University of California
Santa Barbara

# Statistical Methods in Cryptography

A thesis submitted in partial satisfaction
of the requirements for the degree

Master of Science
in
Computer Science

by

Wei Dai

Committee in charge:

Professor Stefano Tessaro, Chair
Professor Huijia (Rachel) Lin
Professor Ömer Eğecioğlu

September 2016

The Thesis of Wei Dai is approved.

Professor Huijia (Rachel) Lin

Professor Ömer Eğecioğlu

Professor Stefano Tessaro, Committee Chair

June 2016

Statistical Methods in Cryptography

# Acknowledgements

This thesis would not materialize without the guidance of my advisor, Prof. Stefano Tessaro. Looking back, every topic explored in this thesis was originally suggested to me by him. His guidance not only helped me navigate the cryptographic literature, but also helped shape my reserach interests and approach.

I would also like to thank my thesis committee for the helpful comments on earlier drafts of this thesis.

# Abstract

Statistical Methods in Cryptography

by

Wei Dai

Cryptographic assumptions and security goals are fundamentally distributional. As a result, statistical techniques are ubiquitous in cryptographic constructions and proofs. In this thesis, we build upon existing techniques and seek to improve both theoretical and practical constructions in three fundamental primitives in cryptography: blockciphers, hash functions, and encryption schemes. First, we present a tighter hybrid argument via collision probability that is more general than previously known, allowing applications to blockciphers. We then use our result to improve the bound of the Swap-or-Not cipher. We also develop a new blockcipher composition theorem that is both class and security amplifying. Second, we prove a variant of Leftover Hash Lemma for joint leakage, inspired by the Universal Computational Extractor (UCE) assumption. We then apply this technique to construct various standard-model UCE-secure hash functions. Third, we survey existing "lossy primitives" in cryptography, in particular Lossy Trapdoor Functions (LTDF) and Lossy Encryptions (LE); we propose a generalized primitive called Lossy Deterministic Encryption (LDE). We show that LDE is equivalent to LTDFs. This is in contrast with the block-box separation of trapdoor functions and public-key encryption schemes in the computational case. One common theme in our methods is the focus on statistical techniques. Another theme is that the results obtained are in contrast with their computational counterparts—the corresponding computational results are implausible or are know to be false.

# Contents

**Abstract** **v**

**1 Introduction** **1**
1.1 Hybrid Argument . . . . . . . . . . . . . . . . . . . . . . . . . . . 1
1.2 Blockcipher Composition Theorems . . . . . . . . . . . . . . . . . . 2
1.3 Leftover Hash Lemma for Joint Leakage . . . . . . . . . . . . . . . . 4
1.4 Unified Lossy Primitive: Lossy Deterministic Encryption . . . . . . . . . 5

**2 Preliminaries** **7**
2.1 Statistical Definitions . . . . . . . . . . . . . . . . . . . . . . . . . 8
2.2 A Generalized Leftover Hash Lemma via Rényi Divergence . . . . . . . 13
2.3 Cryptographic Background . . . . . . . . . . . . . . . . . . . . . . . 19

**3 Tighter Security Proofs Using Collision Probability and Chi-Square Distance** **24**
3.1 Tighter Hybrid Argument . . . . . . . . . . . . . . . . . . . . . . . . 24
3.2 Tighter Composition Theorem via Chi-Square Distance . . . . . . . . . . 29

**4 Leftover Hash Lemma for Joint Leakage** **34**
4.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 34
4.2 Leftover Hash Lemma for Joint Leakage . . . . . . . . . . . . . . . . . 36
4.3 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\mathsf{sup}}]$ with Constant Output Length . . . . . . 38
4.4 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\mathsf{sup}} \cap \mathcal{S}^{\mathsf{splt}}]$ with Polynomial Output Length . 39
4.5 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\mathsf{sup}}]$ with Polynomial Output Length . . . . . 42

**5 Unified Lossy Primitive: Lossy Deterministic Encryption** **46**
5.1 Preliminaries . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 46
5.2 Survey of Existing Results . . . . . . . . . . . . . . . . . . . . . . . 46
5.3 Unifying Definition: Lossy Deterministic Encryption . . . . . . . . . . . 50
5.4 Equivalence of Lossy Deterministic Encryption and Lossy Trapdoor Functions . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 52

**Bibliography** **54**

# Chapter 1

# Introduction

Modern cryptography is almost exclusively based on computational assumptions. The notions such as computational indistinguishablity, one-way functions, pseudorandom generators and pseudorandom functions are of fundamental importance. Coupled with the computational assumptions, cryptographic proofs and constructions often utilize numerous other techniques. Since the cryptographic assumptions and security goals are fundamentally distributional, statistical techniques is ubiquitous. In this thesis, we build upon existing techniques and constructions in three different topics: blockciphers, hash functions, and encryption schemes. One common theme in our methods is the focus on statistical techniques. Another theme is that the results obtained are in contrast with their computational counterparts—the corresponding computational results are implausible or are know to be false.

## 1.1 Hybrid Argument

The hybrid argument, which is essentially the triangle inequality, is arguably the most fundamental tool in security proofs. When bounding the difference in behavior of randomized algorithms with $\{0,1\}$ output, $G_0$, from $G_1$, one can define an intermediate system $G_h$. The triangle inequality tells us that

$$|\Pr[G_0 \Rightarrow 1] - \Pr[G_1 \Rightarrow 1]| \leq |\Pr[G_0 \Rightarrow 1] - \Pr[G_h \Rightarrow 1]| + |\Pr[G_h \Rightarrow 1] - \Pr[G_1 \Rightarrow 1]|.$$

In the statistical case, the statistical distance SD, which we will define formally later, also satisfy the triangle inequality. For $X, Y, Z$ distributions over the same sample space, we have

$$\mathsf{SD}(X, Z) \leq \mathsf{SD}(X, Y) + \mathsf{SD}(Y, Z).$$

Applying this multiple times, one can bound $\mathsf{SD}(X_1, X_n)$ by bounding $\mathsf{SD}(X_i, X_{i+1})$ for all $i = 1, \ldots, n-1$. In particular, if $\mathsf{SD}(X_i, X_{i+1}) \leq \epsilon_i$ for all $i = 1, \ldots, n-1$, then $\mathsf{SD}(X_1, X_n) \leq \sum_i \epsilon_i$. However, in many applications, the individual bounds of $\epsilon_i$ are obtained from $\mathrm{CP}(X_i \mid X_{<i}) \leq \frac{1+\epsilon^2}{M}$, where $M$ is the support size of $X_i$. We show that, with this assumption, a tighter bound can be obtained.

$$\mathsf{SD}(X_1, X_n) \leq \sqrt{\sum_i \epsilon_i^2}.$$

As application of this result, we prove a tighter bound of the Swap-or-Not construction.

## 1.2 Blockcipher Composition Theorems

Blockciphers are fundamental primitives in practical cryptography. It is the building block of symmetrical cryptography. The security game of a blockcipher is, essentially, an interaction of an adversary with the blockcipher. The adversary has access to the blockcipher as a oracle to query values of the cipher at different input points. At the end, the adversary should output a bit. The advantage of the adversary against the blockcipher is the difference in the adversary's behavior between interacting with the blockcipher and interacting with a truly randomly sampled permutation.

**Adaptive versus Non-Adaptive Security** For the bounds to have practical cryptographic importance, we usually want to prove adaptive security. That is to say, the adversary can query the blockcipher depending on the answers it has received. However, non-adaptive security is usually easier to prove. The reason is that, often times, proving non-adaptive security of a blockcipher is similar to bounding the convergence time of a Markov chain. Currently, we do not have techniques to deal with adaptivity directly. Instead, the final security claim is reach by way of a composition theorem. The "two weak make one strong" theorem, proved in [MPR07], states that the composition two non-adaptively secure blockcipher is adaptively secure with roughly the same parameters (the formal statement is given in Chapter 3). However, for practical bounds, this means that one has to double the number of rounds to achieve adaptive security.

In Chapter 3, we define the notion of chi-square distance. We show that, using this new notion of distance, we get both distinguisher-class amplification and security amplification. In particular, under the new definition of security, the composition of two non-adaptively $\epsilon$-secure blockciphers is a adaptively $\epsilon^2$-secure blockcipher.

