arXiv:2611.32174v1 [cs.LG] 24 Nov 2026

联邦学习中梯度反演防御的统一控制框架

预印本

Lianhai Ren
数学系 新加坡国立大学 10 Lower Kent Ridge Road，新加坡119076
lianhairen@u.nus.edu

Chi Zhang
数学系 新加坡国立大学 10 Lower Kent Ridge Road，新加坡119076
czhang24@nus.edu.sg

2026年11月25日

摘要

梯度扰动被广泛认为是抵御联邦学习中梯度反演攻击的有效防御手段，但众所周知，它们同样会损害优化效果，从而在隐私与效用之间形成持续存在的矛盾。本文表明，现有的梯度扰动防御尽管形式各异，却可以统一于受控动力学框架之下，从而对其优化行为形成一致理解，并推导出无需依赖特定案例分析的通用收敛与安全条件。基于这一观察，我们提出“近正交梯度投影（NODE）”，仅沿与原始梯度正交的方向注入扰动，使防御梯度与真实梯度保持固定的余弦相似度。通过保留一阶下降方向，NODE 在维持对梯度反演攻击有效防护的同时，大幅降低了强扰动带来的优化代价。我们在标准优化假设下从理论上建立了收敛保证，并进一步表明，在梯度分布满足温和条件时，NODE 仍能有效防止反演攻击。在包括 CIFAR-10、ImageNet 和胸部X光在内的多个基准数据集上，结合多种梯度反演攻击开展的全面实验表明，在广泛的训练设置下，该框架相较于现有防御实现了更优的隐私—效用权衡。

1 引言

联邦学习中的梯度交换天然会使训练过程暴露于隐私泄露风险之下，使攻击者能够从共享梯度中恢复敏感信息。梯度反演攻击 [1–3],有力地证明了这一漏洞：它们能够根据传输梯度，以极高的保真度重建训练样本。这些发现从根本上挑战了这样一种传统假设：联邦学习仅通过将原始训练数据保留在本地设备上，就能保护数据仅仅通过将原始训练数据保留在本地设备上来保护隐私。

为缓解这一隐私风险，研究人员提出了多种防御机制，包括梯度剪枝 [4],梯度修改 [5], 梯度扰动 [6], 和差分隐私 [7, 8]。尽管这些方法的具体形式各不相同，但它们具有共同目标：在保留联邦学习效用的同时防止梯度反演。现有大多数方法通过有意修改或限制传输梯度中包含的信息来实现这一目标，力求在隐私保护与优化性能之间取得有效平衡。

然而，对这些方法的分析大多仍然零散，各种方法也都是单独评估的。在本研究中，我们从受控动力学的视角出发，证明这些算法可以纳入统一的理论框架。具体而言，通过将梯度修改建模为广义控制函数的具体实例，我们表明许多现有防御方法都自然地符合这一总体结构。这种统一不仅能让我们更深入地理解其基本原理，还能系统地评估其有效性，无需再进行临时性的、针对特定方法的分析。我们对这些算法进行了全面的收敛性与安全性分析，

并展示了设计选择如何影响原始动力学的收敛行为，同时评估这些方法是否为防范潜在攻击提供了理论保证。

在这一统一框架的基础上，我们提出了“近正交梯度投影”（NODE）方法，这是一种能够以更精细的方式向梯度中注入噪声的防御策略。与引入各向同性噪声的传统差分隐私随机梯度下降（DPSGD）方法不同，NODE 有策略地仅在与真实梯度正交的子空间中注入噪声。这种设计确保一阶下降方向保持不变，同时维持真实梯度与防御梯度之间固定的余弦相似度。因此，即使梯度幅值减小，也能确保模型收敛不受影响。此外，在对梯度分布施加温和假设的条件下，我们从理论上证明了 NODE 仍能有效抵御梯度反演攻击，在不影响收敛的情况下，有效保护敏感训练数据。

在多个数据集上开展的大量实验进一步证明了 NODE 的有效性，表明其在保持具有竞争力的测试准确率的同时，能够显著降低包括 GI [1], IG [2], iDLG [9], CI-Net [3], 和 DGGI [10] 在内的多种梯度反演攻击的成功率。

我们的贡献总结如下：

- 我们的工作建立了一个统一的受控动力学框架，涵盖了广泛的现有防御算法，并从统一的理论视角揭示了它们之间的内在联系。这一视角能够系统地分析收敛性质与隐私保证，为评估图像和成像应用中的不同防御策略提供了具有理论依据的方法。
- 我们提出了 NODE，这是一种新的防御机制，采用正交噪声注入，使其与梯度保持近似正交的对齐关系，从而保留核心下降方向，同时增强图像模型抵御梯度反演攻击的能力。
- 我们在标准凸性条件和固定学习率下对 NODE 进行了理论分析，证明其能够确保稳定收敛，同时为抵御反演算法提供强有力的防护。标准图像处理和医学成像基准上的实证结果进一步支持了这一理论基础，展示了其在实际成像场景中的有效性。

2 文献综述

2.1 梯度反演攻击

联邦学习（FL） [11, 12] 已成为分布式模型训练中广泛采用的一种范式，尤其适用于现代过参数化大型模型 [13], 。它能够在将原始数据保留于本地设备的同时实现协作优化。然而，仅靠这种数据本地化机制并不能确保隐私，因为客户端与服务器之间交换的梯度可能包含有关底层训练样本的大量信息。梯度反演攻击 [1] 利用共享梯度重建私有数据，揭示了这种漏洞，也表明模型更新可能成为非预期信息泄露的有效途径。早期方法通过联合优化虚拟输入和标签，使其匹配观测到的梯度，从而证明了恢复训练样本的可行性 [1, 9]。后续方法通过引入先验信息提升了重建保真度，包括自然图像统计信息 [2, 14] 以及目标模型的架构特征。与此同时，解析研究探讨了能够显式反演梯度的理论条件，从而加深了对其潜在泄露机制的理解 [15, 16]。近来，生成模型（如生成对抗网络和扩散模型） [10, 17, 18], 通过引入强大的数据先验 [3, 19, 20], 进一步提升了重建质量与可扩展性。除原始的单样本重建场景外，梯度反演攻击也拓展到了更具挑战性的场景，包括大批量恢复 [21], 、多个通信轮次中累积的隐私泄露 [22], , 以及语言模型等跨领域应用；在这些应用中，梯度可能泄露敏感的文本信息 [23, 24]。

2.2 反演防御

防御策略可分为几类。多方计算（MPC）和同态加密（HE）等密码学解决方案通过避免明文梯度暴露，提供了强有力的理论保护，但会带来较大的通信和计算开销 [25–29]。安全聚合协议通过在不泄露个体梯度的情况下实现聚合，提供了一种切实可行的折中方案 [30–33]。差分隐私（DP）和差分隐私随机梯度下降会引入校准噪声，以掩盖个体贡献，并且更适合大规模联邦学习部署 [6–8, 34–36], 不过，近期研究已探索自适应噪声机制，以改善效用-隐私权衡 [37, 38]。互补方法包括剪枝或稀疏化，以移除信息性梯度分量 [4, 39], 梯度压缩技术可减少信息泄露 [40, 41],以及自适应掩码

或表示变换（例如 Soteria），旨在隐藏可重建特征，同时保留效用 [5, 42]。基于 mixup 的防御等模型扰动方法 [43] 以及输入预处理技术 [44] 也已被提出，以增加重建难度。尽管机制多种多样，这些防御措施都以转化梯度信息、阻碍反演，同时保留图像或视频任务的模型训练为核心目标，但其中许多方法在防御自适应攻击方面仍面临挑战 [45, 46]。

2.3 神经网络训练的控制视角

通过动力系统与控制的视角审视优化和深度网络，可以获得一种统一的观点。残差网络和连续深度模型可被解释为离散时间动力系统，其中参数或梯度修改充当控制输入，引导模型状态的演化 [47–49]。控制与反馈理论为深度学习中的架构设计、稳定性分析和优化方法提供了指导 [50–56]。基于这一观点，我们将广泛的梯度修改防御方法归类为受控动力学实例，从而能够系统地比较其收敛性和隐私属性，并消除对零散的、特定于方法的证明的需求。

3 梯度反演攻击防御的统一框架

本节首先回顾联邦学习中的梯度反演攻击，强调联邦学习范式固有的隐私挑战。随后我们将展示，许多现有的防御方法，如差分隐私、剪枝和 Soteria，都可以被解释为作用于梯度（以及可选地作用于输入数据）的函数。这一观察促成了我们的统一框架，该框架将梯度防御操作视为训练动态中的控制器。

3.1 联邦学习中的梯度反演攻击

在传统的联邦学习系统中，本地训练数据及其相关标签通常无法被外部实体访问，例如好奇服务器或第三方。相反，这些实体可能只能获取上传的本地梯度。梯度反演攻击 [1] 旨在利用模型权重、损失函数与梯度之间的关系，从这些梯度中重构原始数据。

从数学上看，考虑一个参数为 w 的模型及其损失函数 $L(w,\ x)$，其中 x 表示输入数据。在梯度反演攻击中，攻击者试图通过求解一个逆问题来恢复 x。给定梯度 g，攻击者旨在近似原始输入 x，这可以表述为一个优化问题：

$$\hat{x} = \arg\min_{x'} \|\nabla_w L(w, x') - g\|^2.$$

攻击者通常会采用正则化 [2] 或先验 [19], 等高级技术，以提高重构数据的质量，并减少反演过程中固有的不确定性。近期研究 [3] 已证明，即使批大小相对较大，此类攻击仍然可行，这凸显了当前联邦学习训练范式易遭受隐私泄露的脆弱性。

3.2 通过受控动力学进行防御的统一框架

为降低梯度反演攻击的风险，人们提出了各种防御机制来加强通信安全。常见方法包括 DPSGD [6], ，它利用差分隐私保护数据；剪枝技术 [4] ，它有选择地消除大量梯度；以及 Soteria [5], ，它旨在生成经过扰动的数据表示，以保护敏感信息。

