：[匿名模型中心URL]
- **数据集**：TOFU金丝雀样本和评估脚本。可在以下位置获取：[匿名数据集URL]
- **实验**：完整的实验配置文件和分析笔记本。可在以下位置获取：[匿名仓库URL]

所有工件将在接受后公开提供。审稿人可以通过提交系统中提供的匿名链接访问它们。