MoE 路由水印的信息论形式化定义
Information-Theoretic Formalization of MoE Routing Watermarking

Yunhao

Abstract

本文提出了混合专家（MoE）架构中路由水印的信息论形式化框架。我们首先定义了路由状态空间 $\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi)$，其中 $\mathbf{r} \in \Delta^{n-1}$ 为路由分布（位于概率单纯形），$\Sigma \subset [n]$ 为激活专家集合，$\pi \in S_k$ 为激活顺序。基于信息论原理，我们量化了三个主要信息维度的编码容量：组合码（激活模式）、排列码（排列顺序）和连续码（权重量化，考虑概率单纯形约束）。在约束模型性能的条件下，我们建立了率失真（Rate-Distortion）框架，定义了可实现容量和信息-性能权衡函数。针对对抗性攻击，我们采用极小极大（min-max）准则定义鲁棒性，并基于 Le Cam/Pinsker 方法推导了可检性下界。在验证过程中，我们将问题形式化为复合假设检验，采用广义似然比检验（GLRT）或序贯概率比检验（SPRT）构建验证器，并基于 Chernoff/Bhattacharyya 上界间接推导了检测质量的 AUC 下界。最后，我们定义了综合性能指标和帕累托前沿，为 MoE 路由水印的设计和评估提供了严格的理论基础。

1 引言

混合专家（Mixture-of-Experts, MoE）架构通过稀疏激活机制实现了大型神经网络的高效设计。在 MoE 架构中，路由器（Router）负责为每个输入令牌选择性地激活专家子集，这种路由机制为水印嵌入提供了独特的信息载体。然而，目前缺乏对 MoE 路由水印的理论分析框架，特别是从信息论角度对其容量、鲁棒性和检测能力的严格量化。

本文提出了 MoE 路由水印的信息论形式化定义，旨在为路由水印的设计和评估提供理论基础。我们首先建立路由状态空间的数学表示，然后量化不同信息维度的编码容量。在此基础上，我们分析可实现容量与模型性能之间的权衡，推导鲁棒性的信息论下界，并构建验证过程的假设检验框架。本文的贡献包括：

- 定义了路由状态空间 $\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi)$ 的三个信息维度（组合码、排列码、连续码）及其容量计算，考虑了概率单纯形约束和温度/噪声影响
- 建立了率失真（Rate-Distortion）框架，定义了可实现容量与性能权衡的数学形式
- 采用极小极大（min-max）准则定义鲁棒性，并基于 Le Cam/Pinsker 方法推导了可检性下界
- 针对特定攻击类型（路由器重训练、模型蒸馏、量化与剪枝）给出了量化分析
- 构建了基于复合假设检验的验证框架（GLRT/SPRT），并基于错判率上界间接推导了检测能力的 AUC 下界

2 基础符号与系统定义

2.1 MoE 路由机制基础

设 MoE 模型的第 l 层为：

$$\mathrm{MoE}_l(x) = \sum_{i=1}^{n} g_i(x) \cdot E_i(x)$$

其中：

- $x \in \mathbb{R}^d$：输入向量
- $g_i(x)$：第 i 个专家的路由权重（由路由器 Router 产生）
- $E_i(x)$：第 i 个专家的输出
- n：专家总数

路由器输出：

$$\mathbf{r} = \mathrm{Router}(x) = \mathrm{Softmax}(\mathbf{w}^T x + b) \in \Delta^{n-1}$$

其中 $\mathrm{Softmax}(\cdot)$ 为 softmax 函数，得到路由分布 $\mathbf{r} = [r_1, r_2, \ldots, r_n]$，位于概率单纯形 $\Delta^{n-1} = \{\mathbf{r} \in \mathbb{R}^n : r_i \geq 0, \sum_{i=1}^n r_i = 1\}$。温度参数 T 用于控制分布的锐度：$\mathrm{Softmax}_T(\mathbf{z})_i = \exp(z_i/T)/\sum_j \exp(z_j/T)$。

2.2 水印系统定义

水印系统 $\mathcal{W} = (\mathcal{E}, \mathcal{V})$ 包括：

- $\mathcal{E}$：嵌入过程（Embedding）
- $\mathcal{V}$：验证过程（Verification）

3 水印信息的编码空间

3.1 路由状态空间

对于输入 x，在第 l 层的路由状态表示为：

$$\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi) \in \Delta^{n-1} \times 2^{[n]} \times S_k$$

