

## A Zero-Relationship Lattice Algorithm for Preventing AI-Based Secret-Key Recovery Attacks on Lattice Cryptography

### [Abstract]

Recently, attack techniques that exploit AI-based pattern analysis, dimensionality reduction, and statistical imbalance have revealed that lattice-based cryptography is not sufficiently secure when it relies solely on mathematical hardness. This paper proposes a zero-relation lattice algorithm designed to thwart AI-driven secret key recovery attacks on lattice-based cryptosystems. The proposed method applies a Maclaurin-series-based keyed polynomial obfuscation function to both the secret key and the plaintext encoding in the encryption process, transforming the linear traces remaining in the ciphertext into nonlinear distortions. Since the obfuscation coefficients are randomly derived from a secret parameter  $k$  for each key, the statistical structure observed by an adversary concentrates on an obfuscated hallucination key  $s' = f_k(s)$  rather than on the original secret key  $s$ . As a result, Transformer-based learning attacks such as SALSA converge to  $s'$  instead of  $s$ , and without knowledge of the secret parameter  $k$  it becomes computationally difficult to efficiently infer  $s$  from  $s'$ . The plaintext is further protected by an additional masking mechanism that depends on  $k$ , so that message confidentiality is preserved even when both the ciphertext and the obfuscated secret key are exposed. Experimental results show that the proposed scheme reduces the success rate of AI-based pattern-analysis key recovery attacks by more than 70% and degrades the effectiveness of attacks that exploit dimensionality reduction and statistical imbalance, thereby achieving improved security compared with conventional lattice-based cryptosystems.

▶ Key words : lattice-based cryptography, post-quantum security, AI-driven security attacks, data relationality

### [요 약]

최근 인공지능 기반 패턴 분석, 차원 감소 및 통계적 불균형을 활용한 공격 기법들이 제안되면서, 격자 기반 암호가 수학적 난해성만으로는 충분히 안전하지 않다는 문제가 제시되었다. 본 논문은 인공지능 기반 격자 암호 비밀키 복구 공격을 방지하기 위한 영관계 격자 알고리즘을 제안한다. 제안 기법은 격자 기반 암호화 과정에서 비밀키와 평문 인코딩에 맥클로린 급수 기반의 키드 다행 난독화 함수를 적용하여, 암호문에 남는 선형적인 흔적을 비선형 꼬임으로 변환한다. 난독화 계수는 각 키마다 비밀 파라미터  $k$ 에서 무작위로 유도되므로, 공격자가 관측하는 통계적 구조는 원래 비밀키  $s$ 가 아니라 난독화된 hallucination key  $s' = f_k(s)$ 에 집중시킨다. 이를 통해 SALSA와 같은 트랜스포머 기반 학습 공격은  $s$  대신  $s'$ 에 수렴하며, 비밀 파라미터  $k$  없이는  $s'$ 로부터  $s$ 를 효율적으로 추정하기 어렵게 만든다. 평문은  $k$ 에 의존하는 추가 마스킹을 통해 보호되므로, 암호문과 난독화된 비밀키가 노출되더라도 평문 보안성을 유지할 수 있다. 실험 결과, 제안 기법은 인공지능 기반 패턴 분석에 의한 비밀키 복구 성공률을 70 % 이상 낮추고, 차원 감소와 통계적 불균형을 이용하는 공격의 성능을 약화시켜 기존 격자 기반 암호의 기법 대비 향상된 보안성을 보였다.

▶ 주제어 : 격자 기반 암호, 양자 내성 보안, 인공지능 기반 보안 공격, 데이터 관계성

## I. Introduction

양자 컴퓨팅 기술의 발전은 기존 공개키 암호 체계의 보안 가정을 근본적으로 위협하고 있다. 대표적으로 shor 알고리즘은 정수 소인수분해와 이산 로그 문제를 다행 시간에 해결할 수 있음을 보였다 [1]. 이는 Rivest Shamir Adleman (RSA), Digital Signature Algorithm (DSA), Elliptic Curve Digital Signature Algorithm (ECDSA) 등 전통적인 수학적 난제에 기반한 공개키 암호가 향후 양자 컴퓨팅 환경에서 장기적 안전성을 보장할 수 없음을 의미한다 [2]. 따라서, 양자 컴퓨팅 환경에서도 강력한 보안성을 제공할 수 있는 양자 내성 암호에 관한 연구가 활발히 진행되고 있으며 [3], 그 중 격자를 기반으로 한 암호 기법들은 평균-최악 환원을 바탕으로 이론적 안정성을 갖추고 있어 가장 유력한 차세대 보안 후보군으로 평가된다 [4][5]. 격자 기반 암호의 핵심 난제인 Learning with Errors (LWE)와 Ring LWE (RLWE)는 고차원 격자 위에서의 근사 최단 벡터 문제나 근사 가장 가까운 벡터 문제의 난이도에 기반한다 [8]. LWE 문제는 비밀 벡터  $s \in \mathbb{Z}_q^n$  와 가우시안 잡음  $e$ 가 섞인 선형 샘플  $(a, \langle a, s \rangle + e)$ 로부터  $s$ 를 복구하는 문제로 정의되며 [6][7]. RLWE는 이를 다행 환  $R_q = \mathbb{Z}_q[x]/\langle f(x) \rangle$  구조 위로 확장한 형태이다 [9]. 격자 기반 암호는 단순한 이론적 안전성에 머무르지 않고, 클라우드 환경에서의 안전한 키 교환, 분산원장 기술의 보안 강화, 개인정보 보호를 위한 동형 연산 등 다양한 응용 분야에서 실제 구현 및 테스트가 활발히 진행되고 있다. RLWE는 구조적 효율성과 강력한 수학적 기반 덕분에 National Institute of Standards and Technology (NIST)의 양자 내성 암호 표준화 과정에서 주요 후보군으로 채택되었으며, 동형암호 · 키 캡슐레이션 · 디지털 서명 등 다양한 응용에 활용되고 있다 [10]. 실제로 Kyber, Dilithium과 같은 RLWE 기반 알고리즘이 차세대 양자 내성 암호의 최종 후보로 채택되면서 [11], 격자 기반 암호는 양자 위협 방어뿐 아니라 차세대 디지털 인프라 전반의 보안성과 활용성을 뒷받침하는 범용 기술로서 중요성이 점점 커지고 있다 [12].

그러나, 최근 연구들에서 격자 기반 암호가 단순

히 수학적 난제에 의존하는 것만으로는 충분히 안전하지 않음으로 지적하고 있다. 특히, 구현 과정에서 노출되는 구조적 패턴이나 통계적 불균형을 통해 새로운 취약점이 발생할 수 있음을 보여주고 있다 [13]. Wenger et al.은 Soft-Attentional Lattice Secret Attacker (SALSA) 연구에서 LWE 및 RLWE 기반 공개키 및 암호문 샘플을 대규모 데이터셋으로 간주하고 트랜스포머 기반 신경망을 학습하여, 전통적인 격자 감소 없이도 비밀키를 높은 정확도로 복구할 수 있음을 보였다 [14]. 또한, 저자들의 추가적인 연구 결과인 SALSA-FRESCA에서는 Angular embedding과 사전학습 기법을 적용하여 학습 효율과 공격 성능을 더욱 개선하여 격자 기반 암호의 취약성을 입증하였다 [15].

차원 감소와 분산 불균형을 이용한 Cool&Cruel은 비밀키의 hard bits와 easy bits로 분리함으로써, 일부 좌표의 키 비트를 통계적으로 유리한 방식으로 추출할 수 있음을 보였다 [16]. 이와 같은 인공지능 기반 격자 암호 공격들은, 격자 감소 알고리즘의 한계와는 별도로 데이터 구동형 패턴 분석만으로도 실용적인 공격이 가능하다는 점을 보여주며 [17], 격자 기반 암호 체계의 새로운 위협 모델로 부상하고 있다.