However, we note the potential difficulties in bounding the chi-square distance to the Swap-or-Not construction, and we leave it as an open problem.

## 1.3 Leftover Hash Lemma for Joint Leakage

The Leftover Hash Lemma [HILL99] has seen wide-usage in cryptography, such as for key derivation and deterministic encryption. Roughly, it states that if a random variable $X$ has enough "entropy", then

$$(S, h_S(X)) \approx (S, U),$$

where $h$ is a universal hash function, $S$ is a uniform random seed, and $U$ is a uniform element in the range of $h$. The average case variant of Leftover Hash Lemma [DORS08], states essentially the same result with the present of leakage. In particular, if $X$ has enough entropy given leakage $L(X)$, then

$$(S, L(X), h_S(X)) \approx (S, L(X), U).$$

**UCEs** The Universal Composition Extractor (UCE) assumption is an assumption of hash functions [BHK13]. It was designed to instantiate hash functions in random oracle model in the standard model [BHK13]. UCE is a parameterized assumption based on the type of UCE source involved. A UCE source, $\mathcal{S}^{\mathsf{Hash}}$, makes queries to the oracle, $\mathsf{Hash}$, and produces leakage $L$. In this work, we focus on statistically unpredictable sources, that is the queries of $L$ should have high ($\omega(\log(\lambda))$) min-entropy given the leakage $L$. Another classification of sources is a split source, $\mathcal{S} = \mathsf{splt}[\mathcal{S}_0, \mathcal{S}_1]$. Roughly speaking, the leakage of

**Joint Leakage and Construction of UCEs** Following the Universal Computational Extractor assumption [BHK13], we look at the case where the leakage function $L$ takes both the source $X$ and the output $h_S(X)$ or $U$. More concretely, given that $X$ has enough entropy given $L(X, U)$, we ask if

$$(S, L(X, h_S(X))) \overset{?}{\approx} (S, L(X, U)).$$

We prove a variant of Leftover Hash Lemma for the case of joint leakage. We obtain UCE-secure hash function for slightly increasing number of query and constant output length. With help of the recent Extremely Lossy Function (ELF) assumption [Zha16], we show how to get $O(\lambda)$ output length (and hence poly output length) with the same query regime.

## 1.4 Unified Lossy Primitive: Lossy Deterministic Encryption

Lossy trapdoor functions (LTDF) was first proposed by Peikert and Waters [PW11]. In their work, concrete instantiations of LTDF was achieved under decisional Diffie-Hellman (DDH) and learning-with-error (LWE) assumptions. Roughly speaking, a LTDF is a family of efficiently computable functions that can be sampled in two modes: *injective* mode and *lossy* mode. In injective mode, the function is a regular trapdoor function. That is, one can invert the function using a trapdoor. In the lossy mode, the function is compressing, meaning that the image size is smaller than the domain size. The security of LTDFs states that the two modes should be computationally indistinguishable. Peikert and Waters used the abstraction of LTDFs to build the first CCA encryption scheme based on lattice assumptions. Following their seminal work, LTDFs have helped realize multiple security goals in the standard model, including lossy encryption, deterministic encryption, and hedged encryption.

Lossy encryption was first proposed by [PVW08] to construct oblivious transfer

(OT) protocols. In simple terms, lossy encryption has two modes, injective and lossy mode. In injective mode, it should act as a correct encryption scheme. In the lossy mode, encryption of any two messages should be statistically close (under the encryption randomness). As with LTDFs, the security of lossy encryption says that the two modes should be computationally indistinguishable. The first construction of lossy encryption was realized using DDH and LWE. Follow-up works [BBN$^+$09, BHY09] utilize LTDFs as an abstraction to build lossy encryption.

Deterministic encryption and hedged encryption is motivated by randomness subversion in public-key encryption. Deterministic encryption offers security when the encrypted messages has enough entropy (requiring the messages to be either a block-source [BFOR08] or a $q$-block-source [FOR15]). Hedged encryption [BBN$^+$09] offers standard IND-CPA when the encryption randomness is ideal. When the encryption randomness is compromised, it offers security when the message and randomness *jointly* offers contains entropy.

In Chapter 5, we put forth a unified security definition of lossy Deterministic Encryption (LDE) and construct secure LDE using LTDF in a modular way. In short, LDE requires statistical closeness for high-entropy messages in the lossy mode. We show that such assumption is, in some sense, equivalent to LTDFs.

**Theorem** (Informal)**.** *Any Lossy Deterministic Encryption Scheme is also a Lossy Trapdoor Function.*

# Chapter 2

# Preliminaries

We use log to denote the base-2 logarithm. Given a positive integer $n$, we define $[n] = \{1, ...., n\}$, and use uppercase letters to denote the exponential base 2 of the lowercase letters (for example, $N = 2^n$). Let $\Omega$ be a finite sample space, we identify a distribution, $X$, over $\Omega$ with its probability mass function $P_X : \Omega \to [0,1]$[1]. For a distribution or random variable $X$, we use calligraphic letter, $\mathcal{X}$, to denote the support. For notational convenience, we also write $|X|$ as the size of the support of $X$. We use $\|P_X\|_\alpha := (\sum_{\omega \in \Omega} P_X(\omega)^\alpha)^{1/\alpha}$, for $\alpha \geq 1$, to denote the $\alpha$-norm of the pmf, $P_X$. We use $(X, Y)$ to denote the joint probability distribution of $X$ and $Y$. For a set $S$, we let $U_S$ denote the uniform distribution on $S$, and we use $U_n$ to denote $U_{\{0,1\}^n}$, the uniform $n$-bit strings.

**Cryptographic convention** We often specify a distribution by how it is sampled. For instance,

$$\{(x, z) : x \overset{\$}{\leftarrow} \{0,1\}^n, z = x\},$$

[1] Notice that since we are working with a finite sample space, a distribution is uniquely determined given a probability mass function, and vice versa.

denotes a pair of $n$-bit strings, that is equal and both marginally uniform.

**Security parameter** We use $\lambda$ as the security parameter. All randomized algorithm is assumed to take $1^\lambda$ as an implicit input. All parameters will be functions of $\lambda$. If "polynomial" or "negligibility" of a parameter is mentioned, it will be with respect to the security parameter $\lambda$. We use poly to denote the class of polynomials in $\lambda$, and const to denote some constant function of $\lambda$. For a function of $\lambda$, $f(\lambda)$, we will sometimes suppress writing the security parameter $\lambda$ and just write $f$.

## 2.1 Statistical Definitions

**Definition 1.** *The statistical distance of two distributions over $\Omega$, $X, Y$, is defined to be,*

$$\mathsf{SD}(X,Y) := \frac{1}{2}\sum_{\omega\in\Omega} |P_X(\omega) - P_Y(\omega)|.$$

*If $\mathsf{SD}(X,Y) \leq \epsilon$, we also write*

$$X \approx_{s,\epsilon} Y.$$

### 2.1.1 Entropy

There are three notions of entropy that is of interest: Shannon entropy, minimum entropy and collision entropy. Here, we adopt the unified definition of Rényi Entropy [R$^+$61] and a conditional variant [FB14].

**Definition 2** (Rényi Entropy [R$^+$61])**.** *Let $X$ be a distribution, for $\alpha > 1$, the Rényi entropy of $X$ is defined to be*

$$H_\alpha(X) := \frac{\alpha}{1-\alpha} \log \|P_X\|_\alpha.$$

*We also define $H_\infty(X)$ to be the limit of $H_\alpha(X)$ as $\alpha \to \infty$, and $H_1(X)$ to be the limit of $H_\alpha(X)$ as $\alpha \to 1$.*

**Remark.** *Let $X$ be a distribution over $\Omega$. We have that*

$$H_\infty(X) = -\log(\max_{\omega \in \Omega} P_X(\omega)),$$

$$H_1(X) = -\sum_{\omega \in \Omega} P_X(\omega) \log P_X(\omega).$$

**Definition 3** (Conditional Rényi Entropy [FB14]). *Let $\alpha > 1$, we define*

$$H_\alpha(X \mid Y) := -\log \mathsf{Ren}_\alpha(X \mid Y),$$

*where*

$$\mathsf{Ren}_\alpha(X \mid Y) := \mathrm{E}_Y[\|P_{X|Y=y}\|_\alpha]^{\frac{\alpha}{\alpha-1}}.$$

**Lemma 1** ([FB14]). *Let $X, Y$ be jointly distributed and $1 < \alpha < \beta$. Then,*

$$\log(|X|) \geq H_\alpha(X \mid Y) \geq H_\beta(X \mid Y) \geq 0.$$

Additionally, we define two natural notions, *collision probability* and *prediction probability*.