其中：

- $\mathbf{r} \in \Delta^{n-1}$：路由分布向量（位于概率单纯形）
- $\Sigma \subset [n]$：激活的专家集合，满足 $|\Sigma| = k$（top-k 稀疏门控）
- $\pi \in S_k$：激活顺序（排列），表示 Σ 中专家按路由权重降序排列

为避免记号歧义，我们明确区分：Softmax$(\cdot)$ 表示函数，Σ 表示激活集合，π 表示排列。

3.2 可用信息维度

路由状态空间提供了三个主要的信息维度用于编码水印：

维度 **1**：激活模式（组合码）

从 n 个专家中选择 k 个激活，信息容量为：

$$I_{\text{pattern}} = \log_2 \binom{n}{k} = \log_2 \frac{n!}{k!(n-k)!}$$

对于 $n = 128, k = 2$ 的典型配置，使用斯特林近似（Stirling's approximation）：

$$I_{\text{pattern}} \approx 13 \text{ bits}$$

维度 **2**：排列顺序（排列码）

对前 k 个专家的排列，信息容量为：

$$I_{\text{order}} = \log_2(k!)$$

对于 $k = 4$：

$$I_{\text{order}} = \log_2(24) \approx 4.6 \text{ bits}$$

注意：该容量仅在" 前 k 专家可完全排序且可控" 的假设下成立。若路由器含温度缩放或分数权重相近导致排序不稳定，应引入有限分辨率/随机噪声的" 有效容量" 修正。

维度 **3**：权重量化（连续码）

由于 $\mathbf{r} \in \Delta^{n-1}$ 且满足 $\sum_i r_i = 1$，自由度为 $(n-1)$。在 top-k 稀疏门控下，仅对激活的 k 个专家权重进行量化，且满足 $\sum_{i \in \Sigma} r_i = 1$，因此自由度为 $(k-1)$。假设每个路由权重 r_i（$i \in \Sigma$）量化为 b-bit 精度：

$$I_{\text{weight}} \leq (k-1) \cdot b \text{ bits}$$

对于 $k = 4, b = 4$（16 级量化）：

$$I_{\text{weight}} \leq 12 \text{ bits}$$

考虑温度 T 与归一化噪声 σ 的影响，有效容量需乘以系数 $\eta(T, \sigma) \in (0, 1)$：

$$I_{\text{weight}}^{\text{eff}} = \eta(T, \sigma) \cdot (k-1) \cdot b$$

总容量上界：

$$\begin{aligned} C_{\max} &= I_{\text{pattern}} + I_{\text{order}} + I_{\text{weight}}^{\text{eff}} \\ &= \log_2 \binom{n}{k} + \log_2(k!) + \eta(T, \sigma) \cdot (k-1) \cdot b \end{aligned} \tag{1}$$

4 信息容量的量化

4.1 可实现容量（**Achievable Capacity**）与编码本视角

从编码本（Codebook）视角，对每个水印消息 $\mathbf{m}$ 映射为一组偏置向量或门控参数。信道模型为" 干净路由分布 $\mathbf{r} \to$ 受嵌入扰动 $\mathbf{r}' \to$ 攻击与失真 $\mathbf{r}'' \to$ 验证统计 Y"。

在约束模型性能的条件下，可实现的容量定义为：

$$C_{\text{achievable}} = \max_{p(\mathbf{m}), p(\mathbf{r}'|\mathbf{m})} I(\mathbf{m}; Y)$$

其中 Y 为可观测特征（如 (Σ, π) 或归一化权重），$I(\cdot;\cdot)$ 为互信息。

4.2 率失真（**Rate-Distortion**）问题

将" 嵌入强度 δ" 与" 失真 D" 联立，定义率失真函数：

$$\begin{aligned} R(D) = \max_{p(\mathbf{m}), p(\mathbf{r}'|\mathbf{m})} & I(\mathbf{m}; \mathcal{S}(\mathbf{r}')) \\ \text{s.t.} \quad & \mathbb{E}[d(\mathbf{r}', \mathbf{r})] \leq D, \quad \mathbf{r}' \in \Delta^{k-1} \end{aligned} \tag{2}$$