尽管这些方法表面上存在差异，我们证明，从控制理论的角度来看，它们可以纳入一个统一框架。形式上，令 $x_t \in \mathbb{R}^n, w_t \in \mathbb{R}^m$ 表示第 t 次迭代时的输入和权重， g_t 表示相应的梯度， $C(g, x)$: $\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}^m$ 表示防御方法应用于梯度的控制函数。我们可以将一般的防御操作表示为训练中的受控动力学：

$$\frac{dw_t}{dt} = C(g_t, x_t) := g_t'. \tag{1}$$

在控制动力学框架下，许多现有的防御方法都可以视为选择特定控制函数 C的具体实例：

• **差分隐私随机梯度下降**将该变换实现为

$$C_{\text{DP}}(g, x) = g + \xi, \tag{2}$$

其中， $\xi \in \mathbb{R}^m$ 表示从预定义分布中采样的噪声，通常为高斯分布或泊松分布。为简单起见，本文假设DPSGD始终采用高斯分布。

• **剪枝** 可被解释为一种根据梯度的幅值或重要性将其某些分量置零或衰减的函数：

$$C_{\text{prune}}(g, x; k) = g \odot \mathbb{I}\{|g_i| \geq \text{Top}_k(|g_i|)\}, \tag{3}$$

其中， $\odot$ 表示逐元素乘法， $\mathbb{I}$ 表示指示函数， $\text{Top}_k(|g_i|)$返回梯度分量中绝对值第 k大的值。

• **Soteria** 利用从输入到输出特征的梯度范数（具体而言，即输入到最后一个全连接（FC）层的梯度范数）生成掩码，然后将其应用于最后一个FC层的原始梯度。这可以表示为：

$$C_{\text{soteria}}(g, x) = g \odot F_{\text{soteria}}(g, x), \tag{4}$$

其中， $F_{\text{soteria}}(g,\ x)$[1] 是一个用于生成掩码的函数。

这种统一的控制视角表明，尽管防御机制各不相同，这些方法都可以视为对梯度信息施加受控修改。此外，这一视角使我们能够在共享的理论框架内评估各种防御策略，从而无需针对每种方法分别进行分析。将不同的防御方法视为不同的控制策略，有助于该框架推广至各种技术，并为开发新的防御方法奠定基础，详见第 4 节。

3.3 防御算法的一般条件

一般而言，有效的防御算法必须实现两个关键目标：在训练动态过程中保持模型性能，同时保护数据免受梯度反演攻击。借助统一框架，我们现在可以提出一个关键问题：控制函数 C 必须具备哪些性质，才能确保满足这两个条件？

3.3.1 收敛条件

我们首先考察防御操作 $C(g, x)$ 的选择如何影响训练算法的收敛性质。具体而言，修改后的训练动态最终性能应与原始动态的最终性能高度接近。形式上，若 w_∞ 和 w_∞^o 分别表示从修改后的动态和原始动态中获得的最终权重，则我们要求：

$$\mathbb{E}[L(w_\infty) - L(w_\infty^o)] \leq \ell_0, \tag{5}$$

其中， $L:\ \mathbb{R}^m \to \mathbb{R}$ 是权重的损失函数，即整个数据集上的平均损失，期望是对初始权重取的。较小的 ℓ_0 表明该防御方法能够有效地维持训练性能。

例如，我们可以分析在具有 l–Lipschitz 连续梯度的 c–强凸损失函数 $L(w)$ 下， ℓ_0 的值，即：

$$L(w') - L(w) \geq \nabla L(w)^\top (w' - w) + \frac{c}{2}\|w - w'\|_2^2, \tag{6}$$

$$\|\nabla L(w) - \nabla L(w')\|_2 \leq l\|w - w'\|_2. \tag{7}$$

经过简单变换，我们可以由式（6）推导出：

$$L^* \geq L(w) - \frac{c}{2}\|\nabla L(w)\|. \tag{8}$$

此外，式（7）给出了以下界：

$$L(w') \leq L(w) + \nabla L(w)^\top (w' - w) + \frac{l}{2}\|w' - w\|_2^2. \tag{9}$$

为简单起见，本文将重点讨论梯度下降算法。由于防御方法通常会在梯度上添加随机噪声，或有选择地将某些分量置零，因此在下面的假设中，我们需要对防御梯度g' 施加一般的一阶和二阶矩限制。

[1]对于 d 关于 F_{soteria}的详细定义，请参见第 4 节 的 [5]。

假设 3.1。（*[57]*中的假设 *4.3*）假设防御梯度 g' 满足以下一阶和二阶矩约束，$\exists\,\mu>0$、$\mu_G>0$、$M\geq 0$、$M_V\geq 0$，使得：

$$\nabla L(w)^\top \mathbb{E}[g'] \geq \mu\|\nabla L(w)\|_2^2, \tag{10}$$

$$\|\mathbb{E}[g']\|_2 \leq \mu_G\|\nabla L(w)\|_2, \tag{11}$$

$$Var(g') \leq M + M_V\|\nabla L(w)\|_2^2. \tag{12}$$

利用式（8）–（12）中的上述假设，可得如下收敛定理。

定理 3.2. 假设损失函数是 c-强凸的，且具有 *l-Lipschitz* 连续梯度，并且防御梯度 g' 一致满足矩约束 (10)-(12)，其中 w。若学习率固定在范围$0<\eta\leq \frac{\mu}{lM_G}$内，其中 $M_G=M_V+\mu_G^2$，则有如下收敛保证：

$$\begin{aligned}\mathbb{E}[L(w^k)-L^*] &\leq \frac{\eta l M}{2c\mu} + (1-\eta c\mu)^k\left(L(w^0)-L^*-\frac{\eta l M}{2c\mu}\right)\\ &\xrightarrow{k\to\infty} \frac{\eta l M}{2c\mu},\end{aligned} \tag{13}$$

其中，w^k 表示经过 k 步后的权重，期望是以初始权重 w^0为条件计算的。

与梯度下降类似，我们还可以将结果进一步推广到非凸情形：

定理 3.3. 假设 与定理 *3.2* 中相同的条件，但不要求损失函数满足凸性条件，在以下收敛保证成立：

$$\begin{aligned}\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K}\|\nabla L(w^k)\|_2^2\right] &\leq \frac{\eta l M}{2c\mu} + \frac{(L(w^0)-L^*)}{Kc\mu}\\ &\xrightarrow{k\to\infty} \frac{\eta l M}{2c\mu},\end{aligned} \tag{14}$$

注 3.4。上述两个定理表明，如果受保护梯度的方差上界 g'在公式 (12) 中包含常数因子 M，那么在无限时间内我们会得到常数因子 $\frac{\eta l M}{2c\mu}$。此外，当真实梯度较小时，该因子将主导防御梯度，从而导致最终损失与最优损失之间存在非零差距。

注 3.5。请注意，在假设 *3.1* 中，要求防御梯度 g' 与真实梯度之间的角度小于某个值且远离 $90°$。我们已经证明，对于凸损失，使用垂直的防御梯度不会降低该值。[2]

3.3.2 安全条件

现在我们将注意力转向有效防御算法必须满足的安全条件。具体而言，公式（1）中的控制函数必须充分混淆敏感梯度信息，以防御梯度反演攻击。这意味着 $C(g,x)$ 应以受控方式扭曲梯度，使攻击者无法轻易恢复底层数据 x。形式化地，令 G_w^{-1} 表示权重 w[3] 下输入梯度函数的逆映射。那么，对于数据域 $D\subseteq\mathbb{R}^n$ 中的所有 x（例如私有数据集或有界集合），我们要求：

$$\min_{x'\in G_w^{-1}(C'(g,x))}\|x'-x\|_2 \geq \ell_1, \forall x\in D, \forall w\in W\backslash W_\epsilon. \tag{15}$$

其中，$W\subseteq\mathbb{R}^m$ 是参数集，$W_\epsilon=\{w|\|\nabla L(w,x)\|_2\leq\epsilon,\forall x\in D\}$ 表示鞍点附近的集合，此处梯度很小，因此在 W_ϵ [2] 中几乎不可能发起攻击。较大的 ℓ_1 意味着更强的混淆效果，从而带来更好的防御性能。

对 ℓ_1 进行全面分析通常并非易事，而且往往取决于梯度反演攻击的具体形式。不过，我们可以证明，它可以由输入梯度映射的利普希茨常数界定：

[2]在 [58], 中，作者提出了垂直防御梯度。我们注意到其防御函数存在一个错误，导致防御梯度与真实梯度呈现正的余弦相似度。修正这一问题后，我们发现其训练过程可能不再成功。[3]这里我们假设输入梯度函数 G_w 是满射。

定理 3.6. 如果输入梯度函数 G_w： $x \to \nabla L$（w、 x）对所有 $w \in W$ 都是 l_g 利普希茨连续的，则有：

$$\ell_1 \geq \frac{1}{l_g} \min_{x \in D} \|C(g,x) - g\|_2. \tag{16}$$

总而言之，能够实现较小的 ℓ_0（确保训练性能的最低程度下降）和较大的 ℓ_1（确保稳健的隐私保护）的梯度防御方法，被认为是有效的。这一视角统一了对现有防御方法的分析，并为在联邦学习系统中设计更新、更稳健的防御方法提供了明确指导。

3.4 现有防御方法能否满足这些条件?

通过建立统一框架以及相应的收敛和安全条件，我们现在可以评估现有防御算法是否遵循这些标准。

我们首先分析 DPSGD 算法的收敛和安全条件，具体见以下定理 s.