**Definition 4.**

$$\mathrm{CP}(X) := \mathrm{E}_X[P_X(X)],$$

$$\mathsf{Pred}(X) := \max_{x \in \mathcal{X}} P_X(x).$$

*The average case are simply the expectation over the conditioned variable.*

$$\mathrm{CP}(X \mid Y) := \mathrm{E}_Y[\mathrm{CP}(X \mid Y = y)],$$

$$\mathsf{Pred}(X \mid Y) := \mathrm{E}_Y[\mathsf{Pred}(X \mid Y = y)].$$

**Lemma 2** (Jensen's Inequality). *Let $X$ be a discrete real-valued random variable. Let $f$ be a convex function. Then,*

$$f(\mathrm{E}[X]) \leq \mathrm{E}[f(X)].$$

A direct corollary of Jensen's inequality relates Rényi probability of second order and the conditional collision probability.

**Corollary 1.**

$$\mathsf{Ren}_2(X \mid Y) \leq \mathrm{CP}(X \mid Y).$$

*Proof.* We first rewrite both sides.

$$\mathsf{Ren}_2(X \mid Y) = \mathrm{E}_Y[\|P_{X|Y=y}\|_2]^2,$$

$$\mathrm{CP}(X \mid Y) = \mathrm{E}_Y[\|P_{X|Y=y}\|_2^2].$$

The result now follows from the convexity of squaring. □

Another useful notion is Rényi divergence.

**Definition 5.** *Let $1 < \alpha$. The Rényi divergence of order $\alpha$ is defined to be,*

$$D_\alpha(X\|Y) := \frac{1}{\alpha - 1} \log(\sum_\omega P_X(\omega)^\alpha P_Y(\omega)^{1-\alpha}).$$

*Additionally,*

$$D_1(X\|Y) = \lim_{\alpha \to 1} D_\alpha(X\|Y),$$

$$D_\infty(X\|Y) = \lim_{\alpha \to \infty} D_\alpha(X\|Y).$$

*$D_1$ is also known as the Kullbach-Leibler (KL) divergence. For notational convenience, we use $RD$ as the exponentiated Rényi divergence.*

$$RD_\alpha(X\|Y) = 2^{D_\alpha(X\|Y)}.$$

**Proposition 1.**

$$D_1(X\|Y) = \mathrm{E}_X[\log(\frac{P_X}{P_Y})].$$

$$D_\infty(X\|Y) = \log(\sup_\omega \frac{P_X}{P_Y}).$$

**Proposition 2.** *Let $X$ be a distribution on $\mathcal{X}$. Then,*

$$H_\alpha(X) + D_\alpha(X\|U_{\mathcal{X}}) = \log|\mathcal{X}|.$$

**Lemma 3** (Pinsker's Inequality)**.**

$$\mathsf{SD}(X,Y) \leq \sqrt{\frac{\ln(2)}{2} D_1(X\|Y)}.$$

**Lemma 4** (Second Order Pinsker's Inequality [GS02])**.**

$$\mathsf{SD}(X,Y) \leq \frac{1}{2}\sqrt{RD_2(X\|Y) - 1}.$$

**Lemma 5** (Leakage Chain Rule for Rényi Entropy [FB14])**.** *The leakage chain rule states that, conditioning on a random variable $Z$, the Rényi entropy only decreases by at most* $\log(|Z|)$*, i.e.*

$$H_\alpha(X \mid Y, Z) \geq H_\alpha(X \mid Y) - \log(|Z|).$$

**Proposition 3.**

$$RD_2((X,Z)\|(Y,Z)) = \mathrm{E}_Z[RD_2(X \mid Z = z\|Y \mid Z = z)]$$

*Proof.* We derive that

$$\begin{aligned}
RD_2((X,Z)\|(Y,Z)) &= \sum_{(x,z)\in\Omega} \frac{P_{(X,Z)}(x,z)^2}{P_{(Y,Z)}(x,z)} \\
&= \sum_{(x,z)\in\Omega} \frac{P_Z(z)^2 P_{X|Z=z}(x)^2}{P_Z(z) P_{Y|Z=z}(x)} \\
&= \sum_z P_Z(z) \sum_x \frac{P_{X|Z=z}(x)^2}{P_{Y|Z=z}(x)} \\
&= \mathrm{E}_Z[RD_2(X \mid Z = z\|Y \mid Z = z)].
\end{aligned}$$

□

### 2.1.2 Universal Hash Functions and $q$-Wise Independent Hash Functions

**Definition 6.** *A collection of functions*

$$\mathbf{H} = \{H_s : \{0,1\}^n \to \{0,1\}^m \mid s \in \{0,1\}^d\}$$

*is universal, if for $H \xleftarrow{\$} \mathbf{H}$, and two distinct input $x \neq y$,*

$$\Pr[H(x) = H(y)] = \frac{1}{M} = \frac{1}{2^m}.$$

**Definition 7.** *A collection of functions*

$$\mathbf{H} = \{H_s : \{0,1\}^n \to \{0,1\}^m \mid s \in \{0,1\}^d\}$$

*is q-wise independent, if for* $H \overset{\$}{\leftarrow} \mathbf{H}$, *and q pairwise distinct inputs* $x_1, \ldots, x_q$, *the outputs,* $H(x_1), \ldots, H(x_q)$, *are independent as random variables and marginally uniform.*

## 2.2 A Generalized Leftover Hash Lemma via Rényi Divergence

We present a generalized version of Leftover Hash Lemma [HILL99] that generalizes the following variants studies in literature: crooked [DS05, FOR15], average case [DORS08], and multiple-input Leftover Hash Lemma [FOR15].

**Flat sources** A distribution $X$ such that $H_\infty(X) \geq k$ is called a $k$-source. A flat $k$-source is a distribution whose probability mass function is constant on any non-zero inputs. A well-known fact is that any $k$-source is a convex combination of flat $k$-sources [V$^+$].

We prove an analogous fact for (conditionally) simple sources, which we will define below. The motivation is to be able to study *(conditional) q-wise k-sources*, which are jointly distributions of the form $\mathbf{X} = (X_1, \ldots, X_q)$ $((\mathbf{X}, Z) = ((X_1, \ldots, X_q), Z))$ such that the component wise (conditional) entropy is at least $k$, i.e. $H_\infty(X_i) \geq k$ $(H_\infty(X_i \mid Z) \geq k)$. Such sources directly corresponds to statistically unpredictable UCE sources which we will study in Chapter 4.

**Definition 8** ((Conditionally) Simple sources)**.** *A q-wise k-source,* $\mathbf{X} = (X_1, \ldots, X_q)$, *is simple if there exists injective functions,* $f_2, \ldots, f_q$, *such that* $\mathbf{X} = (X_1, f_2(X_1) \ldots, f_q(X_1))$.