其中失真度量 $d(\mathbf{r}', \mathbf{r})$ 可以是：

- Kullback-Leibler 散度：$D_{\text{KL}}(\mathbf{r}' || \mathbf{r}) = \sum_i r_i' \log \frac{r_i'}{r_i}$
- Rényi 散度：$D_\alpha(\mathbf{r}' || \mathbf{r}) = \frac{1}{\alpha - 1} \log \sum_i (r_i')^\alpha (r_i)^{1-\alpha}$

在温度缩放与 top-k 稀疏约束下，$R(D)$ 可通过拉格朗日乘子法求解最优嵌入分布 $p(\mathbf{r}'|\mathbf{m})$，并给出可计算的上下界。

信息-性能权衡曲线：

$$\mathcal{T}(\epsilon) = \{C : \exists D \text{ s.t. } \Delta_{\text{perf}} \leq \epsilon \text{ and } R(D) = C\}$$

其中 $\Delta_{\text{perf}} = \text{Acc}_{\text{clean}} - \text{Acc}_{\text{watermarked}}$ 表示性能下降，ϵ 为性能降级门限。

5 鲁棒性的形式化定义

5.1 对抗性攻击下的鲁棒性（Min-Max 定义）

设攻击为随机变换 $\mathcal{A}:\mathcal{S}\to\mathcal{S}'$，攻击族 $\mathcal{A}$ 包含：

- 微调（fine-tune）：路由器单独/联合微调
- 蒸馏（distill）：蒸馏到无水印模型
- 专家重排/剪枝（expert reindex/prune）
- 温度/门控策略变更：top-k → top-1/2 切换
- 输入域漂移：输入 Noising/域外测试集
- 量化/低比特化
- LoRA 注入、梯度手术（冻结路由层）等

鲁棒性定义（极小极大准则）：

$$\mathcal{R}=\inf_{\mathcal{A}\in\mathcal{A}}\Pr\left[\mathrm{Verify}(f_{\text{watermarked}},\mathcal{A})=1\right]$$

其中 $\mathrm{Verify}(f,\mathcal{A})=1$ 表示在攻击 $\mathcal{A}$ 后仍能成功验证。我们区分随机对手（从 $\mathcal{A}$ 中随机选择攻击）与自适应对手（根据验证器策略选择最优攻击）。

5.2 鲁棒性的下界（基于 Le Cam/Pinsker 方法）

对于独立同分布的攻击，水印的鲁棒性下界基于 **Le Cam** 两点法或信息不等式（**Pinsker/Van Trees**）给出可检性（distinguishability）下界。

设攻击后的路由状态为 $\mathcal{S}'(x)=\mathcal{A}(\mathcal{S}(x))$，水印消息空间大小为 $|\mathcal{M}|$。基于 **Pinsker** 不等式，总变分距离（Total Variation）与 KL 散度的关系为：

$$\mathrm{TV}(p_w,p_c)\le\sqrt{\frac{1}{2}D_{\mathrm{KL}}(p_w||p_c)}$$

其中 $p_w=p_{\text{watermarked}}$ 和 $p_c=p_{\text{clean}}$ 分别为带水印和干净模型的路由分布。

可检性下界：若攻击后的分布满足 $D_{\mathrm{KL}}(p_w||p_c)\ge\delta$，则验证成功率下界为：

$$\mathcal{R}\ge 1-\exp\left(-\frac{\delta}{2}\right)-\alpha$$

其中 α 为假阳性率上界。该下界更贴近" 验证器" 的性质，而非直接使用 Fano 不等式。

5.3 对特定攻击的鲁棒性量化

攻击 **1**：路由器重训练

设路由器参数的扰动为 $\Delta\mathbf{w},\Delta b$，考虑 Kullback-Leibler 散度：

$$D_{\mathrm{KL}}(\mathbf{r}||\mathbf{r}')=\sum_i r_i\log\frac{r_i}{r_i'}\le\beta$$

鲁棒性与 β 的关系：

$$\mathrm{Rob}_{\text{retrain}}=1-\exp(-\lambda D_{\mathrm{KL}})$$

其中 λ 为编码的纠错码强度参数。

攻击 **2**：模型蒸馏

设蒸馏温度为 T，学生模型的路由分布为 $\mathbf{r}_s$：

$$\mathrm{Rob}_{\text{distill}}=\exp\left(-\frac{D_{\mathrm{KL}}(\mathbf{r}^T||\mathbf{r}_s^T)}{H(\mathbf{r})}\right)$$