定理 3.7 (ℓ_0 差分隐私随机梯度下降的界)。 在假设 (8)–(12) 下，对于噪声尺度为 σ 且采用 l_2-范数裁剪阈值 a 的差分隐私随机梯度下降，我们有：

$$\ell_{0,DPSGD} \leq \frac{\eta l a^2 \sigma^2}{2c}. \tag{17}$$

注 3.8。 定理 *3.7* 表明，通常在比较差分隐私随机梯度下降相对于原始动力学的性能时，我们只能得到一个常数因子。因此，使用 $\sigma > 0$ 无法收敛到原始动力学。这一差距源于梯度中添加的噪声。此外，较大的噪声尺度 σ 和 l_2-范数裁剪阈值会导致模型性能变差，这与直觉预期和以往的实验发现一致。

就安全条件而言，对于差分隐私随机梯度下降，我们有 $\min \|C(g,\ x)\|_2 = \min \|\xi\|_2 = 0$。然而，我们可以以 $1-2\delta$ 的置信度得到 ℓ_1 的下界：

定理 3.9(ℓ_1 差分隐私随机梯度下降的界)。 在上述假设下，对于噪声尺度为 σ 且采用 l_2-范数裁剪阈值 a 的差分隐私随机梯度下降，以 $1-2\delta$ 的概率，我们有：

$$\ell_{1,DPSGD} \geq \frac{a\sigma(\sqrt{m} - a_0\sqrt{\log \delta^{-1}})}{l_g}. \tag{18}$$

这里 a_0 是一个绝对常数。

总之，定理 3.9 表明，当噪声尺度 σ 相对较大时，差分隐私随机梯度下降可以被视为一种安全防御方法。然而，这需要付出违反收敛条件的代价：如定理 3.7 所强调的，受控动力学与原始动力学相比可能存在性能差距。在实践中，有必要谨慎选择 σ，以平衡收敛性与安全性之间的权衡。

类似地，对于剪枝方法，我们可以建立以下两个定理。

定理 3.10 (ℓ_0 关于剪枝。 在假设 (8)、(9)成立的条件下，对于剪枝率为 $p < 1$ 的剪枝方法，我们有：

$$\ell_{0,Pruning} = 0. \tag{19}$$

注 3.11。 上述定理表明，在适当的假设下，剪枝方法可以实现与原始优化动力学相同的最终损失，从而满足收敛条件。这一结果是意料之中的，因为即使在仅保留单个元素的极端情形下，剪枝也会退化为坐标下降 *[59],*，而众所周知，坐标下降对于凸函数 *[60, 61]*能够收敛。

为分析剪枝方法的 ℓ_1，我们假设

$$g \sim N(0, \sigma_g^2 I_m). \tag{20}$$

接下来，我们考虑对随机变量 X 进行操作 a 截断；若满足 $|X| > a$，则得到 $\hat{}X = 0$，否则得到 $\hat{}X = X$。于是有如下定理。

定理 3.12 (ℓ_1 阈值剪枝的界。 在假设 (8)、(9) 和 (20) 下，对于剪枝阈值为 a 的阈值剪枝方法，以 $1-2\delta$ 的概率，有：

$$\ell_{1,Threshold\ Pruning} \geq \frac{\sigma_g\sqrt{V(a)}(\sqrt{m} - a_1\sqrt{\log \delta^{-1}})}{l_g}. \tag{21}$$

这里 $V(a)$ 是 a截断标准正态随机变量的方差和 a_1 是一个绝对常数。

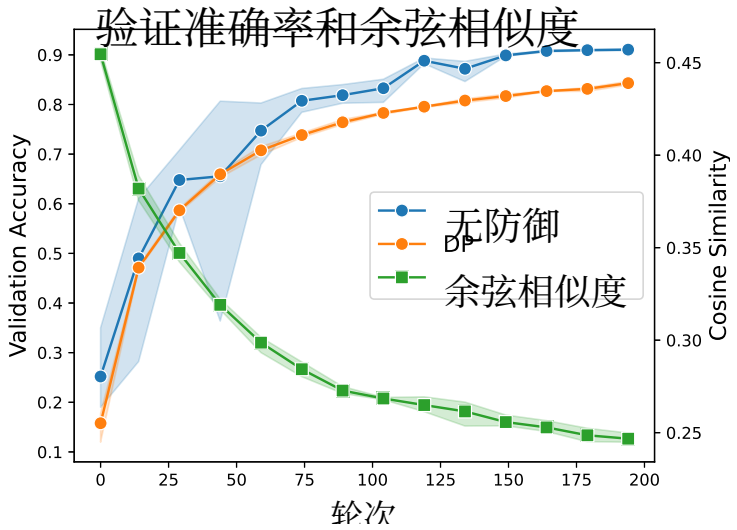

图1：训练期间，随着真实梯度变小（绿色曲线），差分隐私随机梯度下降方法会使梯度偏离得更多，与原始动态相比，导致性能下降。**左轴：**准确率；**右轴：**余弦相似度。

备注 3.13。 通常很难直接分析剪枝方法的 ℓ_1，但借助a截断随机变量的独立性，可以更容易地分析带阈值的剪枝方法。如 $m \to \infty$所示，剪枝的频率（$|X_i| \leq a$）几乎必然以 $O(m^{-\frac{1}{2}})$的速度收敛于剪枝率 p，据此我们可以从期望速率 p推断出a： $a = \Phi^{-1}(\frac{1+p}{2})$。其中， Φ 是标准正态分布的累积分布函数。如 $p \to 0$所示，ℓ_1 范数可能变得相对较小，从而增加攻击者实施本地数据泄露的风险。

总而言之，尽管满足定理 3.10 中的收敛条件，剪枝方法却无法保证定理 3.12 中的安全要求，因为它没有建立足够大的 ℓ_1 损失。

4 具备收敛与安全保证的近正交梯度投影

以往算法违反收敛条件或安全条件中的任一条件，凸显了同时满足二者的防御机制的必要性。因此，本节介绍我们提出的方法——近正交梯度投影（NODE），并提供理论分析，以保证 NODE 在保持训练性能的同时维持稳健的防御能力。

4.1 差分隐私随机梯度下降的关键观察

在介绍新算法之前，我们首先提出两个关键观察结果，用于解释差分隐私随机梯度下降为何无法满足收敛条件。

观察 1：DPNoise噪声尺度权衡。 在传统的差分隐私随机梯度下降方法中，增大噪声尺度可以通过更好地隐藏敏感信息来提升防御性能。然而，较大的噪声尺度也会扭曲梯度方向，从而导致性能显著下降。当需要强大的防御能力，同时又希望保持较高的训练性能时，这种权衡尤其关键。

观察2：训练后期不稳定性。 从实证结果来看，在真实梯度相对较大的训练早期，差分隐私随机梯度下降方法的表现与原始动力学相似。随着训练推进，真实梯度逐渐减小，而固定的噪声水平可能会占据信号的主导地位。这种不平衡会导致防御梯度与原始梯度之间的余弦相似度随时间下降（图1中的绿色曲线）——也就是说，两者之间的角度逐渐增大——从而导致训练后期出现不稳定性和模型性能下降。

算法 1近正交梯度投影（NODE）

要求： 客户端梯度 $g_t \in \mathbb{R}^d$，目标角度 θ**确保：** 防御梯度 g'_t1: 计算梯度的范数： $G_t \leftarrow \|g_t\|_2$2: 采样一个噪声向量 $\xi_t \sim \mathcal{N}(0, I_d)$3: 将 ξ_t 投影到与 g_t 正交的超平面上： $n'_t \leftarrow \xi_t - \frac{\langle \xi_t, g_t \rangle}{G_t^2} g_t$4: 设置噪声幅度： $r_t \leftarrow \frac{G_t}{\|n'_t\|_2} \tan\theta$5: 计算防御梯度： $g'_t \leftarrow g_t + r_t n'_t$6: **返回** g'_t

4.2 近正交梯度投影（NODE）

上述观察表明，在训练后期阶段，梯度本身已变得较小，此时继续使用早期相同的噪声尺度会高估所需的扰动。因此，应以受控方式注入噪声——在确保稳健防御的同时保留下降方向。为此，需要把握两个基石：首先，注入的噪声必须始终与真实梯度方向正交，以确保对下降路径的干扰最小；其次，防御梯度应与真实梯度保持固定的角度关系，从而防止后期阶段出现不稳定。

具体而言，我们提出了"近正交梯度投影"（NODE）方法，该方法严格沿梯度的正交方向注入噪声，并实施受控裁剪机制，以确保修改后的梯度与原始梯度保持近似正交。对于每次迭代 t，防御梯度 g'_t 通过添加噪声向量 n_t 构造，该噪声向量严格与真实梯度 g_t正交。这确保了至少在一阶近似下，下降方向不会受到影响。此外，通过使噪声与 $\|g_t\|$成比例地缩放，NODE 维持了 g_t 与 g'_t之间恒定的余弦相似度（例如 0.2）。数学上，我们约束：

$$C(g_t) = n_t,$$
$$\text{s.t. } \langle g_t, n_t \rangle = 0, \cos\theta = \frac{\langle g_t, g'_t \rangle}{\|g_t\|_2 \|g'_t\|_2} = \alpha.$$

其中， n_t 在接受区域（即一个超平面与超球面的交集）内均匀分布， α （或等价地，固定角度 θ）是预先确定的常数，用于平衡训练性能与防御鲁棒性。

算法 1 概述了针对给定客户端梯度的 NODE 流程。在该算法中，噪声 n_t 首先通过从标准高斯分布中采样生成，然后投影到与 g_t正交的子空间，最后重新缩放，以确保其幅度与 $\|g_t\|_2$ 成正比（具体而言，为 $\|n_t\|_2 = \tan\theta \|g_t\|_2$ ）。将 n_t 添加到g_t后，NODE 在保持下降方向的同时，确保 g_t 与 g'_t 之间的余弦相似度保持不变。这种受控噪声注入旨在阻止梯度反演攻击，同时避免对训练性能造成显著影响。

4.3 理论分析

现在，我们基于定理 3.2 和动力学视角分析 NODE 的收敛性质，并确立其同时保持训练性能与防御有效性的条件。

4.3.1 收敛条件

我们首先分析 NODE 的收敛性质。具体而言，我们可以利用定理 3.2 得到 NODE 在强凸且一阶利普希茨连续损失下的收敛结果：

定理 4.1 (ℓ_0 NODE 的界)。 在定理 *3.2* 的假设下，*NODE* 具有 $\ell_{0,\ NODE} = 0$。

注 4.2。 与剪枝方法类似，*NODE* 方法可以达到与原始动力学相同的最终损失，从而满足收敛条件。

此外，我们将在以下定理中表明，在具有李雅普诺夫函数 [62] 的连续训练动力学中，可以移除假设 10–12，并得到 NODE 的相同收敛结果。

定理 4.3。 $(L(w) - L^*)^2$ 是 *NODE* 方法的李雅普诺夫函数，因此它将渐近达到最小值 0，也就是说，*NODE* 方法将渐近收敛到最小值。

4.3.2 安全条件

由于噪声尺度与梯度尺度成正比，而梯度尺度在训练过程中会不断减小，因此我们需要建立一些温和的附加条件，以确保 NODE 保持防御有效性。

假设 4.4。 假设数据集 D 中每个数据 x 的梯度范数均受数据集平均梯度范数的约束，即 $\exists b > 0$，使得 $\forall x \in D$ 我们有：

$$\|\nabla L(w; x)\|_2 \leq b\|\nabla L(w)\|_2. \tag{22}$$

该假设限制了 训练最终阶段的梯度尺度，此时整体梯度较小。根据这一 s
假设，我们可以推 表明 NODE 与差分隐私随机梯度下降具有相同的差分隐私预算。
定理 4.5。在 假设 *4.4* 下，具有角度 θ 的 *NODE* 与具有
$\sigma = \frac{\tan\theta}{b}$.

