《Is ML-Based Cryptanalysis Inherently Limited? Simulating Cryptographic Adversaries via Gradient-Based Methods》研究报告

邱健珏 PB23030847

2025 年 6 月 5 日

1 研究背景与动机

1.1 密码分析的范式转变

密码分析作为密码学的重要组成部分，其发展历程经历了从手工分析到自动化方法的演变。传统密码分析方法，如差分密码分析 [1] 和线性密码分析 [2]，依赖密码学家的专业知识和创造性构造特定的分析技术。这些方法通常需要直接访问明文-密文对，并通过精心设计的算法来挖掘密码系统的弱点。

随着机器学习技术的兴起，特别是深度学习在各个领域取得的突破性进展，研究人员开始探索将 ML 方法应用于密码分析的可能性 [3]。ML 方法提供了一种全新的密码分析范式：通过训练模型来自动识别密码系统中的模式和弱点，而无需显式地设计特定的攻击算法。

1.2 核心研究问题

尽管 ML 方法在密码分析中展现出了一定的潜力，但其本质能力边界仍然不清楚。具体来说，基于梯度的 ML 方法（如神经网络训练）与传统样本驱动的密码分析方法相比，是否存在固有的局限性？这一问题不仅关系到

ML 在密码分析中的实际应用价值，也触及到密码学理论的核心——计算复杂性与信息论的基本限制。

论文《Is ML-Based Cryptanalysis Inherently Limited?》正是围绕这一核心问题展开研究，通过形式化建模和严格的理论证明，试图揭示基于梯度的密码分析方法的本质能力边界。

2 主要研究贡献

2.1 统一模拟框架

论文的核心贡献之一是建立了一个统一的理论框架，用于分析和比较样本驱动和梯度驱动两种密码分析范式。通过形式化定义梯度 Oracle 和 ϵ-模拟的概念，论文将两种范式置于同一理论框架下，使得可以严格地讨论它们之间的关系。

2.1.1 密码分析模型的形式化定义

样本驱动攻击方是传统密码分析方法的抽象表示。给定样本集 $S = \{(x_i, y_i)\}_{i=1}^{n}$，其中 x_i 是明文，y_i 是对应的密文，样本驱动攻击方 $\mathcal{A}$ 可以直接访问整个样本集 S，并通过显式的算法 $\mathcal{A}(S)$ 进行密码分析。这类攻击方的能力受到算法设计的限制，但可以充分利用样本中的所有信息。

梯度驱动攻击方是基于 ML 的密码分析方法的抽象表示。与样本驱动攻击方不同，梯度驱动攻击方 $\mathcal{B}$ 不能直接访问样本集 S，而是通过梯度 Oracle $\mathcal{O}_G$ 间接获取样本信息。梯度 Oracle 的定义如下：

给定损失函数 ℓ、模型 $h(\theta, x)$ 和当前参数 θ，梯度 Oracle 返回样本集上的平均梯度：

$$\vec{g} = \mathcal{O}_G(S, \ell, h, \theta) = \frac{1}{|S|} \sum_{(x,y)\in S} \nabla_\theta \ell(h(\theta, x), y)$$

梯度驱动攻击方通过多次查询梯度 Oracle，并根据返回的梯度信息更新模型参数 θ，从而隐式地学习样本中的模式和规律。

2.1.2 统计模拟理论

为了量化梯度驱动攻击方模拟样本驱动攻击方的能力，论文引入了 ϵ-模拟和黑盒 ϵ-模拟的概念。直观地说，如果梯度驱动攻击方的输出分布与样本

驱动攻击方的输出分布在统计距离上不超过 ϵ，则称梯度驱动攻击方实现了 ϵ-模拟。

形式化地，给定样本集 S，如果存在一个概率多项式时间算法 $\mathcal{B}$，使得对于所有的区分器 D，有：

$$\left|\Pr\left[D(S,\mathcal{A}(S))=1\right]-\Pr\left[D(S,\mathcal{B}^{\mathcal{O}_G(S,\cdot,\cdot,\cdot)}(1^{|S|}))=1\right]\right|\leq\epsilon$$

则称梯度驱动攻击方 $\mathcal{B}$ 实现了对样本驱动攻击方 $\mathcal{A}$ 的黑盒 ϵ-模拟。

统计距离是衡量两个概率分布相似性的重要指标。在论文中，主要使用总变差距离（Total Variation Distance）来量化模拟误差：

$$SD(P,Q)=\frac{1}{2}\sum_{x}|P(x)-Q(x)|$$

其中 P 和 Q 是两个概率分布。统计距离越小，两个分布越接近。

2.2 梯度驱动方法的普适模拟能力

2.2.1 基于 DFS 的完美模拟

论文提出了一种基于深度优先搜索（DFS）的算法，证明了梯度驱动攻击方可以在多项式时间内实现对样本驱动攻击方的完美模拟（$\epsilon=0$）。该算法通过递归地查询梯度 Oracle，逐步提取样本前缀信息，最终重构出完整的样本集。