其中 $H(\mathbf{r})$ 为路由分布的熵。

攻击 **3**：量化与剪枝

对于 q-bit 量化：

$$\mathrm{Rob}_{\text{quant}}=\left(1-\frac{2^{-q}}{2}\right)^{n\cdot I_{\text{weight}}}$$

6 验证过程与检测能力

6.1 假设检验框架（复合假设）

验证问题可形式化为复合假设检验：

- H_0：模型未被水印化（原始模型），参数 $\theta_0\in\Theta_0$ 未知
- H_1：模型被正确的水印化，参数 $\theta_1\in\Theta_1$ 未知且受攻击扰动

由于参数未知且受攻击扰动，H_0 和 H_1 均为复合假设。给定测试集 $\mathcal{T}=\{x_1,\dots,x_N\}$，提取特征向量：

$$\mathbf{f}=[\mathcal{F}(x_1),\dots,\mathcal{F}(x_N)]$$

其中 $\mathcal{F}$ 为特征提取函数。特征可以是：

- 仅激活集合：$\mathbf{f}=(\Sigma_1,\dots,\Sigma_N)$
- 激活集合 + 排列：$\mathbf{f}=((\Sigma_1,\pi_1),\dots,(\Sigma_N,\pi_N))$
- 激活集合 + 排列 + 权重：$\mathbf{f}=((\Sigma_1,\pi_1,\mathbf{r}_{\Sigma_1}),\dots,(\Sigma_N,\pi_N,\mathbf{r}_{\Sigma_N}))$

需要进行充分统计量分析，确定哪些特征组合是充分的。

6.2 广义似然比检验（GLRT）与序贯概率比检验（SPRT）

由于 H_0 和 H_1 为复合假设，采用广义似然比检验（**GLRT**）以消去未知参数：

$$\Lambda_{\text{GLRT}}(\mathbf{f}) = \frac{\max_{\theta_1 \in \Theta_1} p(\mathbf{f}|\theta_1)}{\max_{\theta_0 \in \Theta_0} p(\mathbf{f}|\theta_0)} \gtrless \tau$$

或采用序贯概率比检验（**SPRT**）以降低测试样本开销：

$$\Lambda_{\text{SPRT}}(\mathbf{f}_t) = \prod_{i=1}^{t} \frac{p(\mathbf{f}_i|H_1)}{p(\mathbf{f}_i|H_0)} \gtrless \tau$$

其中 t 为当前样本数，SPRT 可在达到决策阈值时提前终止。

检测功率（功效）定义为：

$$\beta = \inf_{\theta_1 \in \Theta_1} P(\Lambda > \tau | H_1, \theta_1)$$

假阳性率（False Positive Rate）：

$$\alpha = \sup_{\theta_0 \in \Theta_0} P(\Lambda > \tau | H_0, \theta_0)$$

6.3 ROC 曲线与 AUC 下界（基于错判率上界）

检测质量由接收者操作特征（ROC）曲线量化：

$$\text{AUC} = \int_0^1 \text{TPR}(\text{FPR})\, d(\text{FPR})$$

对于信息论上界，我们采用 **Chernoff/Bhattacharyya** 上界先给出错判率或 **Bayes** 风险的指数型上界，再间接推导对 AUC 的保守下界。

Bhattacharyya 距离：

$$D_B = -\ln \int \sqrt{p(\mathbf{f}|H_0) p(\mathbf{f}|H_1)}\, d\mathbf{f}$$

对于高斯分布族，错判率上界为：

$$P_{\text{error}} \leq \exp(-D_B)$$

进而可推导 AUC 的保守下界（需注明适用分布族，如高斯族）：

$$\text{AUC} \geq 1 - \exp(-D_B) - \alpha$$

其中 α 为假阳性率上界。

7 综合性能指标

7.1 水印质量函数与帕累托前沿

综合定义水印系统的质量为三维指标：

$$\mathcal{Q} = (C_{\text{achievable}}, \mathcal{R}, \text{AUC})$$

其中 $C_{\text{achievable}}$ 为可实现容量，$\mathcal{R}$ 为鲁棒性（min-max 定义），AUC 为检测 AUC。

考虑多目标优化：

$$\max\{C_{\text{achievable}}, \mathcal{R}, \text{AUC}\}$$

$$\text{s.t.} \quad \Delta_{\text{perf}} \leq \epsilon, \quad \alpha \leq \alpha_{\max}$$