备注 4.6。 定理 *4.5* 表明，*NODE* 方法在理论上展现出与差分隐私随机梯度下降方法相似的防御能力。因此，可以选择 θ 的适当值，以确保 ℓ_1 保持足够大，从而满足安全条件。

总之，我们的分析表明，在假设 (8)–(9) 下，NODE 方法能够重塑梯度信息以保护隐私，同时不会削弱学习性能，从而同时满足收敛和安全条件。

5 实验

本节在多个数据集上评估我们的方法，衡量其抵御梯度反演攻击的鲁棒性。

5.1 实验设置

我们的实验设置遵循先前梯度反演工作的常见做法 [1, 3]。我们使用带有 Sigmoid 激活函数的 ResNet-18 模型，已有研究表明这会提高模型对反演攻击的易感性。我们比较以下防御方法：（1）无防御；（2）采用不同噪声尺度的差分隐私随机梯度下降 [6] ；（3）不同稀疏度水平的剪枝 [4] ；（4） Soteria [5]；（5） Outpost [63]；（6）双重梯度剪枝（ DGP ） [64]；以及（7）梯度随机失活 [65]；（8）我们提出的 NODE。为评估防御有效性，我们考虑五种具有代表性的梯度反演攻击：梯度反演（GI） [1], 、改进梯度反演（IG） [2], 、iDLG [9], 、CI-Net [3], 和 DGGI [10]。我们报告模型训练性能以及使用结构相似性指数（SSIM）和特征相似性指数（FSIM）衡量的隐私泄露程度。详细设置见附录 B。

5.2 CIFAR-10 结果

我们在 CIFAR-10 数据集上开始评估防御性能 [66], ，结果见表 1。基线模型取得了 91.21% 的高测试准确率。然而，在反演攻击下，其结构相似性指数（SSIM）值较高，表明恢复图像与原始数据仍过于相似。相比之下，剪枝（p=0.9）和差分隐私（σ=4e-3）等防御方法能够取得更低的 SSIM 值，从而提供更好的隐私保护。但当剪枝去除较大比例的权重或提高差分隐私随机梯度下降中的噪声水平时，通常会以降低测试准确率为代价。值得注意的是，余弦相似度为 0.2 的 NODE 仍能保持具有竞争力的测试准确率（90.30%），同时大幅降低 SSIM 值（例如，在梯度反演下为 0.0256），展现出其混淆梯度的有效性。进一步设置 $\alpha = 0.1$ 可降低 SSIM 值，同时仅导致准确率小幅下降 0.99%。总体而言，NODE 在保持性能与提升安全性之间实现了良好的平衡。

图 2 展示了 CIFAR-10 在不同防御方法下的恢复图像网格。每一列对应一种特定的防御方法，每一行代表一种特定的反演攻击。在 CI-Net 攻击下，基线重建结果保留了大量细节；CI-Net 是一种能够处理大批量大小的生成式方法。值得注意的是，在没有防御的情况下，它可以完全恢复原始训练图像，凸显了梯度反演攻击带来的数据泄露风险。具体而言，噪声尺度为 1e-3 的差分隐私表现出较差的防御性能，而 Outpost 和 DGP 的表现更好。NODE 的重建结果出现了显著退化，表明其能够有效阻碍对抗性恢复，同时保持较高的模型准确率。

5.3 ChestX-ray 结果

我们在胸部X光数据集 [67]上评估了我们的方法；结果汇总于表2。未保护基线的测试准确率为 91.19%，但其预测容易受到反演攻击的重建。以往的防御

表 1：CIFAR-10 上的性能：采用多种防御方法时，四种梯度反演攻击（GI、IG、iDLG、CI-Net 和 DGGI）下的最终测试准确率和平均结构相似性指数。最佳性能以粗体标出，次佳性能以下划线标出。

防御方法	性能	攻击（结构相似性指数）					
	测试准确率 ↑	GI ↓	IG ↓	iDLG ↓	CI-Net ↓	DGGI ↓	Mean
None	**91.21**$_{\pm 0.17}$	0.3751$_{\pm 0.0929}$	0.4701$_{\pm 0.0948}$	0.3974$_{\pm 0.0755}$	0.9866$_{\pm 0.0081}$	0.5017$_{\pm 0.1084}$	0.5462
剪枝（p=0.9）	89.66$_{\pm 0.17}$	0.0317$_{\pm 0.0103}$	0.0367$_{\pm 0.0208}$	0.0222$_{\pm 0.0083}$	0.8513$_{\pm 0.0814}$	0.1577$_{\pm 0.1040}$	0.2199
差分隐私随机梯度下降（σ=1e-3）	87.09$_{\pm 0.28}$	0.0958$_{\pm 0.0484}$	0.4211$_{\pm 0.1235}$	0.3492$_{\pm 0.0752}$	0.9437$_{\pm 0.0196}$	0.4655$_{\pm 0.1038}$	0.4551
差分隐私随机梯度下降 (σ=4e-3)	85.06$_{\pm 0.26}$	0.1284$_{\pm 0.0308}$	0.0462$_{\pm 0.0249}$	0.1446$_{\pm 0.0392}$	0.2933$_{\pm 0.0989}$	0.2576$_{\pm 0.1224}$	0.1740
Soteria	89.20$_{\pm 0.15}$	0.1143$_{\pm 0.0506}$	0.1886$_{\pm 0.0225}$	0.2651$_{\pm 0.0064}$	0.4122$_{\pm 0.1203}$	0.2675$_{\pm 0.0817}$	0.2495
Outpost	61.68$_{\pm 1.35}$	**0.0241**$_{\pm 0.0089}$	0.0389$_{\pm 0.0298}$	0.0311$_{\pm 0.0130}$	0.9402$_{\pm 0.0358}$	0.1657$_{\pm 0.1149}$	0.2400
DGP	89.50$_{\pm 0.12}$	0.0249$_{\pm 0.0051}$	**0.0255**$_{\pm 0.0073}$	**0.0160**$_{\pm 0.0073}$	0.8080$_{\pm 0.1103}$	**0.1515**$_{\pm 0.1074}$	0.2052
梯度丢弃（p=0.3）	87.84$_{\pm 0.41}$	0.0671$_{\pm 0.0146}$	0.1705$_{\pm 0.0533}$	0.1588$_{\pm 0.0495}$	0.1645$_{\pm 0.0678}$	0.2204$_{\pm 0.0296}$	0.1563
NODE (α = 0.2)	90.30$_{\pm 0.14}$	0.0256$_{\pm 0.0096}$	0.0305$_{\pm 0.0169}$	0.0190$_{\pm 0.0048}$	0.2422$_{\pm 0.0770}$	0.1611$_{\pm 0.1099}$	0.0957
NODE（α = 0.1）	89.31$_{\pm 0.02}$	0.0272$_{\pm 0.0086}$	0.0302$_{\pm 0.0120}$	0.0183$_{\pm 0.0058}$	**0.0790**$_{\pm 0.0265}$	0.1561$_{\pm 0.1080}$	**0.0622**

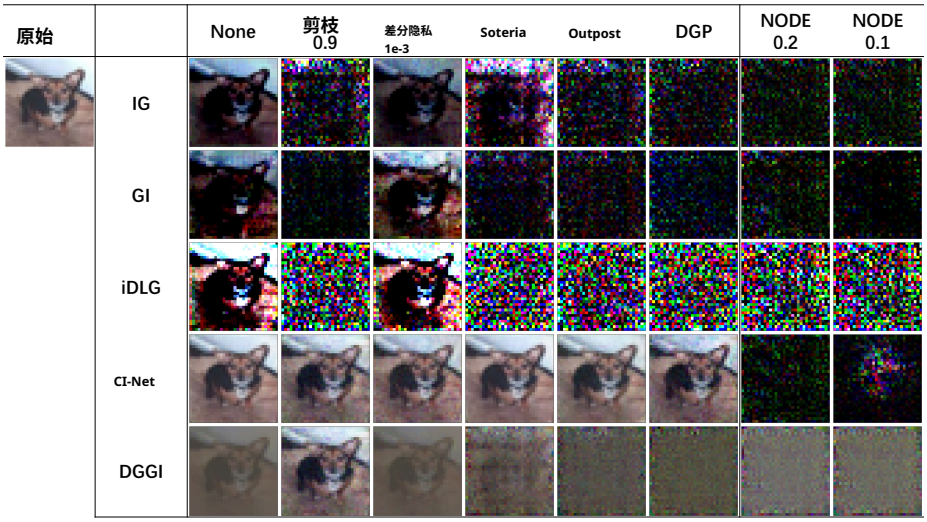

图 2：在 CIFAR-10 上，不同攻击与防御方法组合下原始图像及其重建结果的对比。行表示所采用的攻击，列展示相应的防御策略。

方法在一定程度上降低了可恢复性；然而，它们往往会导致训练不稳定，并显著降低预测性能。相比之下，采用 $\alpha = 0.2$ 的 NODE 实现了良好的平衡，测试准确率达到 89.79%，同时显著降低了各种攻击方法下的结构相似性指数值。将 α 降至 0.1 可大幅抑制重建质量：对于 CI-Net，结构相似性指数从 0.2857 降至 0.0579，不过更强的保护也会使准确率额外下降 0.53%。因此，建议在数据敏感场景中采用 $\alpha = 0.1$ 。

5.4 X-Ray 数据集上的防御可视化

除了以结构相似性指数衡量的定量结果外，图3还对多种防御方法抵御五种具有代表性的梯度反演攻击的效果进行了定性比较。如图所示，差分隐私随机梯度下降未能充分掩盖敏感信息，若干攻击仍能恢复出原始胸部X光图像中可辨认的解剖结构。尽管 Soteria、Outpost 和 DGP 能有效抑制 IG、GI 和 iDLG 下的重建，但面对更先进的攻击时仍易受攻击，尤其是 CI-Net，它能够以相当高的保真度恢复原始解剖结构。相比之下，NODE 在全部五种攻击下均能持续干扰图像重建，生成的图像大多呈噪声状，且不包含清晰可辨的解剖结构。NODE 的优势在抵御 CI-Net 时尤为显著，因为其他所有防御方法都出现了大量视觉信息泄露。这些定性观察进一步印证了结构相似性指数的定量结果，表明 NODE 能有效抵御具有不同重建机制的梯度反演攻击。

表2：ChestX-ray上的性能：在多种防御方法下，针对五种梯度反演攻击（GI、IG、iDLG、CI-Net 和 DGGI）的最终测试准确率和平均结构相似性指数。最佳性能以粗体显示，次佳性能以下划线标示。