*We say that a conditional $q$-wise $k$-source, $(\mathbf{X}, Z)$, is conditionally simple if for all $z$, $\mathbf{X} \mid Z = z$ is simple.*

**Lemma 6.** *A (conditional) $q$-wise $k$-source is a convex combination of (conditionally) simple $q$-wise $k$-sources.*

*Proof.* We first prove the unconditional version. Let $\mathbf{X} = (X_1, \ldots, X_q)$ be a $q$-wise $k$-source. Without loss of generality, we can assume that $|\mathrm{supp}(X_i)| = 2^k$. The proof is by induction on $q$. Base case, let $q = 2$; we show that $(X_1, X_2)$ is a convex combination of simple $k$-sources. We identify the support of $X_1$ and $X_2$ with integers in $[2^k]$. Consider the matrix $A = (a_{i,j} = \Pr\left[(X_1, X_2) = (i, j)\right])$. Notice the constraints of being a flat 2-wise $k$-source says that $\sum_i a_{i,j} = 2^{-k}$ for all $i$ and $\sum_j a_{i,j} = 2^{-k}$ for all $j$. Furthermore, $\sum_{i,j} a_{i,j} = 1$. We note that $2^k A$ is doubly stochastic. Hence, $2^k A$ is a convex combination of permutation matrices of size $2^k$. Notice that permutation matrix of size $2^k$ by $2^k$, say $(p_{ij})$, corresponds to an injective function $f : \mathrm{supp}(X_1) \to \mathrm{supp}(X_2)$, where $f(i) = j$ iff $p_{ij} = 1$. Hence, $(X_1, X_2)$ is a convex combination of simple sources of the form $(X_1, f(X_1))$. We proceed by means of induction. Observe that if $(X_1, \ldots, X_i)$ is simple, an injective function $f : \mathrm{supp}(X_1, \ldots, X_i) \to \mathrm{supp}(X_{i+1})$ corresponds to some injective $f' : \mathrm{supp}(X_1) \to \mathrm{supp}(X_{i+1})$. As the inductive step, we suppose that $\mathbf{X} = (X_1, \ldots, X_{i+1})$ is a $q$-wise flat $k$-source. We know that $(X_1, \ldots, X_i)$ is a convex combination of simple sources. For each of this simple source, say $(X_1', f_2(X_1)', \ldots, f_i(X_i'))$, we apply the argument in base case to the distribution $(X_1', X_{i+1}')$, specified by,

$$\Pr\left[(X_1', X_{i+1}') = (x_1, x_{i+1})\right] = \\ \Pr\left[X_1' = x_1\right] \cdot \Pr\left[\mathrm{X}_{i+1} = x_{i+1} \mid \mathrm{X}_1 = x_1, \mathrm{X}_2 = f_2(x_1), \ldots, \mathrm{X}_i = f_i(x_1)\right].$$

Note that if

$$(X_1, \dots, X_i) = \sum_{f_2,\dots,f_i} c_{f_2,\dots,f_i}(X_1', f_2(X_1'), \dots, f_i(X_1')),$$

then

$$(X_1, \dots, X_i, X_{i+1}) = \sum_{f_2,\dots,f_i} c_{f_2,\dots,f_i}(X_1', f_2(X_1'), \dots, f_i(X_1'), X_{i+1}').$$

Hence, since $(X_i', X_{i+1}')$ is a convex combination of simple ones, so is $(X_1, \dots, X_{i+1})$. This concludes the proof for the unconditional case. The conditional case follows from the fact that the joint distribution is a convex combination of marginal ones,

$$(\mathbf{X}, Z) = \sum_z \Pr\left[Z = z\right](\mathbf{X} \mid Z = z, z).$$

Each $\mathbf{X} \mid Z = z$ is a convex combination of simple sources. We classify the simple sources by the functions $f_2, \dots, f_q$, and write

$$(\mathbf{X} \mid Z = z, z) = \sum_{f_2,\dots,f_q} c_{f_2,\dots,f_q,z}(X_1 \mid Z = z, f_2(X_1 \mid Z = z), \dots, f_q(X_1 \mid Z = z), z).$$

Hence, we can rewrite $(\mathbf{X}, Z)$ as a convex combination of conditionally simple sources.

$$\begin{aligned}(\mathbf{X}, Z) &= \sum_z \Pr\left[Z = z\right] \sum_{f_2,\dots,f_q} c_{f_2,\dots,f_q,z}(X_1 \mid Z = z, f_2(X_1 \mid Z = z), \dots, f_q(X_1 \mid Z = z), z)\\ &= \sum_{f_2,\dots,f_q} c'_{f_2,\dots,f_q} \sum_z c''_{f_2,\dots,f_q,z}((X_1 \mid Z = z, f_2(X_1 \mid Z = z), \dots, f_q(X_1 \mid Z = z)), z),\end{aligned}$$

where $c'_{f_2,\dots,f_q} = \sum_z c_{f_2,\dots,f_q,z} \Pr\left[Z = z\right]$ and $c''_{f_2,\dots,f_q,z} = \frac{c_{f_2,\dots,f_q,z} \cdot \Pr[Z=z]}{c'_{f_2,\dots,f_q}}$. □

**Theorem 1** (Generalized Leftover Hash Lemma)**.** *Let $h$ be a $2q$-wise independent hash function with input space $D$ and output space $R$. Let $f : R \to S$ be any surjective function. Let $\mathbf{X} = (X_1, \dots, X_q)$ be a $q$-wise $k$-source with leakage $Z$ where each $X_i$ has support $D$, and*

*$X_i \neq X_j$ for $i \neq j$. Let $\mathbf{U} = (U_R, \dots, U_R)$. It holds for uniformly sampled seed $S$ that*

$$(S, f(h(S, \mathbf{X})), Z) \approx_{s,\epsilon} (S, f(\mathbf{U}), Z), \tag{2.1}$$

*where $\epsilon = \frac{1}{2}\sqrt{2^{-k}(|S|^q - 1)}$.*

*Proof.* By Lemma 6, it suffices to prove Equation 2.1 for conditionally simple $q$-wise $k$-sources, $(\mathbf{X}, Z)$. However, it will be convenient that we first prove this for the unconditional simple sources. Let $\mathbf{X} = (X_1, g_2(X_1), \dots, g_q(X_1))$ be a simple $q$-wise $k$-source. We will show that

$$A = (S, f(h(S, \mathbf{X}))) \approx_{s,\epsilon} (S, f(\mathbf{U})) = B,$$

where $\epsilon = \frac{1}{2}\sqrt{|S|^q 2^{-k}}$. Here, we follow the technique by [Zha16] and compute the second order Rényi divergence between $A$ and $B$.

$$\begin{aligned} RD_2(A\|B) &= \sum_{s,\mathbf{y}} \frac{(\Pr[S=s]\Pr[f(h(s,\mathbf{X})) = \mathbf{y}])^2}{\Pr[S=s]\Pr[f(\mathbf{U}) = \mathbf{y}]} \\ &= \sum_{\mathbf{y}} \frac{\sum_s \Pr[S=s]\Pr[f(h(s,\mathbf{X}) = \mathbf{y}]^2}{\Pr[f(\mathbf{U}) = \mathbf{y}]}. \end{aligned}$$

Notice that, for $\mathbf{X}_1, \mathbf{X}_2 \overset{\$}{\leftarrow} \mathbf{X}$.

$$\sum_s \Pr[S=s]\Pr[f(h(s,\mathbf{X})) = \mathbf{y}]^2 = \Pr[f(h(S,\mathbf{X}_1)) = y = f(h(S,\mathbf{X}_2))].$$

Also, by the universality and $2q$-wise independence of $h$,

$$\Pr[f(\mathbf{U}) = y] = \Pr[h(S,\mathbf{X}_1) = \mathbf{y}].$$

Hence, we rewrite

$$\begin{aligned}
RD_2(A\|B) &= \sum_{\mathbf{y}} \frac{\Pr\left[f(h(S,\mathbf{X}_1)) = y = f(h(S,\mathbf{X}_2))\right]}{\Pr\left[h(S,\mathbf{X}_2) = \mathbf{y}\right]} \\
&= \sum_{\mathbf{y}} \Pr\left[f(h(S,\mathbf{X}_1)) = \mathbf{y} \mid \Pr\left[f(h(S,\mathbf{X}_2)) = \mathbf{y}\right]\right] \\
&= \sum_{\mathbf{x}_1,\mathbf{x}_2} \Pr\left[\mathbf{X} = \mathbf{x}_1\right] \Pr\left[\mathbf{X} = \mathbf{x}_2\right] \sum_{\mathbf{y}} \Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y} \mid f(h(S,\mathbf{x}_2)) = \mathbf{y}\right].
\end{aligned}$$

We look at two events separately. If $\mathbf{x}_1 = \mathbf{x}_2$, then

$$\Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y} \mid f(h(S,\mathbf{x}_2)) = \mathbf{y}\right] = 1.$$

