MoE水印对抗释义攻击的核心机理优势
简化版

yunhao

1 核心问题

传统词元-逻辑值（Token-Logit）水印在面对释义攻击时存在根本性脆弱性。本文从信息论角度证明，基于MoE专家激活的水印范式具有机理上的根本优势。

2 两种范式的核心差异

2.1 范式A：稠密模型水印（KGW）

信号载体：词元概率分布（词汇表空间）
攻击向量：词元替换（词汇表空间）
核心问题：信号与攻击在同一空间，导致**信号-攻击重合**
衰减机理：线性衰减

$$\Delta Z \propto \gamma \tag{1}$$

其中γ为攻击强度（被替换词元比例）。

2.2 范式B：MoE水印

信号载体：专家激活分布$p(e|x)$（内部激活空间）
攻击向量：词元替换$x \to x'$（外部词元空间）
核心优势：信号与攻击在不同空间，实现**信号-攻击解耦**
衰减机理：次线性衰减

$$\Delta D^* \propto O(\sqrt{\gamma}) \tag{2}$$

3 机理推导

3.1 攻击建模

将释义攻击$x \to x'$建模为输入分布偏移：

$$\gamma = D_{\mathrm{KL}}(D(X')\|D(X)) \tag{3}$$

3.2 核心定理

定理 1 (对抗鲁棒性). *在强度为γ的释义攻击下，MoE水印的检测能力（Chernoff信息）满足：*

$$D^*_{adv} \geq D^*(p_0, p_1) - C\sqrt{\gamma \cdot D^*(p_0, p_1)} - O(\gamma) \tag{4}$$

其中$C \approx 1.5$–2.0为常数。

证明思路：通过Pinsker不等式，输入空间的KL散度γ传播到激活空间时，被约束为$\sqrt{\gamma}$量级的总变差距离，从而检测能力衰减为次线性。

3.3 为什么是$\sqrt{\gamma}$?

关键步骤：

1. **空间解耦：**攻击γ在输入空间，信号D^*在激活空间
2. **Pinsker不等式：**

$$\|p' - p\|_{\text{TV}} \leq \sqrt{\frac{1}{2} D_{\text{KL}}(p' \| p)} \tag{5}$$

3. **传播约束：**输入分布的攻击γ导致激活分布变化：

$$\|p'_i - p_i\|_{\text{TV}} \leq \sqrt{2\gamma} \tag{6}$$

4. **结论：**检测能力衰减ΔD^*被$O(\sqrt{\gamma})$约束

4 对比分析

4.1 数值示例

假设初始$D^* = 0.1$，$C = 1.5$：

- **攻击1：**$\gamma_1 = 0.005$
  - 衰减项：$1.5 \cdot \sqrt{0.005 \cdot 0.1} \approx 0.0335$
  - 剩余$D^*_{\text{adv}} \geq 0.0665$（损失33.5%）
- **攻击2：**$\gamma_2 = 0.020$（4倍于攻击1）
  - 衰减项：$1.5 \cdot \sqrt{0.020 \cdot 0.1} \approx 0.0671$
  - 剩余$D^*_{\text{adv}} \geq 0.0329$（损失67.1%）

关键观察：攻击强度增加4倍，信号损失仅增加2倍（$\sqrt{4} = 2$）。

4.2 线性 vs 次线性衰减

攻击强度γ	范式A（线性）	范式B（次线性）
0.01	信号保留75%	信号保留52.6%
0.02	信号保留50%	信号保留32.9%
0.04	**信号完全丢失**	**仍可检测（5.1%）**

核心优势：次线性衰减在强攻击下仍保持可检测性，而线性衰减会灾难性失效。

5 工程框架

5.1 安全系数c

将水印强度ϵ与预期攻击强度γ关联：

$$\epsilon \approx c \cdot \sqrt{\gamma} \quad 或 \quad D^* \approx c^2 \gamma \tag{7}$$

鲁棒性保证：若$c > C$（约1.5–2.0），则$D^*_{\text{adv}} > 0$，水印理论上保证可检测。

6 核心结论

1. **信号-攻击解耦**是MoE水印鲁棒性的根本原因
2. **次线性衰减**（$O(\sqrt{\gamma})$）vs **线性衰减**（$O(\gamma)$）是机理上的根本差异
3. 通过Pinsker不等式，输入空间的攻击被约束为激活空间的$\sqrt{\gamma}$量级变化
4. 次线性衰减提供了**可检测性下限**，而线性衰减会完全失效

备注

- 实验验证：待完成
- 详细证明：待后续补充
- 本文仅展示核心理论机理