防御方法	性能	攻击（结构相似性指数）				
	测试准确率↑	GI ↓	IG ↓	iDLG ↓	CI-Net ↓	DGGI ↓
None	**91.19**$_{\pm 1.16}$	0.1727$_{\pm 0.1407}$	0.3332$_{\pm 0.0578}$	0.2041$_{\pm 0.0059}$	0.9505$_{\pm 0.0268}$	0.7574$_{\pm 0.1686}$
剪枝	86.11$_{\pm 0.60}$	0.0380$_{\pm 0.0334}$	0.0275$_{\pm 0.0069}$	0.0098$_{\pm 0.0023}$	0.3602$_{\pm 0.1014}$	0.1192$_{\pm 0.0184}$
差分隐私随机梯度下降（σ=1e-3）	85.15$_{\pm 0.66}$	0.1611$_{\pm 0.0475}$	0.2961$_{\pm 0.0838}$	0.1587$_{\pm 0.0473}$	0.8022$_{\pm 0.1507}$	0.6201$_{\pm 0.1186}$
差分隐私随机梯度下降（σ=4e-3）	84.99$_{\pm 0.66}$	0.0831$_{\pm 0.0320}$	0.0325$_{\pm 0.0079}$	0.0378$_{\pm 0.0117}$	0.1778$_{\pm 0.0670}$	0.2764$_{\pm 0.0662}$
Soteria	89.01$_{\pm 0.50}$	0.0203$_{\pm 0.0090}$	0.0369$_{\pm 0.0106}$	0.0148$_{\pm 0.0046}$	0.9624$_{\pm 0.0135}$	0.1827$_{\pm 0.0493}$
Outpost	87.90$_{\pm 0.88}$	0.0261$_{\pm 0.0137}$	0.0290$_{\pm 0.0057}$	0.0126$_{\pm 0.0041}$	0.9171$_{\pm 0.0633}$	0.1511$_{\pm 0.0462}$
DGP	89.18$_{\pm 0.88}$	0.0203$_{\pm 0.0151}$	0.0194$_{\pm 0.0025}$	0.0103$_{\pm 0.0035}$	0.8484$_{\pm 0.0731}$	0.1107$_{\pm 0.0527}$
GradientDropOut	86.24$_{\pm 0.44}$	0.0314$_{\pm 0.0181}$	0.0320$_{\pm 0.0014}$	0.0172$_{\pm 0.0028}$	0.0784$_{\pm 0.0345}$	0.1427$_{\pm 0.0568}$
NODE（$\alpha = 0.2$）	89.79$_{\pm 0.33}$	**0.0143**$_{\pm 0.0045}$	0.0131$_{\pm 0.0037}$	0.0095$_{\pm 0.0027}$	0.2857$_{\pm 0.1184}$	0.1137$_{\pm 0.0593}$
NODE（$\alpha = 0.1$）	89.26$_{\pm 0.39}$	0.0144$_{\pm 0.0054}$	**0.0120**$_{\pm 0.0043}$	**0.0085**$_{\pm 0.0035}$	**0.0579**$_{\pm 0.0267}$	**0.1087**$_{\pm 0.0594}$

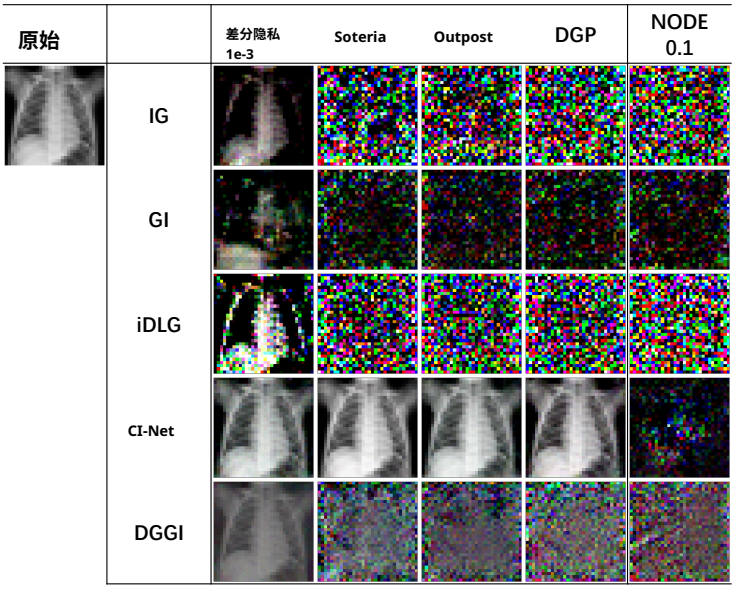

图3：胸部X光在不同攻击和（部分）防御方法下的原始图像与恢复图像对比。

5.5 ImageNet 防御结果

本节进一步介绍我们在 ImageNet数据集 上开展的实验，重点关注防御有效性（表3）。NODE 在各种攻击下均取得较低的结构相似性指数，表明其可扩展至大规模、高分辨率成像场景。NODE 的重建结果表明，它在各种攻击方法下均能有效阻碍对抗性恢复，同时保持较高的模型准确率。本节还将介绍特征相似性指数值以及更多恢复图像，以便更全面地了解防御性能。我们还将讨论各方法的训练动态和时间开销，以评估其效率。此外，我们开展了消融实验，以考察近正交投影的必要性，并凸显其对于提升防御机制抵御梯度反演攻击的鲁棒性的重要作用。

5.6 消融研究：余弦相似度的影响

我们提出的方法通过在原始梯度的垂直方向注入噪声，将原始梯度投影到近乎正交的方向上。为证明这种近乎正交设计的必要性，我们开展了以下消融实验。

我们调整噪声尺度，以操控防御梯度 g' 与原始梯度g之间的余弦相似度 α。如表 4 所示，较高的余弦相似度（即较小的角度）对应于较弱的噪声注入，因此可获得更高的结构相似性指数值，但对反演攻击的防御能力也更弱。例如，CI-Net 几乎可以

表 3：ImageNet 上的性能：在多种防御方法下，针对五种梯度反演攻击（GI、IG、iDLG、CI-Net 和 DGGI）的平均结构相似性指数（含标准差）。最佳性能以粗体标示，次佳性能以下划线标示。

防御方法	攻击（结构相似性指数）					
	GI	IG	iDLG	CI-Net	DGGI	Mean
None	0.1087$_{\pm 0.0318}$	0.0963$_{\pm 0.0230}$	0.0388$_{\pm 0.0073}$	0.6531$_{\pm 0.0896}$	0.3897$_{\pm 0.0382}$	0.2573
剪枝（p=0.9）	0.0841$_{\pm 0.0315}$	0.0548$_{\pm 0.0203}$	0.0049$_{\pm 0.0003}$	0.1091$_{\pm 0.0154}$	0.0259$_{\pm 0.0259}$	0.0558
差分隐私随机梯度下降（σ=1e-3）	0.1698$_{\pm 0.0404}$	0.0142$_{\pm 0.0018}$	0.0045$_{\pm 0.0004}$	0.1430$_{\pm 0.0489}$	0.1813$_{\pm 0.0202}$	0.1026
差分隐私随机梯度下降（σ=4e-3）	0.1157$_{\pm 0.0390}$	0.0054$_{\pm 0.0004}$	0.0048$_{\pm 0.0002}$	0.0336$_{\pm 0.0064}$	~~0.0196~~$_{\pm 0.0010}$	0.0358
Soteria	0.1017$_{\pm 0.0350}$	0.0091$_{\pm 0.0010}$	0.0047$_{\pm 0.0003}$	0.0705$_{\pm 0.0137}$	0.0713$_{\pm 0.0067}$	0.0516
Outpost	0.0877$_{\pm 0.0408}$	0.0564$_{\pm 0.0194}$	0.0060$_{\pm 0.0008}$	0.1715$_{\pm 0.0300}$	0.3388$_{\pm 0.0238}$	0.1321
DGP	0.1012$_{\pm 0.0158}$	**0.0054**$_{\pm 0.0004}$	0.0046$_{\pm 0.0002}$	0.0525$_{\pm 0.0181}$	**0.0106**$_{\pm 0.0014}$	0.0349
NODE ($\alpha = 0.2$)	~~0.0652~~$_{\pm 0.0298}$	0.0072$_{\pm 0.0010}$	~~0.0044~~$_{\pm 0.0002}$	~~0.0208~~$_{\pm 0.0037}$	0.0578$_{\pm 0.0206}$	<u>0.0311</u>
NODE ($\alpha = 0.1$)	**0.0572**$_{\pm 0.0230}$	~~0.0058~~$_{\pm 0.0004}$	0.0046$_{\pm 0.0002}$	**0.0167**$_{\pm 0.0023}$	0.0320$_{\pm 0.0096}$	**0.0233**

表 4：在 CIFAR-10 上，采用不同余弦相似度 α 时 NODE 在 CI-Net 攻击下的测试准确率和平均结构相似性指数。

方法	角度 θ	信噪比 (dB)	测试准确率 ↑	CI-Net ↓
α = 0.7	45.57°	-0.0869	91.23$_{\pm 0.14}$	0.9570$_{\pm 0.0231}$
α = 0.3	72.54°	-5.0240	90.74$_{\pm 0.24}$	0.5610$_{\pm 0.1163}$
α = 0.2	78.46°	-6.9011	90.30$_{\pm 0.14}$	0.2422$_{\pm 0.0770}$
α = 0.1	84.26°	-9.9782	89.31$_{\pm 0.02}$	0.0790$_{\pm 0.0265}$

恢复原始数据，结构相似性指数值达到 0。9570，这违背了定理 4.5 中所述的安全原则，因为这会导致 ℓ_1（或 σ）的下界更小。相比之下，较大的角度（例如， $84.26°$ 且 $\alpha = 0.1$）能够增强对梯度反演攻击的防御能力，将 CI-Net 的结构相似性指数降至 0.0790。虽然这能提供更好的保护，但代价是模型性能略有下降。不过总体而言，建议在大多数应用中将 α 设置为 0.1，尤其是在对隐私敏感的应用中。

6 结论

我们提出了一种统一的受控动力学表述，涵盖了广泛的梯度修改防御方法，并支持统一的收敛性与安全性分析，从而无需再针对具体方法分别进行证明。在此框架下，我们提出了“近正交梯度投影”（NODE），仅在与真实梯度正交的子空间中注入噪声，以维持防御梯度与真实梯度之间固定的余弦相似度。在标准假设和固定学习率下，NODE 保持一阶下降方向，并且可以证明其维持收敛性。实验结果表明，NODE 在不同数据集上显著降低了多种梯度反演攻击的成功率，同时保持较高的任务性能，从而为隐私敏感型应用实现了实用的效用–隐私权衡。