Hence,

$$\sum_{y} \Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y} \mid f(h(S,\mathbf{x}_2)) = \mathbf{y}\right] = |S|^q.$$

If $\mathbf{x}_1 \neq \mathbf{x}_2$, since $\mathbf{X}$ is simple, $\mathbf{x}_1$ and $\mathbf{x}_2$ differ at every component. Hence, by $2q$-wise independence of $h$,

$$\Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y} \mid f(h(S,\mathbf{x}_2)) = \mathbf{y}\right] = \Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y}\right],$$

and

$$\sum_{y} \Pr\left[f(h(S,\mathbf{x}_1)) = \mathbf{y} \mid f(h(S,\mathbf{x}_2)) = \mathbf{y}\right] = 1.$$

Finally, we have

$$\begin{aligned}
RD_2(A\|B) &= \sum_{\mathbf{x}} \Pr\left[\mathbf{X} = \mathbf{x}\right] |S|^q + \sum_{\mathbf{x}_1 \neq \mathbf{x}_2} \Pr\left[\mathbf{X} = \mathbf{x}_1\right] \Pr\left[\mathbf{X} = \mathbf{x}_2\right] \\
&= \mathrm{CP}(\mathbf{X})|S|^q + (1 - \mathrm{CP}(\mathbf{X})) \\
&= 1 + \mathrm{CP}(\mathbf{X})(|S|^q - 1).
\end{aligned}$$

For the conditional case, i.e. $(\mathbf{X}, Z)$ is some conditionally simple $q$-wise $k$ source. We have that $\mathbf{X} \mid Z = z$ is a simple $q$-wise $k_z$-source. Let $A_z = (S, f(h(S, \mathbf{X})), Z) \mid Z = z, B_z = (S, f(\mathbf{U}), Z) \mid Z = z$. We compute that,

$$\begin{aligned} RD_2((S, f(h(S, \mathbf{X})), Z) \| (S, f(\mathbf{U}), Z)) &= \mathrm{E}_z[RD_2(A_z \| B_z)] \\ &= \mathrm{E}_z[1 + \mathrm{CP}(\mathbf{X} \mid Z = z)(|S|^q - 1)] \\ &= 1 + \mathrm{CP}(\mathbf{X} \mid Z)(|S|^q - 1). \end{aligned}$$

We obtain the bound claimed, for any conditionally simple $q$-wise $k$-source $(X, Z)$, using Lemma 4.

$$\begin{aligned} \mathsf{SD}((S, f(h(S, \mathbf{X})), Z); (S, f(\mathbf{U}), Z)) &\leq \sqrt{RD_2((S, f(h(S, \mathbf{X})), Z) \| (S, f(\mathbf{U}), Z)) - 1} \\ &\leq \frac{1}{2}\sqrt{2^{-k}(|S|^q - 1)}. \end{aligned}$$

□

**Comparison with previous results** Our bound, $\frac{1}{2}\sqrt{|S|^q 2^{-k}}$, is tighter for the multi-input case compared to the bound, $\frac{1}{2}\sqrt{|S|^q q^2 2^{-k}}$, obtained in [FOR15] which utilizes a union bound. Our technical contribution is the definition of (conditionally) simple sources and showing that they are the extreme points for (conditional) $q$-wise $k$-sources. However, we note that our version does not handle almost $k$-wise independent hash function directly. Instead, if the statistical distance definition of being almost $k$-wise independent is used, one can still use our result.

## 2.3 Cryptographic Background

Cryptographic primitives and constructions are *systems* with states, inputs, and outputs[2]. One way to formalize them is the framework of random systems [Mau02]. Formally speaking, a $(\mathcal{X}, \mathcal{Y})$-system, $\mathbf{S}$, is a system with input space $\mathcal{X}$ and output space $\mathcal{Y}$. A system's behavior is specified by a infinite list of conditional probability distributions, $P^{\mathbf{S}}_{Y_i|X_i,Y_{i-1}}$. In this thesis, all primitives and constructions of interest are stateless. It is easy to see that such systems are convex combination of deterministic systems. Such systems are call *cc-stateless* in the random systems literature. We will refer to such systems as *keyed* systems.

A quick and efficient way to describe systems in a precise manner is using pseudocode. A randomized procedure taking input shall specify a random system in the following manner: variable shall be shared between successive calls of the procedure, and the return value shall be the output of the system. Cryptographic definitions and reductions are also easily captured with pseudocode. In particular, we use the framework of code-based game playing [BR08]. A game, $G$, is consists of procedures and named oracles. The code shall always be specified in pseudocode. The entry point of the game shall be implicitly the procedure "Main". State between different oracles are not shared. At the end, a game usually outputs a bit, $G \Rightarrow b, b \in \{0,1\}$. In addition to its brevity, it works naturally to the style of "hybrid" proofs. One can change part of the pseudocode of game $G$ to game $G'$. By carefully selecting the changed code, one can bound the distance between games, $|\Pr[G \Rightarrow 1] - \Pr[G \Rightarrow 0]|$.

[2] While some system can be realized as a randomized Turing machine, not all systems need to be efficient implementable using Turing machines.

### 2.3.1 Computational Indistinguishability

We follow the standard definition in literature. A function, $f(\lambda)$, is said to be negligible if for any $n \in \mathbb{N}$ there exists $m \in \mathbb{N}$ such that $f(\lambda) \leq \frac{1}{\lambda^n}$ for all $\lambda > m$. Two distribution ensembles, $\{A_\lambda\}, \{B_\lambda\}$, are said to be computationally indistinguishable if for any polynomial time algorithm, $\mathcal{D}$, the function

$$\left|\Pr\left[\mathcal{D}(A_\lambda) \Rightarrow 1\right] - \Pr\left[\mathcal{D}(B_\lambda) \Rightarrow 1\right]\right|,$$

is negligible in $\lambda$ (recall that $\mathcal{D}$ takes $1^\lambda$ as an implicit input).

### 2.3.2 Hybrid Argument

The hybrid argument is a technique to bound the closeness of two distributions, $D_0$ and $D_n$, via a sequence of "hybrids", $D_1, \ldots, D_n$. Two consecutive hybrids, $D_i$ and $D_{i+1}$, usually differ in one feature. Using the triangle inequality (for statistical distance and computational distance), one then obtains a bound of the distance between $D_0$ and $D_n$ by bounding the distance between $D_i$ and $D_{i+1}$ for all $i = 0, \ldots, n-1$.

**Statistical Hybrid Argument** Suppose we have some joint distribution $\mathbf{X} = (X_1, \ldots, X_n)$ with each $X_i$ distributed on some set $\mathcal{X}$. We would like to bound the distance of the distribution, $\mathcal{X}$, from uniform, $\mathbf{U} = (U_\mathcal{X}, \ldots, U_\mathcal{X})$. Conventional hybrid method utilizes bounds

$$\mathsf{SD}(X_i; U_\mathcal{X} \mid X_1, \ldots, X_{i-1}) \leq \epsilon_i, \quad \text{for all } 1 \leq i \leq n,$$

to argue that

$$\mathsf{SD}(\mathbf{X}; \mathbf{U}) \leq \epsilon_1 + \ldots + \epsilon_n.$$

Such method can be seen as bounding between the distance between $n + 1$ hybrids, i.e.

$$\begin{aligned}(X_1, \ldots, X_n) &= \mathbf{X} \\ (U_{\mathcal{X}}, \ldots, X_n) & \\ \vdots & \\ (U_{\mathcal{X}}, \ldots, U_{\mathcal{X}}) &= \mathbf{U}.\end{aligned}$$

For simplicity, suppose all $n$ bounds are the same, say $\epsilon$. This method yield an overall error of $n\epsilon$, which grows $O(n)$ as $n$ increases.

### 2.3.3 Blockciphers

Blockciphers, which are key-ed permutations, are widely used primitives in cryptography. Practical constructions of blockciphers such as DES or AES consists of multiple rounds of simpler key-ed permutations. The round keys are, in general, derived from the blockcipher key. The heuristics is that compositing different permutations will make the construction appear more “random”, hence amplifying the security.