기존 방어 기법 연구들에서는 주로 파라미터를 조정하여 공격 난이도를 증가시키거나, 샘플 수를 제한하고 학습 데이터에 노이즈를 추가하는 등 비교적 수동적인 방법에 초점을 맞추어 왔다 [18][19]. 그러나 이와 같은 방식은 공격자 모델이 진화할수록 방어 효과가 빠르게 저하될 수 있으며, 근복적으로 암호 구성 자체에서 발생하는 통계적 흔적을 어떻게 설계적으로 제어할 것인가라는 문제에 충분히 대응하지 못한다. 현재까지의 많은 격자 기반 암호는 LWE 샘플이 충분히 많더라도, 격자 감소 없이  $s$ 를 직접 학습하기는 어렵다는 직관에 의존해 왔으나, SALSA 공격은 이 직관이 실제 구현 환경에서는 더 이상 안전한 가정이 아닐 수 있음을 보여준다.

본 논문은 인공지능 기반 비밀키 복구 공격에 대응하기 위해 암호문과 비밀키 사이에 형성되는 학습 가능한 상관관계를 효율적으로 최소화하는 영관계 격자 알고리즘을 제안한다. 제안 기법의 핵심 아이디어는 기존 격자 기반 암호에서 비밀키를 단일 벡터  $s$ 로 두는 대신, 비밀키를  $(s, k)$ 의 쌍으로 확장하고,

맥클로린 급수 기반의 키 의존적인 비선형 난독화 함수  $f_k$ 를 적용한다. 키 생성 시 비밀 파라미터  $k$ 에서 난독화 계수  $a_i(k)$ 와 전개 지수 집합  $I(k)$ 를 무작위로 유도하고, 이를 통해

$$s' = f_k(s) = \sum_{i \in I(k)} a_i(k)s^{*i}$$

형태의 난독화된 hallucination key  $s'$ 를 정의한다. 공개키 성분은 기존의  $b = As + e$  대신  $b' = Af_k(s) + e$ 로 대체되며, 평문 인코딩 과정에서도  $k$ 에 의존하는 추가 마스킹을 적용한다. 이를 통해 공격자가 관찰 및 학습하는 통계적 구조에서는  $s$ 가 아니라 난독화된  $s' = f_k(s)$ 에 결합되도록 설계하고, 공개 정보와  $s$  사이의 직접적인 선형 및 통계적 상관관계를 최대한 제거한다. 제안하는 기법에서 SALSA와 같은 인공지능 기반 공격은 충분한 샘플을 학습하더라도 원래  $s$ 가 아닌  $s'$  방향으로 수렴시킨다. 본 논문에서 제안하는 기법의 보안 목표는, 공격자가  $s'$ 를 상당히 정확하게 회수하더라도,  $s'$ 와 공개 정보를 바탕으로 비밀 파라미터  $k$ 와  $s$ 를 함께 복구하는 문제는 고차원 역상 탐색 문제로 남아 현실적으로 해결이 어렵도록 설계하는 것이다. 또한 평문은  $k$ 에 의존하는 마스크를 통해 한 번 더 보호되므로, 암호문과 난독화된 비밀키가 노출되더라도 평문 기밀성이 유지될 수 있도록 한다. 이를 통해 인공지능 기반 격자 암호 비밀키 복구 공격을 암호 구성 차원에서 방지하기 위한 새로운 영관계 격자 알고리즘을 제안한다.

## II. Related works

본 장에서는 격자 기반 암호에 구조적 취약점을 제기한 최신 공격 기법들을 살펴본다. 표 1은 격자 기반 암호의 관련 연구들을 보여준다. 특히 LWE와 RLWE 공개키와 암호문을 대규모 데이터셋으로 간주하여 비밀키를 학습하는 인공지능 기반 비밀키 복구 공격과 차원 감소 이후의 통계적 불균형을 이용해 비밀키를 부분적으로 복구하는 통계적 분리 공격에 초점을 맞춘다.

### 1. Pattern-Based Secret Key Recovery Attacks

Wenger et al.이 제안한 SALSA는 트랜스포머 모델을 이용해 LWE 샘플의 계수 분포에 존재하는 구조적 패턴을 학습하여 비밀키 복구가 실제로 가

능함을 입증하였다 [14]. SALSA는 다수의 LWE 샘플  $a, b = as + e$ 에 대해  $b$ 를 예측하는 모델을 학습한 뒤, 선택된 입력  $a$ 를 넣어 모델이 출력하는 반응을 이용해  $s$ 의 좌표를 추정하고, 잔차 분산 분석을 통해 올바른 후보를 검증하는 과정을 따른다.

Stevens et al.은 SALSA를 확장한 SALSA FRESCA를 제안하였다 [15]. SALSA FRESCA는 공격 프레임워크를 크게 개선하여 각도 임베딩을 이용해 입력 표현을 강화하고, 대규모 사전학습을 통해 다양한 비밀키 분포에 대해 재사용 가능한 초기화 모델을 확보하였다. 그 결과 전처리 단계가 기존보다 25배 가속되고 샘플 효율성은 10배 증가하였다. 차원  $n = 1024$ 의 희소 이진키를 복구하는데 성공하며, 기계학습이 기존에는 불가능하다고 여겨졌던 고차원 격자 문제에까지 실질적 위협을 가할 수 있음을 입증했다.

패턴 기반 인공지능 비밀키 공격 기법 연구는 격자 기반 암호가 단순히 이론적 난해성에 의존할 뿐 아니라, 암호문에 남는 미세한 통계적 흔적이 존재할 경우 인공지능 모델이 이를 빠르게 포착하고 학습하여 보안 취약점 공격에 활용할 수 있음을 입증하였다. 따라서 암호문과 비밀키 사이의 학습 가능한 상관구조를 최소화하거나, 가능하다면 원천적으로 차단하는 설계가 요구된다. 본 논문에서 제안하는 영관계 격자 알고리즘은 상관 구조를 비선형 꼬임으로 변환하여 SALSA와 같은 인공지능 기반 비밀키 복구 공격을 약화시키는 것을 목표로 한다.

### 2. Statistical-Vulnerability-Based Secret Key

#### Recovery Attacks

Nolte et al.은 Cool&Cruel 공격을 통해, 차원 감소 과정을 거친 LWE 행렬의 열 분산이 불균질하게 분포한다는 점을 이용하였다 [16]. 비밀키 좌표를 복구가 어려운 부분 hard bits와 통계적으로 유리한 부분 easy bits로 나누고, 잔차 분산 비교를 통해 순차적으로 복구하는 통계적 분리 공격을 제안하였다. 이외에도 부분 키 누출[13], 작은 에러 분포[12], 힌트 정보를 활용하는 다양한 변형 공격 [19]들이 보고되면서, 격자 기반 암호의 안전성이 수학적 난해성만으로는 충분치 않을 수 있음이 지적되고 있다. 이와 같은 연구들은 암호문과 비밀키 사이에 형성되는 학습 가능한 상관 구조가 존재할 경우, 공격자가 이를 기계학습 또는 통계적 방법으