参考文献

[1] Ligeng Zhu、Zhijian Liu 和 Song Han。梯度深度泄漏。《神经信息处理系统进展》论文集，2019。[2] Jonas Geiping、Hartmut Bauermeister、Hannah Dröge 和 Michael Moeller。梯度反演——在联邦学习中破坏隐私有多容易？《神经信息处理系统进展》论文集，33：16937–16947，2020。[3] Chi Zhang、Zhang Xiaoman、Ekanut Sotthiwat、Yanyu Xu、Ping Liu、Liangli Zhen 和 Yong Liu。联邦学习中过参数化网络的生成式梯度反演。载于《IEEE/CVF国际计算机视觉会议》论文集，第5126–5135页，2023。[4] Yangsibo Huang、Samyak Gupta、Zhao Song、Kai Li 和 Sanjeev Arora。联邦学习中的梯度反演攻击与防御评估。《神经信息处理系统进展》，34：7232–7241，2021。[5] Jingwei Sun、Ang Li、Binghui Wang、Huanrui Yang、Hai Li 和 Yiran Chen。Soteria：从表征视角出发的联邦学习隐私泄露可证明防御。载于《IEEE/CVF计算机视觉与模式识别会议》论文集，第9311–9319页，2021。[6] Martin Abadi、Andy Chu、Ian Goodfellow、H Brendan McMahan、Ilya Mironov、Kunal Talwar 和 Li Zhang。带差分隐私的深度学习。载于《ACM SIGSAC计算机与通信安全会议》论文集，第308–318页，2016。[7] Cynthia Dwork。差分隐私：研究成果综述。载于《国际计算模型理论与应用会议》，第1–19页。Springer，2008。[8] Cynthia Dwork、Aaron Roth 等。差分隐私的算法基础。《理论计算机科学基础与趋势》，9(3–4)：211–407，2014。[9] Bo Zhao、Konda Reddy Mopuri 和 Hakan Bilen。iDLG：改进的梯度深度泄漏。CoRR，2020。URL https://arxiv.org/abs/2001.02610。[10] Liwen Wu、Zhizhi Liu、Bin Pu、Kang Wei、Hangcheng Cao 和 Shaowen Yao。DGGI：基于扩散模型的深度生成式梯度反演。《信息融合》，113：102620，2025。[11] Jakub Koneˇcn`y、H Brendan McMahan、Felix X Yu、Peter Richtárik、Ananda Theertha Suresh 和 Dave Bacon。联邦学习：提升通信效率的策略。arXiv预印本 arXiv:1610.05492，2016。[12] Brendan McMahan、Eider Moore、Daniel Ramage、Seth Hampson 和 Blaise Aguera y Arcas。基于去中心化数据的深度网络通信高效学习。载于《人工智能与统计学》。PMLR，2017。[13] Chi Zhang 和 Qianxiao Li。过参数化产生的退化损失函数的分布式优化。《人工智能》，301：103575，2021。[14] Hongxu Yin、Arun Mallya、Arash Vahdat、Jose M Alvarez、Jan Kautz 和 Pavlo Molchanov。透视梯度：通过gradinversion进行图像批次恢复。载于《IEEE/CVF计算机视觉与模式识别会议》论文集，第16337–16346页，2021。[15] Jianping Geng、Yanmin Mou、Qian Li、Quanzheng Li 和 Haonan Zhu。通过差分隐私数据合成实现隐私保护联邦学习。载于《AAAI人工智能会议》，2021。[16] Franziska Boenisch、Adam Dziedzic、Roei Schuster、Ali Shahin Shamsabadi、Ilia Shumailov 和 Nicolas Papernot。当好奇者放弃诚实：联邦学习并不私密。载于《IEEE欧洲安全与隐私研讨会》（EuroS&P），2023。[17] Ali Hatamizadeh、Hongxu Yin、Holger R Roth、Wenqi Li、Jan Kautz、Daguang Xu 和 Pavlo Molchanov。Gradvit：视觉Transformer的梯度反演。载于《IEEE/CVF计算机视觉与模式识别会议》论文集，第10021–10030页，2022。[18] Ekanut Sotthiwat、Liangli Zhen、Chi Zhang、Zengxiang Li 和 Rick Siow Mong Goh。从梯度进行生成式图像重建。《IEEE神经网络与学习系统汇刊》，2024。[19] Jinwoo Jeon、Kangwook Lee、Sewoong Oh、Jungseul Ok 等。利用生成式图像先验进行梯度反演。《神经信息处理系统进展》，34：29898–29908，2021。[20] Hao Chen、Zhikun Huang、Tao Huang、Shaowei Wang 和 Ming Zhao。Gifd：采用特征域优化的生成式梯度反演方法。载于《IEEE/CVF计算机视觉与模式识别会议》（CVPR），2024。[21] Mislav Balunovi´c、Dimitar I Dimitrov、Nikola Jovanovi´c 和 Martin Vechev。Lamp：利用语言模型先验从梯度中提取文本。载于《神经信息处理系统进展》（NeurIPS），2022。

[22] Liam Fowl、Jonas Geiping、Wojciech Czaja、Micah Goldblum 和 Tom Goldstein。《Decepticons：被篡改的 Transformer 如何在语言模型联邦学习中侵犯隐私》。载于国际学习表征会议（ICLR），2022 年。[23] Samyak Gupta、Yangsibo Huang、Zexuan Zhong、Tianyu Gao、Kai Li 和 Danqi Chen。《在语言模型联邦学习中恢复私有文本》。载于神经信息处理系统进展（NeurIPS），2022 年。[24] Jieren Deng、Yijue Zhou、Milad Nasr、Amir Houmansadr、Luca Melis 和 Vitaly Shmatikov。《Tag：针对基于 Transformer 的语言模型的梯度攻击》。载于计算语言学协会论文集：EMNLP Findings，2021 年。[25] Andrew C Yao。《安全计算协议》。载于第 23 届计算机科学基础年度研讨会（SFCS 1982）。IEEE，1982 年。[26] Ronald L Rivest、Len Adleman、Michael L Dertouzos 等。《关于数据库与隐私同态》。载于《安全计算基础》，4(11):169–180，1978 年。[27] Chuan Zhao、Shengnan Zhao、Minghao Zhao、Zhenxiang Chen、Chong-Zhi Gao、Hongwei Li 和 Yu-an Tan。《安全多方计算：理论、实践与应用》。《信息科学》，476:357–372，2019 年。[28] Hossein Fereidooni、Samuel Marchal、Markus Miettinen、Azalia Mirhoseini、Helen Möllering、Thien Duc Nguyen、Phillip Rieger、Ahmad-Reza Sadeghi、Thomas Schneider、Hossein Yalame 等。《SafeLearn：面向隐私保护联邦学习的安全聚合》。载于 2021 年 IEEE 安全与隐私研讨会（SPW），第 56–62 页。IEEE，2021 年。[29] Jing Ma、Sheikh Ahnaf Naas、Stephan Sigg 和 Xixiang Lyu。《基于多密钥同态加密的隐私保护联邦学习》。《国际智能系统杂志》，第 37 卷，第 5880–5901 页，2022 年。[30] Keith Bonawitz、Vladimir Ivanov、Ben Kreuter、Antonio Marcedone、H Brendan McMahan、Sarvar Patel、Daniel Ramage、Aaron Segal 和 Karn Seth。《面向隐私保护机器学习的实用安全聚合》。载于 ACM SIGSAC计算机与通信安全会议（CCS），2017 年。[31] James Henry Bell、Kallista A Bonawitz、Adrià Gascón、Tancrède Lepoint 和 Mariana Raykova。《具有（多）对数开销的安全单服务器聚合》。载于 ACM SIGSAC计算机与通信安全会议（CCS），2020 年。[32] Chi Zhang、Sotthiwat Ekanut、Liangli Zhen 和 Zengxiang Li。《抵御联邦学习中梯度泄露的增强型多方计算》。《IEEE大数据汇刊》，2022 年。[33] Ekanut Sotthiwat、Liangli Zhen、Zengxiang Li 和 Chi Zhang。《用于联邦学习的部分加密多方计算》。载于 2021 年 IEEE/ACM 第 21 届集群、云与互联网计算国际研讨会（CCGrid），2021 年。[34] Ilya Mironov。《Rényi差分隐私》。载于 2017 年 IEEE 第 30 届计算机安全基础研讨会（CSF），第 263–275 页。IEEE，2017 年。[35] Kang Wei、Jun Li、Ming Ding、Chuan Ma、Howard H Yang、Farhad Farokhi、Shi Jin、Tony QS Quek 和 H Vincent Poor。《差分隐私联邦学习：算法与性能分析》。《IEEE信息取证与安全汇刊》，15:3454–3469，2020 年。[36] Zhen Ling Teo、Xiaoman Zhang、Yechao Yang、Liyuan Jin、Chi Zhang、Stanley Shuoh Jieh Poh、Weihong Yu、Youxin Chen、Jost B Jonas、Ya Xing Wang 等。《利用联邦学习与区块链技术保护视网膜成像免受对抗攻击》。《眼科学》，132(4):484–494，2025 年。[37] Jordan Noble、Sam Adams、Gueorgui Georgiev、Raluca Munteanu 和 Tom Chothia。《联邦学习中的自适应差分隐私：一种基于优先级的方法》。载于国际可用性、可靠性与安全会议，2022 年。[38] Alexey Kurakin、Steve Chien、Shuang Song、Roxana Geambasu、Andreas Terzis 和 Abhradeep Thakurta。《迈向使用差分隐私进行 ImageNet 规模训练》。arXiv 预印本 arXiv:2201.12328，2022 年。[39] Junyi Zhu 和 Matthew Blaschko。《梯度解聚：通过重建用户参与者矩阵打破联邦学习中的隐私保护》。载于国际机器学习会议（ICML），2021 年。[40] Zihan Wang、Meisam Song、Zhibo Zhang、Yang Song、Qian Wang 和 Hairong Qi。《Sapag：一种基于梯度的自适应隐私攻击》。载于 arXiv 预印本 arXiv:2009.06228，2021 年。[41] Daniel Scheliga、Patrick Mäder 和 Marco Seeland。《Precode：一种用于防止深度梯度泄露的通用模型扩展》。载于 IEEE/CVF冬季计算机视觉应用会议（WACV），2022 年。[42] Yangsibo Huang、Zhao Song、Kai Li 和 Sanjeev Arora。《InstaHide：面向隐私保护分布式学习的实例隐藏方案》。载于国际机器学习会议（ICML），2021 年。