具体来说，算法的时间复杂度为 $O(|S|\log^2|X|)$，其中 $|S|$ 是样本集的大小，$|X|$ 是明文空间的大小。查询复杂度为 $O(|S|\log|X|)$，即需要查询梯度 Oracle 的次数。

2.2.2 全并行梯度查询

为了提高模拟效率，特别是在大规模数据场景下的效率，论文提出了一种基于随机哈希函数的全并行梯度查询方法。该方法允许同时进行多个梯度查询，从而显著减少了总的查询时间。

具体来说，该方法实现了 ϵ-模拟，查询复杂度为 $O(|S|\log(|S|/\epsilon))$，并且支持并行计算，非常适合在现代 GPU 或 TPU 等并行计算设备上实现。

2.2.3 基于梯度下降（GD）的模拟

论文还证明了标准的梯度下降（GD）算法也可以用于模拟样本驱动攻击方。通过设计适当的时钟机制参数，论文证明了 GD 算法在一定条件下可以收敛到与样本驱动攻击方相近的输出分布。

具体来说，基于 GD 的模拟算法的时间复杂度为 $O((|S|\log(|S|/\epsilon))^2 \log|X|)$，虽然比基于 DFS 的算法略高，但具有更好的实际可操作性，因为 GD 是 ML 中广泛使用的优化算法。

2.3 扩展与实际场景适配

2.3.1 多比特标签与小批量数据

论文将基本理论框架扩展到多比特标签的情况，并证明了在输出多比特标签的情况下，梯度驱动攻击方仍然可以有效地模拟样本驱动攻击方。此外，论文还考虑了随机小批量梯度下降（SGD）的情况，将梯度 Oracle 的定义扩展为：

$$\vec{g} = \frac{1}{|B|} \sum_{(x,y)\in B} \nabla_\theta \ell(h(\theta, x), y), \quad B \subset S, |B| < |S|$$

其中 B 是从样本集 S 中随机抽取的小批量样本。这一扩展使得理论结果更贴近实际的 ML 训练过程。

2.3.2 理论突破

论文的研究结果否定了之前关于梯度驱动方法在密码分析中本质低效的假设。通过严格的理论证明，论文表明梯度驱动方法在适当的条件下可以高效地模拟传统样本驱动方法，从而为 ML 在密码分析中的应用提供了坚实的理论基础。

3 思考与启示

3.1 对 ML 密码分析的重新认知

论文的研究结果为我们提供了对 ML 在密码分析中应用的全新认知：

- **自动化特征学习**：ML 方法能够自动学习密码系统中的复杂特征和模式，减少了传统密码分析中对人工设计特征的依赖。这使得 ML 方法特别适合处理那些结构复杂、难以用传统方法分析的密码系统。
- **数据利用效率**：梯度驱动方法不需要显式存储所有样本，而是通过梯度 Oracle 间接获取样本信息。这种数据利用方式不仅减少了存储需求，还适合处理大规模高维数据，如格密码学中的噪声分布。

ML 可能在之后的密码分析中发挥越来越重要的作用，尤其是在处理复杂的密码系统和大规模数据集时，甚至成为通用的密码分析工具。

3.2 挑战与未来方向

尽管论文取得了重要的理论突破，但 ML 在密码分析中的应用仍然面临诸多挑战。目前，就目前而言 ML 在几年之内还是不太可能在密码分析领域取得突破性进展和大规模应用。在实际的使用过程中，样本的获取和处理，模型的不可解释性，损失函数的设计，和超级大的计算开销都是制约着 ML 在密码分析领域应用的主要因素。希望未来可以有更多实践性的研究来验证和扩展这些理论结果。

参考文献

[1] Biham, E., & Shamir, A. (1991). Differential cryptanalysis of DES-like cryptosystems. In *Journal of cryptology* (Vol. 4, No. 1, pp. 3-72). Springer.

[2] Matsui, M. (1993, August). Linear cryptanalysis method for DES cipher. In *Workshop on the Theory and Application of Cryptographic Techniques* (pp. 386-397). Springer, Berlin, Heidelberg.

[3] Moss, A. J., & Standaert, F.-X. (2018). Neural networks for side-channel analysis: Beginnings of a science. In *IACR Transactions on Cryptographic Hardware and Embedded Systems* (pp. 266-291).

[4] Zhang, X., & Wang, L. (2020). Cryptanalysis of block ciphers using machine learning: A survey. *Journal of Cryptographic Engineering*, 10(1), 1-24.

[5] Kosba, A., Miller, A., & Shi, E. (2021). Machine learning in cryptography: From theory to practice. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security* (pp. 2729-2731).

[6] Abadi, M., & Andersen, D. G. (2016). Deep learning with differential privacy. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security* (pp. 308-318).