표 1 Related works

연도	저자	공격 기법	취약 파라미터	특징
2025	Cao et al. [19]	힌트를 이용한 가우시안 소거 기반 격자 구성 및 복구	<ul style="list-style-type: none"> <li>힌트에 의존함.</li> <li>사이드채널/디코딩 오류/추가 정보가 있어야 효율적</li> </ul>	<ul style="list-style-type: none"> <li>충분하면 가우시안 소거와 격자 기법 결합으로 효율적 복구 가능</li> </ul>
2024	E Wenger et al. [15]	SALSA 계열 확장하여 각도 임베딩과 사전학습으로 표현력 · 샘플 효율 개선	<ul style="list-style-type: none"> <li>희소/이진 비밀키 대상</li> <li>확장된 차원</li> </ul>	<ul style="list-style-type: none"> <li>프리트레이닝과 임베딩으로 SALSA 대비 샘플 ·시간 효율이 개선되어 실무 파라미터로의 확장성이 증가함.</li> <li>전처리 · 사전학습 설계에 민감하고 재현성 및 파라미터 의존성 문제가 있을 수 있음.</li> </ul>
2024	N. Nolte et al. [16]	비밀의 일부 비트를 분리해 단계적 복구	<ul style="list-style-type: none"> <li>비밀 분포 및 차원 감소 이후 남아있는 분산 불균형</li> </ul>	<ul style="list-style-type: none"> <li>전처리 후 컬럼별 분산 차이를 이용해 일부 비트부터 복구하고 확장</li> <li>통계적 편차와 격자 리덕션 성능에 크게 의존함.</li> </ul>
2023	May et al. [21]	부분 정보가 충분하면 Lenstra-Lenstra-Lovász lattice basis reduction algorithm(LLL) 등으로도 복구 가능한 경계 제시	<ul style="list-style-type: none"> <li>충분한 수의 정확한 힌트</li> </ul>	<ul style="list-style-type: none"> <li>힌트 유출 모델의 정량적 경계 제공, 힌트 유출 방지의 중요성 강조</li> </ul>
2022	E Wenger et al. [14]	Transformer 기반 ML 공격 LWE/RLWE 샘플의 계수 패턴을 학습해 비밀키를 추정	<ul style="list-style-type: none"> <li>작은/중간 차원(예: <math>n \leq 128-256</math>)</li> <li>희소 이진 비밀키, 비교적 작은 <math>q</math></li> </ul>	<ul style="list-style-type: none"> <li><math>q = 251</math> 등 소규모 모듈러스 및 중간 차원에서 성공 사례 보고 대량 학습 데이터와 연산 자원 필요</li> <li>희소성에 민감하여 엔트로피가 낮은 경우에 특히 위협적</li> </ul>
2020	Dachman-Soled et al. [13]	RLWE의 Number Theoretic Transform (NTT) 좌표 일부 노출 모델	<ul style="list-style-type: none"> <li>NTT 좌표의 구조적/지정된 누출</li> </ul>	<ul style="list-style-type: none"> <li>NewHope 등의 표준 파라미터에 대해 1/4 누출로 전체 복구를 실험적으로 입증</li> <li>실제 공격은 누출 패턴이 구조적일 때에 한정되는 점이 한계</li> </ul>
2017	Martin R. Albrecht et al. [18]	작은/희소 비밀에 특화된 듀얼 격자 공격 변형 및 파라미터 권고	<ul style="list-style-type: none"> <li>binary/ternary 등 작은/희소 비밀 분포</li> <li>small-secret LWE용 (<math>n, q, o</math>) 선택 구간</li> </ul>	<ul style="list-style-type: none"> <li>실무 라이브러리 파라미터 평가에 영향</li> <li>작은/희소 비밀키 사용 주의</li> </ul>
2015	Albrecht et al. [20]	BumKalai-Wasserman algorithm(BKW) 알고리즘의 복잡도 · 샘플 요구량 분석	<ul style="list-style-type: none"> <li>많은 LWE 샘플 수 <math>m</math></li> <li>작은 <math>q</math>와 중간 크기 <math>n</math> 조합에서 BKW 활용</li> </ul>	<ul style="list-style-type: none"> <li>특정 인자에서는 BKW가 실용적일 수 있음을 입증</li> <li>파라미터 설계 시 BKW 관점 고려 필요함.</li> </ul>

로 빠르게 포착하고 비밀키 복구에 활용할 수 있음을 보여준다. 따라서 LWE 및 RLWE와 같은 양자 내성 암호 기법의 설계 과정에서 공개 파라미터와 암호문에 남는 통계적 흔적을 체계적으로 제어 및 완화할 수 있는 방어 메커니즘이 요구된다.

### III. Preliminary

본 장에서는 기존 격자 기반 암호의 문제를 정의하고, 격자 기반 암호의 취약점을 이용한 보안 공격의 위험성을 보인다. 표 2는 논문에서 사용된 표

기법을 보여준다.

### 1. Learning With Errors (LWE)

보안 매개변수를  $\lambda$ 라 할 때, 모듈러스  $q = q(\lambda)$  와 차원  $n = n(\lambda)$ , 그리고 샘플 수  $m = \text{poly}(n)$ 이 주어진다. 비밀 벡터는  $s \in \mathbb{Z}_q^n$ 에서 선택되고, 공개 행렬  $A \in \mathbb{Z}_q^{m \times n}$ 은  $\mathbb{Z}_q^{m \times n}$ 에서 균등 분포로 임의 샘플링 된다. 암호문 벡터  $b \in \mathbb{Z}_q^m$ 은 잡음 벡터  $e \leftarrow \chi^m$ 에 대해 아래와 같이 정의 된다.

표 2 Notations

기호	설명
$\lambda$	보안 매개변수
$q$	모듈러스 (정수 량 크기)
$n$	비밀 벡터 차원
$m$	샘플 개수 ( $\text{poly}(n)$ )
$\chi$	LWE 오차 분포
$\mathbb{Z}_q$	모듈러스 $q$ 에 대한 정수 량
$A \in \mathbb{Z}_q^{m \times n}$	균등 분포에서 샘플링된 공개 행렬
$s \in \mathbb{Z}_q^n$	원래 LWE 비밀 벡터
$e \in \mathbb{Z}_q^m$	$\chi^m$ 에서 선택된 오차 벡터
$b = As + e \pmod{q}$	기존 LWE 암호문 성분
$\mathcal{A}$	확률적 다향 시간 공격자
$\epsilon(\lambda)$	무시할 수 없는 함수
$\text{negl}(\lambda)$	무시할 수 있는 함수
$M_\theta$	학습 모델
$D_s = \{(a_i, b_i)\}_{i=1}^m$	$s$ 에 대해 수집된 LWE 학습 데이터셋
$\epsilon_{ML}(\lambda)$	비밀키 복구 성공확률 하한
$k \in \{0, 1\}^k$	비밀 파라미터
$G(k)$	$k$ 에서 맥클로린 난독화 파라미터를 생성하는 난수 생성기
$I(k)$	맥클로린 전개 지수 집합
$a_i(k)$	맥클로린 난독화 계수
$f_k(s)$	맥클로린 난독화 함수
$s' = f_k(s)$	hallucination key
$b' = As' + e \pmod{q}$	난독화된 비밀키에 결합된 공개키 성분
$PK = (A, b')$	공개키

$SK = (s', k)$	비밀키
$m$	평문
$c \in \{0, 1\}^l$	$m$ 의 내부 비트열 표현
$c^* \in \{0, 1\}^l$	$k$ 기반 마스킹이 적용된 인코딩 표현
$\Delta$	메시지 임베딩 스케일링 인자
$D_{n,q}()$	비트열 인코딩을 $\mathbb{Z}_q$ 영역으로 매핑하는 임베딩 함수
$D_{n,q}^{-1}()$	$\mathbb{Z}_q$ 영역에서 비트열 인코딩으로 복원하는 역임베딩 함수
$r \in \{0, 1\}^m$	암호화 시 사용되는 난수 벡터
$e_1 \in \mathbb{Z}_q$	1차 LWE 성분에 사용되는 오차 벡터
$e_2 \in \mathbb{Z}_q$	2차 LWE 성분에 사용되는 스칼라 오차
$u = A^T r + e_1 \pmod{q}$	암호문 1차 성분
$v$	암호문 2차 성분
$CT = (u, v)$	최종 암호문
$\rho = KDF(k, u)$	복호 시 평문 마스크 제거에 사용되는 키 유도 값
$t = v - \langle u, s' \rangle \pmod{q}$	복호화 과정의 중간 값
$\eta$	복호화 시 남는 잔여 잡음항
$\tilde{c}^*, \tilde{c}, \tilde{m}$	라운딩 및 역임베딩 후 복원된 값을 나타내는 표기

$$b \equiv As + e \pmod{q}, \quad e \leftarrow \chi^m$$

위에서  $\chi$ 는 평균 0, 작은 분산을 갖는 LWE 잡음 분포이다.