[43] Zhe Shen、Juan Cervino、Hamed Hassani 和 Alejandro Ribeiro。《如何避免个性化联邦学习中的客户端性能不佳？》载于 arXiv 预印本 arXiv:2201.11942，2022 年。[44] Bingchen Li、Hanlin Gu、Lingjuan Pang、Song Wang、Jiannong Cao 和 Jun Li。《AuditFL：用于隐私保护人脸识别的可审计联邦学习框架》。载于《IEEE可信与安全计算汇刊》，2022 年。[45] Jonas Geiping、Hartmut Bauermeister、Hannah Dröge 和 Michael Moeller。《攻破 FedMD：通过配对逻辑值反演攻击恢复图像》。载于 arXiv 预印本 arXiv:2108.13673，2021 年。[46] Florian Tramer 和 Dan Boneh。《Truth Serum：通过投毒机器学习模型揭示其秘密》。载于 ACM SIGSAC计算机与通信安全会议（CCS），2022 年。[47] Weinan E。《基于动力系统的机器学习提案》。《数学与统计学通讯》，1(5):1–11，2017 年。[48] Qianxiao Li、Long Chen、Cheng Tai 等。《用于深度学习的基于最大值原理的算法》。arXiv 预印本 arXiv:1710.09513，2017 年。[49] Kaiming He、Xiangyu Zhang、Shaoqing Ren 和 Jian Sun。《用于图像识别的深度残差学习》。载于 IEEE 计算机视觉与模式识别会议论文集，第 770–778 页，2016 年。[50] Gene F Franklin、J David Powell、Abbas Emami-Naeini 和 J David Powell。《动态系统的反馈控制》，第 4 卷。2002 年。[51] John C Doyle、Bruce A Francis 和 Allen R Tannenbaum。《反馈控制理论》。Courier Corporation，2013 年。[52] Martin Benning、Elena Celledoni、Matthias J Ehrhardt、Brynjulf Owren 和 Carola-Bibiane Schönlieb。《作为最优控制问题的深度学习：模型与数值方法》。arXiv 预印本 arXiv:1904.05657，2019 年。[53] Paulo Tabuada 和 Bahman Gharesifard。《通过非线性控制理论论证深度残差神经网络的通用逼近能力》。arXiv 预印本 arXiv: 2007.06007，2020 年。[54] Chi Zhang、Cheng Jingpu、Yanyu Xu 和 Qianxiao Li。《结合控制的参数高效微调》。载于第四十一届国际机器学习会议，2024 年。[55] Chi Zhang、Jingpu Cheng 和 Qianxiao Li。《LoRA 与视觉 Transformer二元控制器设计的最优控制视角》。载于欧洲计算机视觉会议，第 144–160 页。Springer，2024 年。[56] Chi Zhang、Lianhai Ren、Jingpu Cheng 和 Qianxiao Li。《从基于权重到基于状态的微调：通过并行控制进一步减少 LoRA 的内存占用》。载于 ICML，2025 年。[57] Léon Bottou、Frank E Curtis 和 Jorge Nocedal。《大规模机器学习的优化方法》。《SIAM评论》，60(2):223–311，2018 年。[58] Kaiyuan Zhang、Siyuan Cheng、Guangyu Shen、Bruno Ribeiro、Shengwei An、Pin-Yu Chen、Xiangyu Zhang 和 Ninghui Li。《Censor：通过正交子空间贝叶斯采样防御梯度反演》。arXiv 预印本 arXiv:2501.15718，2025 年。[59] Yurii Nesterov 等。《凸优化讲义》。Springer，2018 年。[60] Stephen J Wright。《坐标下降算法》。《数学规划》，151(1):3–34，2015 年。[61] Paul Tseng。《不可微最小化的块坐标下降方法的收敛性》。《优化理论与应用杂志》，109:475–494，2001 年。[62] Aleksandr Mikhailovich Lyapunov。《运动稳定性的一般问题》。《国际控制杂志》，55(3):531–534，1992 年。[63] Fei Wang、Ethan Hugh 和 Baochun Li。《过犹不及：针对生产环境联邦学习中梯度泄露的自适应防御》。[64] Lulu Xue、Shengshan Hu、Ruizhi Zhao、Leo Yu Zhang、Shengqing Hu、Lichao Sun 和 Dezhong Yao。《重新审视梯度剪枝：抵御梯度攻击的双重实现》，2024 年 1 月。URL http: //arxiv.org/abs/2401.16687。arXiv:2401.16687 [cs]。[65] Ekanut Sotthiwat、Chi Zhang、Xiaokui Xiao 和 Liangli Zhen。《通过梯度随机失活保护联邦学习免受数据重构攻击》。《IEEE信息取证与安全汇刊》，2026 年。[66] Alex Krizhevsky、Geoffrey Hinton 等。《从微小图像中学习多层特征》。2009 年。[67] Kermany 等。《通过基于图像的深度学习识别医学诊断与可治疗疾病》。《Cell》，2018 年。ISSN 0092-8674，1097-4172。doi:10.1016/j.cell.2018.02.010。出版商：Elsevier。[68] Roman Vershynin。《高维概率》。2018 年。URL https://api.semanticscholar.org/CorpusID:196128750。

A 定理证明

A.1 定理 3.2 和定理 3.3 的证明

定理 3.2 和 3.3 在标准凸性假设下建立了一个收敛结果，其证明基于凸分析的基本原理。有关相关概念的详细论述，可参见 [57]，中的一系列凸分析经典著作，例如第 4.2 节和第 4.3 节，其中全面概述了所用的技术。

A.2 备注 3.5 的证明

证明 根据损失函数 L 的凸性，我们有：

$$L(y) \geq L(x) + \nabla L(x)^T (y - x), \forall x, y \in \mathbb{R}^n \tag{23}$$

若防御梯度与真实梯度垂直，即 $\nabla L(x)^T g = 0$，则可令 $y = x - \eta g$，从而得到所需结果：

$$L(x - \eta g) \geq L(x). \tag{24}$$

A.3 定理 3.6 的证明

证明 根据利普希茨连续性，我们有

$$\|G_w(x) - G_w(x')\|_2 \leq l_g \|x - x'\|_2. \tag{25}$$

将 $G_w(x')$ 替换为 $g' = C(g, x)$，并将 x' 替换为 $G_w^{-1}(C(g, x))$ 中的一个元素，可得：

$$\frac{1}{l_g} \|g - g'\|_2 \leq \|x - x'\|_2, \quad \forall x' \in G_w^{-1}(g') \tag{26}$$

thus

$$\frac{1}{l_g} \|g - g'\|_2 \leq \min_{x' \in G_w^{-1}(g')} \|x - x'\|_2, \tag{27}$$

对两边的 $x \in D$ 取最小值，可得：

$$\min_{x \in D} \frac{1}{l_g} \|g - C(g, x)\|_2 \leq \min_{x \in D} \min_{x' \in G_w^{-1}(g')} \|x - x'\|_2 = \ell_1 \tag{28}$$

A.4 定理 3.7 的证明

证明 由于差分隐私随机梯度下降所添加的噪声均值为零，我们有 $\mathbb{E}[g'] = \nabla L(w)$，因此：

$$\nabla L(w)^\top \mathbb{E}[g'] = \|\nabla L(w)\|_2^2, \tag{29}$$

即 $\mu = 1$， $\mu_g = 1$。此外，防御梯度 g' 的方差就是噪声的方差，因此：

$$\text{Var}(g') = \text{Var}(\xi) = a^2 \sigma^2, \tag{30}$$

这意味着 $M = a^2\sigma^2$， $M_V = 0$ 将 μ， M 代入公式 (14)，即可得到结果。

A.5 定理 3.9 的证明

证明 记 $\xi_t = a\sigma X$，其中 $X \sim N(0, I_m)$。根据 [68], 中的定理 3.1.1，我们得到

$$\mathbb{P}\{|\|X\|_2 - \sqrt{m}| \geq t\} \leq 2\exp(-\frac{kt^2}{K^4}), \tag{31}$$

其中 K 是 $X_i \sim N(0, 1)$ 的独立同分布范数，为一个常数。 k 是另一个绝对常数。将 $\exp(-\frac{kt^2}{K^4})$ 替换为 δ，我们知道

$$\|X\|_2 \geq \sqrt{m} - \frac{K^2}{\sqrt{k}} \sqrt{\log \delta^{-1}} \tag{32}$$

以 $1 - 2\delta$ 的概率成立，即

$$\|C(g, x) - g\|_2 = a\sigma \|X\|_2 \geq a\sigma \left(\sqrt{m} - a_0 \sqrt{\log \delta^{-1}} \right) \tag{33}$$

以概率 $1 - 2\delta$ 的概率，其中 $a_0 = \frac{K^2}{\sqrt{k}}$ 是一个常数。由此，定理 3.6 的证明完成。 .

A.6 定理 3.10 的证明

证明 由于 w　我们总是剪枝较小的值，并且剪枝方法中不存在随机性，因此可以得到:

$$\nabla L(w)^\top \mathbb{E}[g'] \geq p\|\nabla L(w)\|_2^2, \tag{34}$$

$$\|E[g']\|_2 \leq \|\nabla L(w)\|_2, \tag{35}$$

$$Var(g') = 0, \tag{36}$$

其中 p 是剪枝率。于是我们可以得到 $\mu \geq p > 0$， $\mu < \mu_G \leq 1$， $M = 0 = M_G$，从而得到 $\ell_{0,\ \text{Pruning}} = 0$。即使考虑随机剪枝，我们也有 $Var(g') \leq \|g'\|_2 \leq \|g\|_2^2$，其他系数保持不变，因此仍然可以推导出相同的结果。

A.7 定理 3.12 的证明

证明 与定理 3.9 的证明类似，我们记 $g = \sigma X$，其中 $X \sim N(0,\ I_m)$。对 $\hat{X}$ 使用 [68] 中的定理 3.1.1，可得

$$\mathbb{P}\left\{\left|\frac{\|\hat{X}\|_2}{\sqrt{V(a)}} - \sqrt{m}\right| \geq t\right\} \leq 2\exp\left(-\frac{kt^2}{K^4}\right), \tag{37}$$

其中 K和 k是定理 3.9 证明中相同的常数。令 $s = t\sqrt{V(a)}$ $\delta = \exp(-\frac{kt^2}{K^4})$，并使用定理 3.6，我们以相同的方式完成证明。