**Random Permutation** In this thesis, by a secure blockcipher we mean indistinguishability from a random permutation, under a uniformly random chosen key that is hidden from the adversary. Let $\mathcal{A}$ be any oracled randomized algorithm. Let **RP** denote the random permutation, whose code is specified in Figure 2.1. Let **P** be any key-ed permutation. We write $\mathcal{A}^{\mathbf{P}}$ to mean that we run $\mathcal{A}$ and offering it two sided oracle (both forward and backward queries) access of **P**.

Security proofs of blockciphers are generally offered in two steps. First, one uses a computational assumption to replace a underlying primitive with a information the-

```
Proc $\mathbf{RP}[D](x)$ :
if $T[x] = \perp$ then
  $T[x] \xleftarrow{\$} D \setminus B$
  $B = B \cup \{T[x]\}$
return $T[x]$
```

Figure 2.1: Pseudocode implementing $\mathbf{RP}[D]$, where $D$ is the domain.

oretical counterpart. Second, the construction is proved secure assuming information theoretical primitives. The classic example is the security proof of the Feistal construction assuming PRF security of the primitive [LR88].

Let $\mathbf{S}, \mathbf{T}$ be two keyed permutations. We consider two types of distinguishers, non-adaptive and adaptive distinguisher. A non-adaptive distinguisher is specified by the set of query points. An adaptive distinguisher can adaptively select query points based on the output of the cipher. Without loss of generality, we do not count redundant queries by a distinguisher. By convention, we will use $\mathcal{D}$ to denote a non-adaptive distinguisher and $\mathcal{A}$ to denote an adaptive one. Usually, we will also use $q$ to denote the number of queries a distinguisher makes.

### Transcripts and Security of Blockcipehrs

We associate each interaction, e.g. $\mathcal{D}^{\mathbf{S}}$, with a transcript of the form

$$\tau = \big((x_1, y_1), (x_2, y_2), \ldots, (x_q, y_q)\big) ,$$

where we assume that without loss of generality the distinguisher makes no duplicate queries. Now, we fix two systems, say $\mathbf{S}, \mathbf{U}$, where $\mathbf{U}$ is the ideal system (e.g. a PRP), and a distinguisher $\mathcal{D}$. We denote the set of potential transcripts (attainable with non-zero probability) as $\mathbf{T}(\mathcal{D}^{\mathbf{S}})$ and $\mathbf{T}(\mathcal{D}^{\mathbf{U}})$ respectively. We now make the assumptions that $\mathbf{T}(\mathcal{D}^{\mathbf{S}}) \subseteq \mathbf{T}(\mathcal{D}^{\mathbf{U}})^3$. For each set of potential transcripts, $\mathbf{T}$, we have a natural probability assignment, $P_{\mathbf{T}}(\tau)$, assigning each transcript to the probability of attaining it. We define the distinguishing advantage of $\mathcal{D}$ against two systems to be.

$$\mathsf{Adv}^{\mathrm{dist}}_{\mathbf{S},\mathbf{U}}(\mathcal{D}) := \|P_{\mathbf{T}(\mathcal{D}^{\mathbf{S}})}(\cdot) - P_{\mathbf{T}(\mathcal{D}^{\mathbf{U}})}(\cdot)\|_1,$$

Now, we recall the standard advantage of non-adaptively security of weak PRP and adaptive security of strong PRP.

$$\mathsf{Adv}^{\mathrm{nprp}}_{\mathbf{S}}(q) := \max_{\text{non-adaptive q-query } \mathcal{D}} \mathsf{Adv}^{\mathrm{dist}}_{\mathbf{S},\mathbf{U}}(\mathcal{D}),$$

$$\mathsf{Adv}^{\pm\mathrm{prp}}_{\mathbf{S}}\, q := \max_{\text{adaptive q-query } \mathcal{A}} \mathsf{Adv}^{\mathrm{dist}}_{\mathbf{S},\mathbf{U}}(\mathcal{A}).$$

[3] This will be true for our setting, and can always be made true by conditioning on some good event.

# Chapter 3

# Tighter Security Proofs Using Collision Probability and Chi-Square Distance

Proofs using statistical distance is ubiquitous in cryptography. The main advantage of statistical distance is two-fold. First, they are the statistical analog of the computational distance. Second, they satisfy data processing inequality and triangular inequality. In this chapter, we explores proof techniques using collision probability, CP, and chi-squared distance,

$$\chi(X;Y) := \sqrt{\sum_{\omega\in\Omega} \frac{(P_X(\omega) - P_Y(\omega))^2}{P_Y(\omega)}} = \sqrt{RD_2(X\|Y) - 1}.$$

## 3.1 Tighter Hybrid Argument

It was first observed by Chung and Vadhan [CV08] that the bound of statistical distance $(X_1, \ldots, X_n)$ to $(U_1, \ldots, U_n)$ only needs to grow $\sqrt{n}$ larger than the individual distance between $X_i$ and $U_i$. Their proof uses Hellinger distance and repetitive applications of Holder's inequality. Here, we prove a more general version via Rényi divergence. The main different is, in our Theorem 2, one can use component bound of the form $\mathrm{CP}(X_i \mid X_{<i}) \leq \frac{1+\epsilon}{M}$ where $M$ is the support size of $X_i$ conditioned on some given value of $X_{<i}$ ($X_{<i}$ denotes $X_1, \ldots, X_{i-1}$). This enables us to apply this result to the case where the joint distribution describes a permutation, i.e. $X_i \neq X_j$ for $i \leq j$.

**Theorem 2.** *Let $\mathcal{X}$ be some finite set. Let $\mathbf{U} = (U_1, \ldots, U_n)$ be some uniform distribution, where each component, $U_i$, is distributed over $\mathcal{X}$. Let $m_i = \max_{x_{<i}} |U_i \mid U_{<i} = x_{<i}|$. Let $\mathbf{X} = (X_1, \ldots, X_n)$ be jointly distributed over the support of $\mathbf{U}$. Suppose that $\mathrm{CP}(X_i \mid X_{<i}) \leq \frac{1+\epsilon_i}{m_i}$, for all $i = 1, \ldots, n$. Then,*

$$\mathsf{SD}(\mathbf{X}; \mathbf{U}) \leq \sqrt{\frac{\ln(2)}{2} \sum_{i=1}^{n} \epsilon_i}.$$

Before we offer the proof, let us look at two special cases of the Lemma. First, let $m_i = |\mathcal{X}|$ for all $i$. In this case, the Lemma captures the same result as [CV08]. Second, take $m_i = |\mathcal{X}| - i$, and that $\mathbf{U}$ represents the (possibly partial) function table of a permutation. This case will be crucial for application to block ciphers. Third, if we take $\epsilon_i = \epsilon$ for all $i$, our final bound is $\sqrt{\frac{\ln(2)}{2} n\epsilon}$, which is $O(\sqrt{n})$ with respect to $n$.

*Proof (of Theorem 2).* The key idea here is to bound the KL divergence, or the total Shannon entropy deficiency, with bounds on collision probabilities. By Lemma 1 and Corollary 1,

$$-\log(\mathrm{CP}(X \mid Y)) \leq H_2(X \mid Y) \leq H_1(X \mid Y).$$

```
Proc $E_{\mathbf{K},}(X)$ :
  for $i = 0$ to $r$ do
    $X' \leftarrow \mathbf{K}[i] \oplus X$; $\hat{X} \leftarrow \max(X, X')$
    if $_i(\hat{X}) = 1$ then $X \leftarrow X'$
  return $X$
```

Figure 3.1: The Swap-or-Not Cipher, $\mathsf{SN}[N, r]$, which takes $r$ rounds keys $\mathbf{K}[1], \ldots, \mathbf{K}[r]$ and round functions $_{1,r}$.