Search-LWE 문제에서 공격자의 목표는 주어진  $(A, b)$ 로부터  $s$ 를 직접 복구하는 것이다.

해당 공격의 성공 확률은 아래와 같이 무시할 수 없는 수준이다.

$$\Pr[\mathcal{A}(A, b) = s] \geq \epsilon(\lambda)$$

위에서  $\mathcal{A}$ 는 다향시간 공격자,  $\epsilon(\lambda)$ 는 non-negligible 함수이다.

Decision-LWE에서는  $(A, As + e)$ 와  $(A, w)$  (단,

$w \leftarrow \mathbb{Z}_q^m$ )를 구별하는 문제가 고려된다. 공격자의

어드벤티지는 아래와 같이 표현된다.

$$\Pr[\mathcal{A}(A, As + e) = 1] - \Pr[\mathcal{A}(A, w) = 1] \geq \epsilon(\lambda)$$

실제로는 특정 매개변수 설정에서 이 값이 무시할 수 없는 수준이다.

## 2. Ring Learning With Errors (RLWE)

RLWE는 LWE 문제를 다항식 링 위로 확장한 형태이다. 다항식  $f(x) \in \mathbb{Z}[x]$ 에 대해 정의된 몫 링  $R_q = \mathbb{Z}_q[x]/\langle f(x) \rangle$  위에서 정의된다. 비밀키  $s(x) \in R_q$ 에서 선택되고, 공개 다항식  $a(x) \in R_q$ 는  $R_q$ 에서 균등분포로 샘플링된다. 잡음 다항식  $e(x) \leftarrow \chi_R$  대해 RLWE 샘플은 아래와 같다.

$$b(x) \equiv a(x) \cdot s(x) + e(x) \pmod{q}, e(x) \leftarrow \chi_R$$

위에서  $\chi_R$ 은  $R_q$  위의 오차 분포를 나타낸다.

Search-RLWE 문제에서 공격자는 주어진  $(a_i(x), b_i(x))_{i=1}^m$ 으로부터  $s(x)$ 를 복구하는 것을 목표로 하며, 공격 성공 확률은 다음과 같이 표현된다.

$$\Pr[\mathcal{A}((a_i(x), b_i(x))) = s(x)] \geq \epsilon(\lambda)$$

Decision-RLWE에서는 RLWE 분포 샘플과 균등분포 샘플을 구별하는 문제를 고려하며, 공격자의 어드밴티지는 아래와 같다.

$$|\Pr[A(RLWE samples) = 1] - \Pr[A(Uniform samples) = 1]| \geq \epsilon(\lambda)$$

이론적으로 RLWE는 LWE와 유사한 평균-최악 환원 성질을 가지며 양자 내성을 제공하는 것으로 알려져 있으나, 실제 구현 및 파라미터 설정에 따라 부분 키 노출이나 작은 에러, 구조적 누출에 기반한 공격들이 제안되고 있다.

## IV. Method

본 장에서는 격자 기반 암호문에 남아 있는 흔적이나 패턴을 인공지능 및 통계적 정보를 기반으로 비밀키를 복구하는 공격을 방지하기 위한 영관계 격자 알고리즘을 정의한다.

### 1. Threat Model

본 절에서는 LWE 및 RLWE 기반 암호를 대상으로 하는 위협 모델을 정의한다. 공격자는 앞서 정의한 공개 행렬  $A \in \mathbb{Z}_q^{m \times n}$ , 비밀키  $s \in \mathbb{Z}_q^n$ , 오차  $e \leftarrow \chi^m$ , 암호문  $b = As + e \pmod{q}$  또는 대응하는 RLWE 샘플  $(a(x), b(x))$ 에 접근할 수 있다고 가정

한다. 공격자의 목표는 search-LWE 및 search-RLWE 문제에서  $s$ 를 직접 복구하거나, decision-LWE 및 decision-RLWE 문제에서 LWE 및 RLWE 분포 샘플과 균등 분포를 구별하는 비트를 출력하는 것이다. 이때 공격자의 성공 확률 및 어드밴티지는 방정식 (1)과 방정식 (2)와 같이 정의된다.

$$\Pr[\mathcal{A}(A, b) = s] \geq \epsilon(\lambda) \quad (1)$$

$$|\Pr[\mathcal{A}(A, As + e) = 1] - \Pr[\mathcal{A}(A, w) = 1]| \geq \epsilon(\lambda) \quad (2)$$

방정식 (2)에서  $w \leftarrow \mathbb{Z}_q^m$ 은 균등 분포 샘플이며,  $\epsilon(\lambda)$ 는 보안 매개변수  $\lambda$ 에 대해 무시할 수 없는 확률이다.

본 논문은 고전적인 격자 감소 알고리즘이 아닌 인공지능 모델을 사용하는 공격자를 고려한다. 즉, 공격자는 방정식 (1)과 방정식 (2)를 만족하는 임의의 다향 시간 알고리즘일 수 있으나, 구체적인 인스턴스로서 SALSA와 같은 트랜스포머 기반 학습 공격을 가정한다.

공격 모델은 LWE 샘플을 입력 및 출력 쌍  $(a_i, b_i)$ 로 보는 트랜스포머 기반 인공지능 공격 기법이다. 공격자는 비밀키  $s \in \mathbb{Z}_q^n$ 과 잡음 분포  $\chi$ 에 대해  $b_i = \langle a_i, s \rangle + e_i \pmod{q}$ ,  $e_i \leftarrow \chi$ 를 만족하는 샘플들을 수집하여 데이터셋  $D_s = \{(a_i, b_i)\}_{i=1}^m$ 을 구성한다. 인공지능 기반 모델은 방정식 (3)와 같이 정의되는 트랜스포머 모델  $M_\theta$ 를 학습시켜, 내적  $\langle a, s \rangle$ 에 해당하는 관계를 근사한다.

$$b_i = M_\theta(a_i) \approx \langle a_i, s \rangle + e_i \pmod{q} \quad (3)$$

모델은 방정식 (3)을 만족하도록 학습된다. 학습이 완료된 후, 공격자는 각 비밀키 좌표  $s_j$ 에 민감하게 반응하도록 설계된 특수 입력 벡터  $a^{(j)}$ 들을 모델에 주입하고, 모델 출력  $M_\theta(a^{(j)})$ 를 해석하여 좌표별 후보 값을 얻는다. 이와 같이 얻은 좌표들을 모아 후보 비밀키  $\tilde{s}$ 를 구성한 뒤, 원래 샘플들과의 잔차를 방정식 (4)와 같이 계산한다.

$$r = b - As \pmod{q} \quad (4)$$

방정식 (4)를 통해 후보 비밀키  $\tilde{s}$ 에 대해 잔차의 분산을 계산하고, 충분히 작은 경우 이를 정답으로 채택한다. 인공지능 기반 공격의 성공 확률은 방정식 (5)와 같이 표현된다.

$$\Pr[\mathcal{A} \text{ SALSA}(M_\theta; A, b) = s] \geq \epsilon_{ML}(\lambda) \quad (5)$$