帕累托前沿定义为非被支配的解集合，即不存在其他解在所有目标上都不劣于当前解且至少在一个目标上更优。

权重设定：质量函数的加权形式 $Q_{\text{normalized}} = \alpha C_n + \beta R_n + \gamma A_n$（其中 $\alpha + \beta + \gamma = 1$）仅用于政策选择（根据具体应用需求选择折中点），而非理论结论。权重来源与任务依赖需明确：不同数据集/任务（如生成式 vs 分类式 MoE）可能需要不同的权重配置。建议以 **Pareto** 前沿为主图呈现不加权的解集，再由应用需求选择折中点。

8 数学推导示例

8.1 示例：基于激活模式的容量计算

假设设计仅基于激活模式 Σ 进行水印：

信息来源：从 n 个专家中选择 k 个激活

$$I_{\text{pattern}} = \log_2 \binom{n}{k} = \log_2 \frac{n!}{k!(n-k)!}$$

编码方案：使用修改路由权重使特定专家组合优先激活

$$\tilde{\mathbf{w}} = \mathbf{w} + \delta \mathbf{b}_{\text{target}}$$

其中 $\mathbf{b}_{\text{target}}$ 为目标激活模式 Σ_{target} 的偏置向量。

性能约束：KL 散度限制

$$D_{\text{KL}}(\mathbf{r}_{\text{clean}} || \mathbf{r}_{\text{watermarked}}) \leq \delta_{\max}$$

其中 $\mathbf{r}_{\text{clean}}$ 和 $\mathbf{r}_{\text{watermarked}}$ 分别为干净和带水印模型的路由分布。

鲁棒性下界：对路由器重训练攻击，基于 Pinsker 不等式，若攻击后的 KL 散度满足 $D_{\text{KL}}(\mathbf{r}' || \mathbf{r}_{\text{watermarked}}) \leq \beta$，则检测成功率下界为：

$$P_{\text{detect}} \geq 1 - \exp\left(-\lambda D_{\text{KL}}\right) - \alpha$$

其中 λ 为编码的纠错码强度参数，α 为假阳性率上界。

9 开放的理论问题

尽管本文建立了 MoE 路由水印的信息论框架，但仍存在一些开放的理论问题：

1. 稀疏门控与单纯形容量的精确化：在 top-k 稀疏门控与概率单纯形约束下，路由水印的有效自由度与容量是多少？如何用类型法（method of types）为离散部分给出可靠编码区与错误指数？

2. 嵌入-性能的率失真闭环：在给定性能降级门限下，能否得到显式的容量上/下界？如何将性能损失上界映射为对路由分布的散度球约束（KL 或 Rényi 散度）？

3. 复合假设检验的稳健验证器：如何构造稳健最优的验证器以处理攻击与域漂移？如何比较以 Σ、π、$\mathbf{r}_\Sigma$ 为特征的不同充分性与功效？

4. 隐蔽性（**Steganographic Security**）与可检性下界：路由分布的统计隐蔽性如何量化？如何定义对手的检测器族并给出最小可辨性下界？

5. 联合攻击：多种攻击同时进行时的鲁棒性如何分析？是否存在攻击之间的协同效应？

10 结论

本文提出了 MoE 路由水印的信息论形式化框架，为路由水印的设计和评估提供了严格的理论基础。我们定义了路由状态空间 $\mathcal{S}(x) = (\mathbf{r}, \Sigma, \pi)$，量化了三个信息维度的编码容量（考虑了概率单纯形约束和温度/噪声影响），建立了率失真框架以分析可实现容量与性能权衡。采用极小极大准则，我们定义了鲁棒性并基于 Le Cam/Pinsker 方法推导了可检性下界。在验证过程中，我们构建了基于复合假设检验的框架（GLRT/SPRT），并基于错判率上界间接推导了检测质量的 AUC 下界。这些理论结果为 MoE 路由水印的实践应用提供了指导，并为未来的研究指明了方向，包括稀疏门控与单纯形容量的精确化、嵌入-性能的率失真闭环、复合假设检验的稳健验证器以及隐蔽性分析等开放问题。