Suppose that $\mathrm{CP}(X_i \mid X_{<i}) \leq \frac{1+\epsilon_i}{m_i}$. We compute

$$\begin{aligned}
D_1(\mathbf{X} \| \mathbf{U}) &\leq \log(|(X_1, \ldots, X_n)|) - H_1(X_1, \ldots, X_n) && \text{(Lemma 2)} \\
&\leq \sum_{i=1}^{n} \log(m_i) - \sum_{i=1}^{n} H_1(X_i \mid X_{<i}) \\
&\leq \sum_{i=1}^{n} \log(m_i) + \sum_{i=1}^{n} \log(\mathrm{CP}(X_i \mid X_{<i})) \\
&\leq \sum_{i=1}^{n} \log(m_i) + \sum_{i=1}^{n} \log(\frac{1+\epsilon_i}{m_i}) \\
&\leq \sum_{i=1}^{n} \log(1+\epsilon_i) \\
&\leq \sum_{i=1}^{n} \epsilon_i && (\log(1+x) = x - \frac{x^2}{2} + O(x^3)).
\end{aligned}$$

Finally, by Pinsker's inequality (Lemma 3),

$$\mathsf{SD}(\mathbf{X}; \mathbf{U}) \leq \sqrt{\frac{\ln(2)}{2} D_1(\mathbf{X} \| \mathbf{U})} = \sqrt{\frac{\ln(2)}{2} \sum_{i=1}^{n} \epsilon_i}.$$

□

### 3.1.1 Applications to Swap-Or-Not

The motivation for the swap-or-not construction (see Figure 3.1) was format-preserving encryption [HMR12]. In format-preserving encryption, one assumes some group struc-

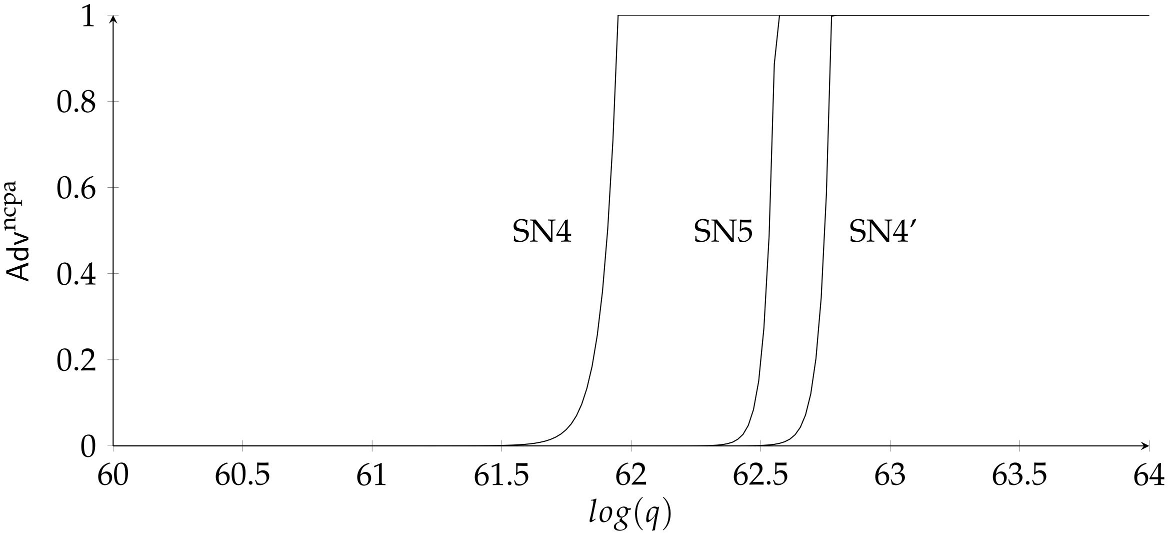

Figure 3.2: Comparison of bounds of Swap-or-Not. The domain size is $N = 2^{64}$. The plit “SN4” and “SN5” represents bounds of npca advantage of npca advantage of npca advantage of npca advantage obtained in [HMR12] for $\mathsf{SN}[N, 4n]$ and $\mathsf{SN}[N, 5n]$, respectively. “SN4′′ represents the bound obtained in this work for $\mathsf{SN}[N, 4n]$. One can see the our bound saves at least one pass in this setting of parameters.

ture on the domain and some total ordering. The goal is to generate a key-ed permutation on the domain that is indistinguishable from a random permutation with the maximum amount of queries.

We use the Theorem 3 proved in [HMR12], together with Theorem 2 to prove the following result.

**Theorem 3.**

$$\mathsf{Adv}^{\mathrm{nprp}}_{\mathsf{SN}[N,r]}(q) \leq \sqrt{\frac{\ln(2)}{2} q N (\frac{1}{2} + \frac{q}{2N})^{r/2}}.$$

We first observe the following identity.

**Claim 1.** *Let $X$ be a distribution over $\{0,1\}^n$. Then,*

$$\|X - U_d\|_2^2 = \|X\|_2^2 - \frac{1}{N}.$$

*Proof (of Claim 1).* Compute that

$$\begin{aligned}\|X - U_n\|_2^2 &= \|X\|_2 + \|U_n\|_2 - 2\langle X, U_n\rangle \\ &= \|X\|_2 - \frac{1}{N}.\end{aligned}$$

□

*Proof (of theorem 3).* We re-write Equation (5) obtained in [HMR12] in our notation.

$$\mathrm{E}_{X_{t,<i}}[\|X_{t,i} \mid X_{t,<i} - U_{t,i}\|_2^2] = (\frac{i+N}{2N})^t.$$

We note that this translates to a bound of $\mathrm{CP}(X_{t,i} \mid X_{t,<i}) \leq \frac{1+(\frac{i+N}{2N})^t \cdot N_i}{N_i}$ via Claim 1. We apply Theorem 2 to obtain the final bound.

$$\begin{aligned}\mathsf{Adv}^{\mathrm{ncpa}}_{\mathsf{SN}[N,r]}(q) &\leq \mathsf{SD}(\mathbf{X}_r, \mathbf{U}) \\ &\leq \sqrt{\frac{\ln(2)}{2}\sum_{i=1}^{q}(\frac{i+N}{2N})^r \cdot N_i} \\ &\leq \sqrt{\frac{\ln(2)}{2} qN(\frac{1}{2}+\frac{q}{2N})^r}.\end{aligned}$$

□

## 3.2 Tighter Composition Theorem via Chi-Square Distance

### 3.2.1 Adaptive Versus Non-Adaptive Security

In the previous section, we proved tighter bound for non-adaptive security of Swap-Or-Not.

One line of work [MPR07, CPS14, MP04] studies the security properties of block ciphers under composition assuming information theoretical secure rounds. In [HMR12], the adaptive security of Swap-Or-Not is proved using the below theorem. Intuitively, it says that, with double the rounds, you get adaptive security with essentially the same parameters.

**Theorem 4.** *[MPR07]*

$$\mathsf{Adv}^{\mathrm{prp}}_{\mathbf{S}^{-1}\circ\mathbf{T}}(q) \leq \mathsf{Adv}^{\mathrm{nprp}}_{\mathbf{S}^{-1}}(q) + \mathsf{Adv}^{\mathrm{nprp}}_{\mathbf{T}}(q).$$

Here, we offer evidence that working 2-norm related distances can improve such a bound. In particular, we will define the Chi-Square distance. We fix some domain $\mathcal{X}$. Let $\mathbf{S}$ be a keyed permutation on $\mathcal{X}$, and $\mathbf{U} = \mathbf{RP}[\mathcal{X}]$. Recall that, for some distinguisher $\mathcal{D}$, we use $\mathbf{T}(\mathcal{D}^{\mathbf{U}})$ to denote the distribution of transcripts of an interaction between $\mathcal{D}$ and $\mathbf{U}$. We define

$$\mathsf{Adv}^{\mathrm{2nprp}}_{\mathbf{S}}(q) := \max_{\text{non-adaptive q-query } \mathcal{D}} \chi(\mathbf{T}(\mathcal{D}^{\mathbf{S}}); \mathbf{T}(\mathcal{D}^{\mathbf{U}})),$$

$$\mathsf{Adv}^{\mathrm{2prp}}_{\mathbf{S}}(q) := \max_{\text{adaptive q-query } \mathcal{A}} \chi(\mathbf{T}(\mathcal{A}^{\mathbf{S}}); \mathbf{T}(\mathcal{A}^{\mathbf{U}})).$$

We prove a stronger version using the Chi-Square advantage.