방정식 (5)에서  $\epsilon_{ML}(\lambda)$ 는 보안 매개 변수  $\lambda$ 에 대해 무시할 수 없는 학습 기반 공격 성공률을 의미한다. 방정식 (6)은 제안하는 난독화 구조에서 파라미터  $\theta$ 를 선택하는 과정의 방정식 (6)으로 나타낼 수 있다.  $F$ 는 선택할 수 있는 난독화 함수 혹은 파라미터들의 후보 집합이다. 주어진  $\theta$ 에 대해  $M'_\theta$ 는  $(A, b'_\theta)$ 를 입력으로 사용할 때 공격자가 취할 수 있는 최적의 인공지능 기반 공격 학습 전략을 의미한다.

$$\min_{\theta \in F} \mathbb{E}_{A, e} [\Pr(\mathcal{A}^{SALSA}(M'_\theta; A, b'_\theta) = s | A, e) - 2^{-H_\infty(s)}] \leq negl(\lambda) \quad (6)$$

$\mathbb{E}_{A, e} [\Pr(\mathcal{A}^{SALSA}(M'_\theta; A, b'_\theta) = s | A, e) - 2^{-H_\infty(s)}]$ 은 공개 행렬  $A$ 와 잡음  $e$ 를 랜덤으로 선택했을 때의 평균적인 SALSA 공격자의 비밀키 복구 성공 확률에서, 비밀키의 최소 엔트로피  $H_\infty(S)$ 로부터 정의되는 순수 랜덤 추측 성공 확률  $2^{-H_\infty(S)}$ 를 뺀 값을 나타낸다.  $2^{-H_\infty(S)}$ 는 공격자가 비밀키의 사전분포를 모두 알고 있다는 가정 하에, 가능성이 높은 키를 한번 대입하는 경우 달성할 수 있는 추측 확률이며, 이 차이는 무작위 추측에 비해 SALSA 공격자가 추가로 얻는 어드벤티지이다. 따라서 방정식 (6)은 설계자가  $F \in \theta$ 를 올바른 복호가 유지되는 범위에서 SALSA가 얻는 추가 어드벤티지  $\mathbb{E}[\text{성공률}] - 2^{-H_\infty(s)}$ 가  $negl(\lambda)$  이하가 되며 동시에 복호 오차 확률이  $negl(\lambda)$ 이 되도록 선택한다. 랜덤 추측과 같은 수준의 성공 확률을 보안 매개변수  $\lambda$ 에 대해  $negl(\lambda)$  수준까지 줄인다.

## 2. Proposed System Architecture

본 장에서는 제안하는 영관계 격자 알고리즘의 전체 구

조를 설명한다. 제안하는 알고리즘은 키 생성, 암호화 및 복호화의 세 과정으로 이뤄진다. 모든 연산은 모듈러  $q$ 에서 수행되며, 내적은  $\langle \cdot, \cdot \rangle$ , RLWE에서는 링곱  $*$ 로 정의한다. 제안 기법은 기본적으로 LWE 기반 공개키 구조를 따르지만, 비밀키를  $(s, k)$ 로 정의하고 맥클로린 급수 기반의 키 의존 난독화 함수  $f_k$ 를 정의하여  $s' = f_k(s)$ 를 생성한다. 공개키는  $(A, b')$  형태로 유지되지만,  $b'$ 은 원래  $s$ 가 아니라  $s'$ 에 결합 되도록 구성한다. 암호화 과정에서는 추가로  $k$ 에 의존하는 추가 마스킹을 삽입하여 공격자가 암호문과 비밀키 사이의 통계적 상관관계를 학습하더라도 평문이 직접 노출되지 않도록 기밀성을 강화한다.

### 2.1 Hallucination Key Generation

키 생성 단계에서 보안 매개 변수  $\lambda$ 에 대해 LWE 파라미터  $(n, q, m, \chi)$ 를 선택한다.

비밀 벡터  $s$ 는  $s \in \mathbb{Z}_q^n$  위의 오차 분포  $\chi$ 에서 독립적으로 샘플링된 좌표들로 구성되며 방정식 (7)과 같이 정의한다.

$$s \leftarrow \chi \subset \mathbb{Z}_q^n \quad (7)$$

방정식 (7)과 동시에 길이  $k$ 의 비밀 파라미터  $k \leftarrow \{0, 1\}^k$ 를 샘플링한다. 이때  $k$ 는 난독화 함수의 내부 파라미터를 결정하는 키로 사용되며, 수신자와 송신자 사이에 사전 공유된 비밀값이라고 가정한다.

비밀 파라미터  $k$ 로부터 난수 생성기  $G$ 를 이용해 맥클로린 난독화 계수와 지수 집합을 방정식 (8)과 같이 유도한다.

$$(a(k), I(k)) = G(k) \quad (8)$$

방정식 (8)에서  $a(k) = (a_i(k))_{i \in I(k)}$ 는 각 전개 차수  $i$ 에 대응하는 계수들의 집합이고,  $I(k) \subseteq \mathbb{N}$ 는 실제로 사용되는 전개 차수들의 인덱스 집합이다. 공격자는  $G$ 의 알고리즘과 함수 형태는 알고 있으나, 입력  $k$ 와  $k$ 로부터 유도되는  $a(k), I(k)$ 는 알 수 없다고 가정한다. 이때 맥클로린 난독화 함수  $f_k : \mathbb{Z}_q^n \rightarrow \mathbb{Z}_q^n$ 를 방정식 (9)과 같이 정의한다.

$$f_k(s) = \sum_{i \in I(k)} a_i(k) s^{*i} \quad (9)$$

방정식 (9)에서  $s^{*i}$ 는  $s$ 의  $i$ -fold convolution

power를 의미하며, RLWE 환경에서는 다항식 곱을 이용한 convolution, 벡터 LWE 환경에서는 좌표별 곱과 합으로 구현될 수 있다. 난독화된 hallucination key는 방정식 (10)으로 정의한다.

$$s' = f_k(s) \quad (10)$$

제안 기법의 공개키는  $s'$ 에 결합된 형태로 생성된다. 공개 행렬  $A$ 와 잡음 벡터  $e$ 는 각각  $A \leftarrow \mathbb{Z}_q^{m \times n}, e \leftarrow \chi^m$ 에서 균등 및 오차 분포에 따라 샘플링된다. 이후 방정식 (11)을 계산한다.

$$b' = Af_k(s) + e = As' + e \pmod{q} \quad (11)$$

방정식 (11)을 계산하여 난독화 공개키 성분을 얻는다.

$$PK = (A, b'), SK = (s', k) \quad (12)$$

최종적으로 공개키와 비밀키는 방정식 (12)로 정의한다.

공격자는  $A$ 와  $b'$  및 LWE 파라미터를 알고 있으나,  $s, s'k$ 와  $(a(k), I(k))$ 는 모두 비밀로 유지된다. 이와 같은 조건 하에서 인공지능 기반 공격이 관찰하는 통계 구조는  $s$ 가 아니라  $s' = f_k(s)$ 에 결합되며,  $s'$ 와 공개 정보를 알고 있더라도 비밀 파라미터  $k$  없이는  $s$ 를 효율적으로 추정하기 어렵게 설계한다.