A.8 定理 4.1 的证明

证明 与定理 4.1 的证明类似，我们需要分析式 (14) 中的参数。由于噪声服从均匀分布，其期望为 0，因此有。

$$\nabla L(w)^\top \mathbb{E}[g'] = \|\nabla L(w)\|_2^2, \quad \|\mathbb{E}[g']\|_2 = \|\nabla L(w)\|_2 \tag{38}$$

即 $\mu = 1 = \mu_G$。此外，防御梯度 g' 与 g 之间的角度是固定的，因此防御梯度 g' 的范数是固定的，即

$$\|g'\|_2 \equiv \sec\theta \|g\|_2. \tag{39}$$

因此，

$$Var(g') = \mathbb{E}[\|g'\|_2^2] - \|\mathbb{E}[g']\|_2^2 = \tan^2\theta \|g\|_2^2, \tag{40}$$

这意味着 $M = 0$， $M_V = \tan^2\theta$。根据定理 3.2 和式 14，我们有 $\ell_{0,\ \text{NODE}} = 0$。

A.9 定理4.3的证明

证明 李雅普诺夫函数是一个标量函数 $V(w): \mathbb{R}^m \to \mathbb{R}$，用于研究动力系统的稳定性。对于如下形式的动力系统

$$\frac{dw}{dt} = f(w),$$

其中， w 表示系统的状态， $f(w)$表示系统随时间的演化，李雅普诺夫函数 $V(w)$ 满足以下条件:

1. 正定定：对于所有 w， $V(w) \geq 0$，且仅在平衡点（通常为 $w = 0$）处， $V(w) = 0$).
2. 沿系统轨迹递减： $V(w)$ 的时间导数记为

$$\frac{dV(w)}{dt} = \nabla V(w) \cdot f(w),$$

应为负定或负半定。该条件意味着李雅普诺夫函数会随时间递减，表明系统正在“趋向”平衡点。

如果满足这些条件，则平衡点是稳定的。具体而言:

- 渐近　c 稳定性：当 $\frac{dV(w)}{dt}$ 为负定时成立，这意味着 $V(w)$ 随时间严格递减　e.

- 李雅普诺夫意义下的稳定性：当 $\frac{dV}{dt}(w)$ 为负半定时成立，这意味着 $V(w)$ 要么递减，要么保持不变。

对于训练动态方程 (1)，很容易证明，其平衡点是损失函数$L(w)$的鞍点；在强凸情形下，这些平衡点也是极小值点。因此，收敛到最优损失等价于控制理论中的稳定性性质。

为证明 $V := (L(w) - L^*)^2$ 是 NODE 方法的李雅普诺夫函数，我们需要验证：(1) $V(w^*) = 0$；(2) $V(w) > 0$, $\forall w \neq w^*$；(3) $\frac{dV}{dt} < 0$。
对于强凸的 $L(w)$，很容易验证前两个条件均成立。我们只需证明第三个条件：

$$\begin{aligned}\frac{dV}{dt} &= 2(L(w(t)) - L^*)\frac{dL}{dt} \\ &= 2(L(w(t)) - L^*)\nabla L(w(t))^\top \frac{dw}{dt} \\ &= -2(L(w(t)) - L^*)\|\nabla L(w(t))\|_2^2 < 0.\end{aligned} \tag{41}$$

因此，$(L(w) - L^*)^2$ 是 NODE 方法的李雅普诺夫函数，并且会渐近达到最小值 0，正如定理所述。

A.10 定理 4.5 的证明

证明： 根据假设 4.4，梯度范数阈值可取为 $a = b\|\nabla L(w)\|_2$，噪声尺度为$a^2\sigma^2 = \tan^2\theta\|\nabla L(w)\|_2^2$。因此 $\sigma = \tan\theta\|\nabla L(w)\|_2 \geq \tan\theta$。

B 补充实验

本部分包含用于支持第 5 节的补充实验。

B.1 详细实验设置

我们在模拟联邦学习环境中评估所提出的方法。遵循以往关于梯度反演攻击的研究，我们采用带有 Sigmoid 激活函数的 ResNet-18 模型；已有研究表明，与常用的激活函数相比，这类模型更容易发生梯度泄露。联邦设置由三个模拟客户端组成，每个客户端在与服务器通信梯度之前执行一个本地优化步骤。模型采用不带动量的随机梯度下降（SGD）进行训练，批大小为 16。初始学习率设为 0.1，并在整个训练过程中根据余弦衰减计划逐步降低。

B.2 攻击的特征相似性指数结果

除了结构相似性指数（SSIM）外，我们还使用特征相似性指数（FSIM）评估防御方法，以衡量重建图像的质量。

表5：在 CI-Net 攻击下，CIFAR10、Tiny-ImageNet、胸部X光和 ImageNet 上的平均特征相似性指数。最佳结果以粗体显示，次优结果加下划线。

防御方法	CIFAR-10	Tiny-ImageNet	胸部X光	ImageNet	Mean
None	$0.9843_{\pm 0.0083}$	$0.9899_{\pm 0.0032}$	$0.6725_{\pm 0.1704}$	$0.6483_{\pm 0.1217}$	0.8238
剪枝 (p=0.9)	$0.9117_{\pm 0.0257}$	$0.9313_{\pm 0.0146}$	$0.6225_{\pm 0.1364}$	$0.5732_{\pm 0.0817}$	0.7597
剪枝（p=0.99）	$0.6602_{\pm 0.0363}$	$0.7012_{\pm 0.0417}$	$0.6429_{\pm 0.0153}$	$0.5721_{\pm 0.0838}$	0.6441
差分隐私 (σ=1e-3)	$0.9394_{\pm 0.0256}$	$0.9607_{\pm 0.0062}$	$0.6452_{\pm 0.1321}$	$0.5589_{\pm 0.0900}$	0.7761
差分隐私（σ=4e-3）	$0.6796_{\pm 0.0451}$	$0.7486_{\pm 0.0188}$	$0.5569_{\pm 0.0425}$	$0.5102_{\pm 0.0814}$	0.6238
Soteria	$0.9864_{\pm 0.0064}$	$0.9901_{\pm 0.0024}$	$0.6408_{\pm 0.1680}$	$0.6408_{\pm 0.1680}$	0.8145
Outpost	$0.6353_{\pm 0.1613}$	$0.7244_{\pm 0.1297}$	$0.6296_{\pm 0.1493}$	$0.5783_{\pm 0.0836}$	0.6419
DGP	$0.6284_{\pm 0.1324}$	$0.6490_{\pm 0.1472}$	$0.6176_{\pm 0.1350}$	$\mathbf{0.4993}_{\pm 0.0787}$	0.5986
NODE（α=0.2）	$0.6521_{\pm 0.0479}$	$0.7507_{\pm 0.0426}$	$0.5803_{\pm 0.0533}$	$0.5236_{\pm 0.0795}$	0.6267
NODE (α=0.1)	$\mathbf{0.5731}_{\pm 0.0342}$	$\mathbf{0.5993}_{\pm 0.0240}$	$\mathbf{0.5546}_{\pm 0.0279}$	$0.5081_{\pm 0.0721}$	**0.5588**

表5报告了 CI-Net 攻击下不同防御算法的 FSIM值。在 CIFAR-10 数据集上，未采用任何防御措施时的基线性能对应的 FSIM值高达 0.9843，这表明模型容易受到梯度反演攻击，进而导致数据泄露。如此高的值说明，若缺乏适当的防御机制，模型梯度会保留过多信息，可能使原始数据被重建。在测试的防御方法中，NODE技术表现最为出色，其 FSIM值最低，为 0.5731，表明它能有效防御反演攻击。它优于其他方法，包括紧随其后的 Soteria，其 FSIM值高达 0.9864。因此，NODE 在 CIFAR-10 上展现出最佳防御效果，在维持稳健防御性能的同时显著降低了数据泄露风险。

在 Tiny-ImageNet 数据集中，采用 $\alpha = 0.1$ 的 NODE 在防御梯度反演攻击方面继续表现出显著提升，取得了最佳特征相似性指数值 0.5993。这表明其大幅降低了重构原始数据的能力，凸显了该方法的有效性。然而，较大剪枝比例（$p = 0.99$）的剪枝方法和较高噪声水平（σ=4e-3）的差分隐私（DP）可能在防御强度方面优于采用 $\alpha = 0.2$ 的 NODE，其特征相似性指数值分别为 0.7012 和 0.7486。采用 $\alpha = 0.2$ 的 NODE 的防御性能略高于这两种方法，但在此场景下取得了最佳测试准确率 53。27%。

B.3 效率分析

除了准确率和防御有效性之外，我们还从训练时间的角度评估了我们的方法。图4比较了不同防御方法每轮次的训练时长。

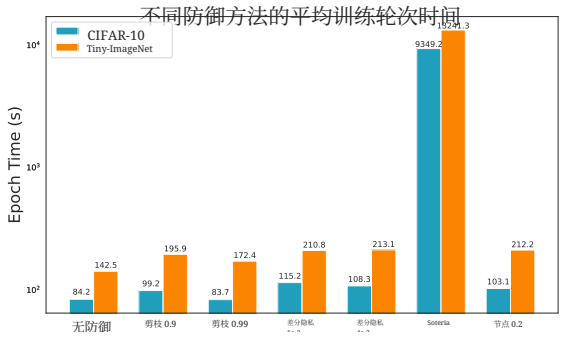

图4：每种防御方法单轮次的平均运行时间。

在 CIFAR-10 数据集上，我们得出以下结论：

- 基线方法（无防御）耗时最少（84.2 秒），几乎是效率最高的算法。
- 剪枝（p=0.9）（99.2 秒）和剪枝（p=0.99）（83.7 秒）都会带来小幅的时间增加。虽然剪枝本身可能会增加额外的计算时间，但使用剪枝后的梯度进行反向传播通常所需时间更少。因此，剪枝并不会导致防御梯度反演的时间显著增加。
- 差分隐私方法 DP (σ=1e-3) 和 DP (σ=4e-3) 的耗时明显增加，分别为 115.2 秒和 108.3 秒。
- Soteria 的计算开销显著增加，每轮次耗时 9349.2 秒。
- 与 Soteria 不同，NODE 方法的耗时适度增加，平均训练轮次时间为 103.1 秒。这使 NODE 的时间开销与其他方法（如差分隐私和剪枝）相当。

Tiny-ImageNet 数据集上的结果与上述结论一致。总体而言，我们的结果表明，NODE 的开销与差分隐私随机梯度下降方法相当或略低。这种效率使 NODE 特别适用于对性能和速度都有严格要求的实际联邦学习应用。