## 2.2 Encryption

암호화 알고리즘은 LWE 기반 공개키 암호 구조를 따르면서, 평문에 대해  $k$ 에 의존하는 추가 마스킹을 적용한다. 송신자는 수신자의 공개키  $PK = (A, b')$ 와 기존에 공유된 비밀키 파라미터  $k$ 를 이용하여 평문  $m$ 을 암호문  $(u, v)$ 로 변환한다. 평문  $m$ 을 내부 비트열  $c \in \{0, 1\}^l$ 로 인코딩한다. LWE 임베딩 함수  $D_{n, q}$ 와 스케일링 계수  $\Delta = \lfloor q/2 \rfloor$ 를 이용하여, 비트열  $c$ 를  $\mathbb{Z}_q$ 상의 값으로 매핑하는  $D_{n, q}(c)$ 를 정의한다. 암호화 시에는 새로운 임시 벡터와 잡음 샘플을  $r \leftarrow \{0, 1\}^m, e_1 \leftarrow \chi^n, e_2 \leftarrow \chi$ 에서 각각 샘플링하고, 이를 이용해 기존 LWE 기반 PKE와 같이 방정식 (13)과 같이 계산한다.

$$u = A^T r + e_1 \in \mathbb{Z}_q^n \quad (13)$$

계산 후, 송신자는 비밀 파라미터  $k$ 와  $u$ 를 입력으로 하는 키 유도 함수 Key Derivation Function (KDF)를 적용하여 방정식 (14)을 계산한다.

$$\rho = KDF(k, u) \in \{0, 1\}^l \quad (14)$$

평문 인코딩  $c$ 에 대해 방정식 (15)과 같이 정의한다.

$$c^* = c \oplus \rho \quad (15)$$

이를 다시 LWE 임베딩을 통해 방정식 (16)과 같이 변환한다.

$$M = D_{n, q}(c^*) \in \mathbb{Z}_q \quad (16)$$

변환한 후, 난독화된 공개키 성분  $b' = Af_k(s) + e$ 와 임시 벡터  $r$ 를 이용하여 방정식 (17)과 같이 계산한다.

$$v = \langle b', r \rangle + e_2 + M \pmod{q} \quad (17)$$

계산 후, 최종 암호문  $CT = (u, v)$ 를 출력한다. 이때 암호문과 공개키를 관찰하는 공격자는 평문  $m$ 뿐 아니라 마스킹된 내부 표현  $c^*$ 에도 직접 접근할 수 없으며,  $k$ 를 모르면  $\rho$ 를 계산할 수 없다. 따라서 선형적으로 복원할 수 있는 관계식을 구성하기 어렵다.

## 2.3 Decryption

복호화 알고리즘은 비밀키  $SK$ 를 이용하여 LWE에서 마스킹된 메시지  $c^*$ 를 복원하고, 이후  $k$ 기반 마스크를 제거함으로써 평문을 복호화한다. 수신자는 암호문  $CT$ 를 입력으로 받아 아래와 같은 절차를 수행한다.

$s' = f_k(s)$ 를 이용하여 방정식 (18)과 같이 계산한다.

$$t = v - \langle u, s' \rangle \pmod{q} \quad (18)$$

키 생성 및 암호화 과정에서의 정의를 전개하면 방정식 (19)가 된다.

$$t = v - \langle A^T r + e_1, s' \rangle = M + \eta \pmod{q} \quad (19)$$

방정식 (19)에서 잡음 항  $\eta$ 는 방정식 (20)과 같이 표현된다.

$$\eta = e_2 + \langle e, r \rangle - \langle e_1, s' \rangle \quad (20)$$

방정식 (20)에서 잡음 분포의 표준 편차, 벡터  $r$ 과  $s'$ 의 해밍 무게 상한, 모듈러스  $q$  및 스케일링 인자  $\Delta$ 를 선택할 때, 암호화 과정에서 누적되는 합성 잡음  $\eta$ 가  $\|\eta\|_\infty < \Delta/2$ 를 만족하도록 설정한다. 이 경우  $t$ 를  $\Delta$  단위로 라운딩하여  $M = D_{n, q}(c^*)$ 를 정확히 복원할 수 있다. 수신자는 역임베딩  $D_{n, q}^{-1}$ 를 적용하여 방정식 (21)과 같이 된다.

$$\hat{c}^* = D_{n, q}^{-1}(M) \quad (21)$$

잡음이 허용 범위 내이면  $\hat{c}^* = c^*$ 가 된다. 이후, 송신자와 동일하게  $u$ 를 입력으로 KDF를 적용하여 방정식 (22)과 같이 계산한다.

$$\rho = KDF(k, u) \quad (22)$$

방정식 (22)이후, 마스크 제거 연산 방정식 (23)를 수행한다.

$$\hat{c} = \hat{c}^* \oplus \rho \quad (23)$$

마스크 제거 후  $\hat{c}$ 를 평문 영역으로 디코딩하여  $\hat{m} = \text{Decode}(\hat{c})$ 를 얻는다. 잡음  $\eta$ 가 설계된 한계 내에 있는 경우  $\hat{m} = m$ 이 성립하며, 제안 기법이 LWE 복호화 정확성과  $k$  기반 평문 보호를 동시에 만족함을 확인할 수 있다.

## IV. experiment results

### 1. Simulation setup

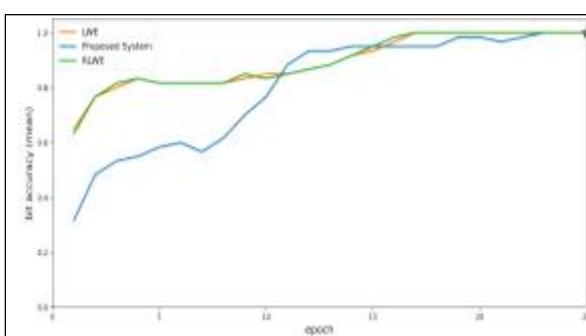
본 시뮬레이션은 NVIDIA RTX 4060 GPU, 64GB RAM CUDA 12.2 환경에서 수행되었으며, 모델 구현과 학습은 PyTorch 2.1을 기반으로 진행하였다. 데이터셋은 LWE 및 RLWE 인스턴스로 구성되었으며, 차원  $n \in \{10, 30\}$ , 모듈러스  $q = 842,779$ , 잡음 분포의 표준편차  $\sigma = 3.0$ 을 사용하였다.  $s$ 는 이진 분포에서 샘플링하였고, 해밍 무게는  $h = 3$ 으로 제한하였다. 각 실험 반복에서는 epoch 당 20,000 개의 샘플을 사용하였으며, 결과의 재현성을 보장하기 위해 동일한 난수 시드를 고정하였다. 암호화 기법의 성능 평가를 위한 취약점 공격 모델은 [14]에서 Wenger et.al.이 제안한 Transformer 기반의 SALSA를 구현하여 사용하였다. 학습 설정은 배치 크기 16, 최대 50 epoch으로 수행하였고, 각 epoch 별 학습 로그를 기록하였다. 성능 평가는 복구된 키의 단위 정확도 bitwise accuracy, 실제 비밀키 복구 여부인 true secret accuracy 및 공

격자가 진짜 비밀키를 회수하지 못한 비율 defense success rate를 사용하였다. 실험에서는 개념 증명과 공격 및 방어 알고리즘의 동작 원리 검증에 초점을 맞춰 간소화된 실험을 위해 차원 파라미터를  $n=10, 30$ 을 사용하여 수행하였다.

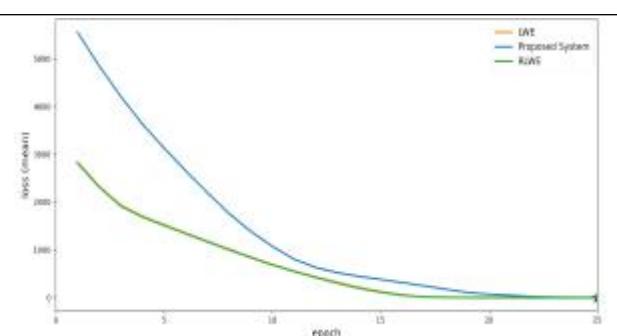
### 2 Performance evaluation

성능 평가는 제안하는 영관계 격자 알고리즘의 효과를 정량적으로 입증하기 위해 LWE, RLWE 및 영관계 격자 알고리즘에서 동일한 조건인  $n \in 10, 30, h = 3, \sigma = 3.0, q = 842,779$ 으로 학습 및 평가한 결과를 비교하였다. 평가에 사용한 지표는 복구된 키의 단위 정확도인 bitwise accuracy와 실제 비밀키 복구 여부는 true secret accuracy이며 공격자가 진짜 비밀키를 회수하지 못한 비율은 defense success rate이다. 조건별 평균 성능은 표3에 나타나 있다.

LWE와 RLWE는 bitwise accuracy와 true secret accuracy가 동시에 1.0에 수렴하여 SALSA 계열 공격이 작은 차원에서도 실제 비밀키 복구에 성공하였음을 확인할 수 있다. 반면, 제안하는 영관계 격자 알고리즘은 bitwise accuracy는 1.0으로 높게 유지되지만, true secret accuracy는  $n=10$ 일 때 0.3,  $n=30$ 일 때 0.1로 낮아, 모델이 복구한 값이 난독화된  $s'$ 임을 보여준다. 이와 같은 경향은 학습 추이를 나타낸 그림 1의 (a)에서도 동일하게 확인된다. 영관계 격자 알고리즘은 epoch 이 진행될수록 빠르게 높은 bitwise accuracy에 도달하지만 true secret accuracy는 상승하지 않아 공격이 hallucination key에 수렴함을 증명한



(a) 인공지능 기반 공격 bit-accuracy



(b) 인공지능 기반 공격 loss

그림 1 Comparison of the Learning Performance of LWE, RLWE, and the Proposed System

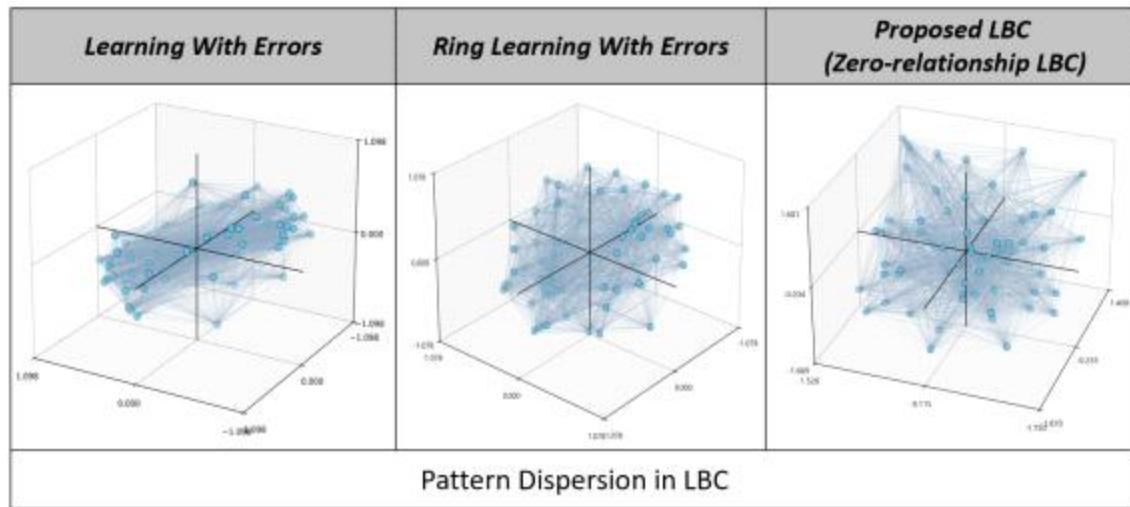


그림 2 Edge Density Matched Pattern Dispersion in LBC

다. 또한 그림 1의 (b)는 손실이 안정적으로 감소함에 따라 학습이 정상적으로 진행은 되었으나, 제안하는 기법의 경우  $s'$ 에 수렴하였기 때문에 실제로  $s$ 에 대한 공격은 이루어지지 않았음을 확인된다. LWE 및 RLWE의 경우 실제 비밀키가 노출되고 이를 이용해 평문을 성공적으로 복호화하였음을 확인했다. 그러나 제안 기법은  $s'$ 을 학습하도록 유도함으로써 비밀키 복구 공격에 대한 성공률을 효과적으로 억제하였다는 사실을 보여준다.

### 3. Security analysis

본 절에서는 제안 기법이 공격자의 관측 분포를 변화시켜 실제 비밀키 복구 시도를 효과적으로 무력화하는지 분석한다. 그림 2는 기존 격자 기반 암호들과 제안하는 기법의 데이터 간 관계성을 시각적으로 보여준다. 그림 2에서 기존 LWE 및 RLWE는 특정 방향으로 패턴이 정렬되어 있고 덩어리처럼 모여있어 데이터 간의 패턴을 쉽게 알아내어 공격에 활용할 수 있었다. 반면, 제안하는 기법은 패턴이 불규칙적으로 분산된 형태로 데이터들 간의 관계성을 약화시켜 패턴 분석 및 통계적 분석 기법

에 내성을 강화하였다. 이는 제안하는 기법이 맥클로린 기반 난독화를 이용해 통계적 상관관계가 고여버린 상태라는 것을 시각적으로 보여준다. 기존 공격자는  $(A, b)$  또는 RLWE 샘플로부터  $s$ 를 직접 추정하거나, SALSA와 같이  $b \approx \langle a, s \rangle + e$ 의 학습 가능한 상관 구조를 포착해 좌표별 후보를 추정한 뒤 잔차 분산으로 검증한다. 하지만 제안하는 기법인 영관계 격자 알고리즘에서는 키 생성 시 비밀키를 맥클로린 기반 난독화 함수  $f(\cdot)$ 로 변환하여  $s' = f(s)$ 를 사용하고, 공개키 성분도  $b' = Af(s) + e$ 로 대체한다. 이와 같은 절차를 통해 평문 내부 표현  $c$ 는 인코딩 후 비밀 마스크로 가려진다. 따라서 공격자가  $s'$ 을 완전히 복구하더라도  $k$ 가 없으면  $KDF(k, u)$ 로 생성되는 마스크 값을 계산할 수 없으므로, 마스크 제거를 수행하지 못해 평문을 얻지 못한다. 그 결과 공격자가 학습하는 분포는  $(A, b')$ 이며, 이는  $(A, s)$ 가 아니라  $(A, f(s))$ 에 의해 결정된다. 즉, 모델이 포착하는 신호는  $s$ 가 아닌  $s'$ 에 수렴하고,  $f$ 가 비선형 조합으로 설계되어  $s$ 와의 좌표별 선형 관계를 체계적으로 약화시키므로  $s' \rightarrow s$  추정은 중간 과정에서의 정보 손실로

표 3 Key Recovery Accuracy Comparison

Mode	$n$	hamming	$\sigma$	bitwise acc	true acc	obfuscated acc
LWE	10	3	3.0	1.000	1.000	-
LWE	30	3	3.0	1.000	1.000	-
RLWE	10	3	3.0	1.000	1.000	-
RLWE	30	3	3.0	1.000	1.000	-
Proposed System	10	3	3.0	1.000	0.300	0.300
Proposed System	30	3	3.0	1.000	0.100	0.100

인해 성공 확률이 매우 희박하다. 성능 평가의 그림 1과 표 3에 따르면, 영관계 격자 알고리즘의 bitwise accuracy는 높지만 true secret accuracy가 낮게 유지되어 공격자가 얻은 것은 실제 키  $s$ 가 아닌  $s'$ 임이 확인된다. 따라서 제안하는 영관계 격자 알고리즘은 인공지능을 이용한 비밀키 공격의 패턴 학습을 구조적으로 약화시키며, 실제 비밀키를 복구하는 것을 효과적으로 방어한다.

## V. Conclusion

본 논문은 기존 격자 기반 양자 내성 암호의 취약성을 보완한 영관계 격자 암호를 제안한다. 제안 기법은 격자 기반 암호에 맥클로린 난독화 함수를 적용한 이중 난독화 방식을 제시한다. 비밀키를  $(s, k)$ 로 확장하고  $s' = f_k(s)$ 를 도입함으로써, 공개키 성분이  $s$ 가 아니라  $s'$ 에 결합되도록 설계하였다. 또한 평문 인코딩 단계에서 비밀 파라미터  $k$ 에 의존하는 마스킹을 추가하여 암호문과 난독화된 비밀키가 유출되더라도 평문 기밀성을 유지할 수 있는 구조를 제시하였다.

인공지능 기반 패턴 분석 공격 모델을 이용한 실험 결과, 제안 기법은 동일한 파라미터와 학습 조건 하에서 기존 격자 기반 암호 기법들에 비해  $s$ 에 대한 복구 성공률을 10%까지 유의미하게 감소시켰다. 또한 학습된 모델이 수렴하는 대상이  $s$ 에서  $s'$ 으로 유도하는 것을 확인하였다. 이는 모델 정확도가 높을수록 실제 비밀키가 그대로 노출된다는 기존 인공지능 공격의 직관을 완화시키면서, 관측 가능한 통계 구조를 설계적으로 조정하여 학습 방향을 다른 키 공간으로 유도할 수 있음을 보여준다. 즉, 제안 기법은 인공지능 모델의 성능을 제한하는 방식이 아니라, 비밀키와 암호문 사이의 상관 구조 자체를 조절함으로써 학습 기반 공격 성능을 저하시켜 보다 강력한 안전성을 제공할 수 있음을 입증하였다.

## REFERENCES

- [1] P. W. Shor, "Polynomial-Time Algorithms for Prime Factorization and Discrete Logarithms on a Quantum Computer," SIAM Journal on Computing, Vol. 26, No. 5, pp. 1484-1509, 1997. DOI: 10.1137/S0097539795293172.
- [2] D. J. Bernstein, J. Buchmann, and E. Dahmen, Post-Quantum Cryptography, Springer, 2009. DOI: 10.1007/978-3-540-88702-7.
- [3] O. Regev, "Lattice-Based Cryptography," Advances in Cryptology - CRYPTO 2006, Lecture Notes in Computer Science, Vol. 4117, pp. 131-141, Springer, 2006. DOI: 10.1007/11818175\_8.
- [4] C. Peikert, "A Decade of Lattice Cryptography," Foundations and Trends in Theoretical Computer Science, Vol. 10, No. 4, pp. 283-424, 2016. DOI: 10.1561/0400000074.
- [5] M. R. Albrecht, R. Player, and S. Scott, "On the Concrete Hardness of Learning with Errors," Journal of Mathematical Cryptology, Vol. 9, No. 3, pp. 169-203, 2015. DOI: 10.1515/jmc-2012-0016.
- [6] O. Regev, "The Learning with Errors Problem," 2010 IEEE 25th Annual Conference on Computational Complexity (CCC), pp. 191-204, June 2010. DOI: 10.1109/CCC.2010.26.
- [7] Z. Brakerski, A. Langlois, C. Peikert, O. Regev, and D. Stehlé, "Classical Hardness of Learning with Errors," Proceedings of the 45th ACM Symposium on Theory of Computing (STOC '13), pp. 575-584, 2013. DOI: 10.1145/2488608.2488680.
- [8] O. Regev, "On Lattices, Learning with Errors, Random Linear Codes, and Cryptography," Journal of the ACM, Vol. 56, No. 6, Article 34, pp. 1-40, 2009. DOI: 10.1145/1552285.1552303.
- [9] V. Lyubashevsky, C. Peikert, and O. Regev, "On Ideal Lattices and Learning with Errors over Rings," Advances in Cryptology - EUROCRYPT 2010, Lecture Notes in Computer Science, Vol. 6110, pp. 1-23, Springer, 2010. DOI: 10.1007/978-3-642-13190-5\_1.
- [10] J. W. Bos, L. Ducas, E. Kiltz, T. Lepoint, V. Lyubashevsky, J. M. Schanck, P. Schwabe, G. Seiler, and D. Stehlé, "CRYSTALS-Kyber: A CCA-Secure Module-Lattice-Based KEM," 2018 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 353-367, 2018. DOI: 10.1109/EuroSP.2018.00032.
- [11] D. J. Bernstein, S. Dahlum, T. Lange, C. van Vredendaal, and C. van Veen, "KyberSlash: Exploiting Secret-Dependent Division in Kyber," IACR Transactions on Cryptographic Hardware and Embedded Systems, Vol. 2025, No. 2, pp. 209-234, 2025. DOI: 10.46586/tches.v2025.i2.209-234.
- [12] H. Chen, K. Lauter, and K. E. Stange, "Attacks on the Search RLWE Problem with Small Errors," SIAM Journal on Applied Algebra and Geometry, Vol. 1, No. 1, pp. 665-682, 2017. DOI: 10.1137/16M1096566.

- [13] D. Dachman-Soled, H. Gong, M. Kulkarni, and A. Shahverdi, "(In)Security of Ring-LWE Under Partial Key Exposure," *Journal of Mathematical Cryptology*, Vol. 15, No. 1, pp. 72-86, 2021. DOI: 10.1515/jmc-2020-0075.
- [14] E. Wenger, M. Chen, F. Charton, and K. Lauter, "SALSA: Attacking Lattice Cryptography with Transformers," *Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. (Proc. of the 36th Conference on Neural Information Processing Systems).
- [15] S. Stevens, E. Wenger, C. Y. Li, N. Nolte, E. Saxena, F. Charton, and K. E. Lauter, "SALSA FRESCA: Angular Embeddings and Pre-Training for ML Attacks on Learning With Errors," *Cryptology ePrint Archive*, Report 2024/150, 2024. Available: <https://eprint.iacr.org/2024/150>.
- [16] N. Nolte, M. Malhou, E. Wenger, S. Stevens, C. Li, F. Charton, and K. Lauter, "The Cool and the Cruel: Separating Hard Parts of LWE Secrets," in *Progress in Cryptology - AFRICACRYPT 2024*, Lecture Notes in Computer Science, Vol. 14861, pp. 428-453, Springer, 2024. DOI: 10.1007/978-3-031-64381-1\_19.
- [17] H. Chen, L. Chua, K. E. Lauter, and Y. Song, "On the Concrete Security of LWE With Small Secret," *La Matematica*, Vol. 3, pp. 1032-1068, 2024. DOI: 10.1007/s44007-024-00111-3.
- [18] M. R. Albrecht, "On Dual Lattice Attacks Against Small-Secret LWE and Parameter Choices in HElib and SEAL," *Advances in Cryptology - EUROCRYPT 2017*, Lecture Notes in Computer Science, Vol. 10210, pp. 103-129, Springer, 2017. DOI: 10.1007/978-3-319-56614-6\_4.
- [19] J. Cao, H. Jiang, and Q. Cheng, "Refined Attack on LWE with Hints: Constructing Lattice via Gaussian Elimination," *Advances in Cryptology - CRYPTO 2025*, Lecture Notes in Computer Science, pp. 385-416, Springer, 2025. DOI: 10.1007/978-3-032-01855-7\_13.
- [20] M. R. Albrecht, C. Cid, J.-C. Faugère, R. Fitzpatrick, and L. Perret, "On the Complexity of the BKW Algorithm on LWE," *Designs, Codes and Cryptography*, Vol. 74, No. 2, pp. 325-354, 2015. DOI: 10.1007/s10623-013-9864-x.
- [21] A. May and J. Nowakowski, "Too Many Hints - When LLL Breaks LWE," *Advances in Cryptology - ASIACRYPT 2023*, Lecture Notes in Computer Science, Vol. 14441, pp. 106-137, Springer, 2023. DOI: 10.1007/978-981-99-8730-6\_4.