**Theorem 5.**

$$\mathsf{Adv}^{\text{2prp}}_{\mathbf{S}\circ\mathbf{T}}(q) \le \mathsf{Adv}^{\text{2nprp}}_{\mathbf{S}^{-1}}(q) \cdot \mathsf{Adv}^{\text{2nprp}}_{\mathbf{T}}(q).$$

*Proof.* For any transcripts of the form,

$$\tau = \left((x_1, y_1), (x_2, y_2), \ldots, (x_q, y_q)\right).$$

We consider the *decomposition*, $\tau = (\mathbf{x}, \mathbf{y})$, where $\mathbf{x}$ consists of the $q$ input values (could be a result from a backward query), and $\mathbf{y}$ consists of $q$ output values (could be the input of a backward query). Notice that, for non-adaptive distinguishers, the probability of attaining $\mathbf{x}$ is specified by $\mathcal{D}$ and *independent* of $\mathbf{y}$. We use $\mathbf{T}[\mathbf{x}, \mathbf{y}]$ to denote the probability that $\mathbf{T}$ maps $\mathbf{x}$ to $\mathbf{y}$. We let $p^* = \frac{1}{(M)_q}$ be the probability mass of each transcript in $\mathbf{T}(\mathbf{U})$. Notice that for any $\mathbf{x}, \mathbf{y}$, $\sum_{\mathbf{z}}(\mathbf{T}[\mathbf{x}, \mathbf{z}] - p^*) = \sum_{\mathbf{z}}(\mathbf{S}[\mathbf{z}, \mathbf{y}] - p^*) = 0$. If we fix a *deterministic* $q$-query distinguisher $\mathcal{A}$. Then, the set of potential transcript for $\mathcal{A}^{\mathbf{U}}$ is of size $M(M-1)\cdots(M-q+1) := (M)_q$, since conditioned on $t$ queries, there are always $M - t + 1$ possibilities for the next $(x, y)$ regardless of the direction of the queries. Hence, non-adaptive and adaptive $q$-query distinguisher will have potential transcripts of the same cardinality. We use $\mathbf{T}(\mathbf{U})$ to denote $\mathbf{T}(\mathcal{A}^{\mathbf{U}})$. We compute that,

$$
\begin{aligned}
\mathsf{Adv}^{2\pm\mathrm{prp}}_{\mathbf{S}^{-1}\circ\mathbf{T}}(q) &= \sqrt{|\mathbf{T}(\mathbf{U})| \sum_{\mathbf{x},\mathbf{y}} \left(\mathbf{S}^{-1} \circ \mathbf{T}[\mathbf{x},\mathbf{y}] - \mathbf{U}[\mathbf{x},\mathbf{y}]\right)^2} \\
&= \sqrt{|\mathbf{T}(\mathbf{U})| \sum_{\mathbf{x},\mathbf{y}} \left( \sum_{\mathbf{z}} \mathbf{T}[\mathbf{x},\mathbf{z}] \cdot \mathbf{S}^{-1}[\mathbf{z},\mathbf{y}] - p^* \right)^2} \\
&= \sqrt{|\mathbf{T}(\mathbf{U})| \sum_{\mathbf{x},\mathbf{y}} \left( \sum_{\mathbf{z}} (\mathbf{T}[\mathbf{x},\mathbf{z}] - p^* + p^*) \cdot (\mathbf{S}^{-1}[\mathbf{z},\mathbf{y}] - p^* + p^*) - p^* \right)^2} \\
&= \sqrt{|\mathbf{T}(\mathbf{U})| \sum_{\mathbf{x},\mathbf{y}} \left( \sum_{\mathbf{z}} (\mathbf{T}[\mathbf{x},\mathbf{z}] - p^*) \cdot (\mathbf{S}^{-1}[\mathbf{z},\mathbf{y}] - p^*) \right)^2} \\
&\leq \sqrt{|\mathbf{T}(\mathbf{U})| \sum_{\mathbf{x},\mathbf{y}} \left( \sqrt{\sum_{\mathbf{z}} (\mathbf{T}[\mathbf{x},\mathbf{z}] - p^*)^2} \sqrt{\sum_{\mathbf{z}} (\mathbf{S}^{-1}[\mathbf{z},\mathbf{y}] - p^*)^2} \right)^2} \\
&\leq \sqrt{|\mathbf{T}(\mathbf{U})|^2 \cdot \left( \frac{1}{\sqrt{|\mathbf{T}(\mathbf{U})|}} \mathsf{Adv}^{2-\mathrm{nprp}}_{\mathbf{T}}(q) \cdot \frac{1}{\sqrt{|\mathbf{T}(\mathbf{U})|}} \mathsf{Adv}^{2-\mathrm{nprp}}_{\mathbf{S}^{-1}}(q) \right)^2} \\
&\leq \mathsf{Adv}^{2-\mathrm{nprp}}_{\mathbf{S}^{-1}}(q) \cdot \mathsf{Adv}^{2-\mathrm{nprp}}_{\mathbf{T}}(q).
\end{aligned}
$$

□

### 3.2.2 Properties of Chi-Square Distinguishing Advantage

**Behavior under Conditioning**

Let $X, Y, Z$ be random variables. For notational convenience, we define

$$\chi(X, Y \mid Z) := \chi((X, Z);\ (Y, Z)).$$

Unlike $\|\cdot\|_2$, $\chi$ behaves well under conditioning.

**Lemma 7.** *If $Z$ is independent of $X, Y$, then*

$$\chi(X, Y) = \chi(X, Y \mid Z).$$

*Proof.*

$$\begin{aligned}
\chi(X, Y \mid Z) &= \sqrt{\sum_{a,b} \frac{(P_X(a)P_Z(b) - P_Y(a)P_Z(b))^2}{P_Y(a)P_Z(b)}} \\
&= \sqrt{\sum_{a,b} Z[b] \frac{(P_X(a) - P_Y(a))^2}{P_Y(a)}} \\
&= \sqrt{\sum_{a} \frac{(P_X(a) - P_Y(a))^2}{P_Y(a)} (\sum_b P_Z(b))} \\
&= \chi(X, Y).
\end{aligned}$$

□

**Double-Sided Point-Wise Proximity**

In the section, we show how to bound $\chi$ using double-sided point-wise proximity.

**Lemma 8.** *Let $X$ be a random variable, and let $U$ be the uniform over the same support. We say that $X$ satisfy $(\epsilon, \delta)$-proximity if for all $x$,*

$$(1 - \epsilon)\frac{1}{|U|} \le \Pr[X = x] \le (1 + \delta)\frac{1}{|U|}.$$

*If $X$ satisfy $(\epsilon, \delta)$-proximity, then*

$$\chi(X, U) \le \sqrt{\epsilon\delta}.$$

*Proof.* We rewrite $\chi$ in terms of variance, and apply Bhatia–Davis inequality [BD00].

$$\begin{aligned}
\chi(X,U) &= \sqrt{|U|(\sum_{\omega} P_X(\omega(^2 - \frac{1}{|U|})} \\
&= \sqrt{|U|^2(\sum_{\omega} \frac{1}{|U|} P_X(\omega)^2 - (\sum_{\omega} \frac{1}{|U|} P_X(\omega))^2} \\
&= \sqrt{|U|^2 \mathsf{Var}[P_X(U)]} \\
&\leq \sqrt{|U|^2(\max_x \Pr[X = x] - \frac{1}{|U|})(\frac{1}{|U|} - \min_x \Pr[X = x])} \\
&\leq \sqrt{|U|^2 \epsilon \frac{1}{|U|} \delta \frac{1}{|U|}} \\
&\leq \sqrt{\epsilon\delta}.
\end{aligned}$$

□

### 3.2.3 Potential Application to Swap-Or-Not

To apply our new composition theorem to Swap-Or-Not, we need a bound on $\mathsf{Adv}^{\mathrm{2nprp}}$. In our previous setup of notation, it amounts to bounding

$$\chi(\mathbf{X}_t, \mathbf{U}).$$

However, the technique from [HMR12] and [Tes14] does not apply directly to $\chi$. The main difficulty is that $\chi$ does not satisfy triangle inequality and one cannot apply the hybrid technique. We leave it as an open problem.

# Chapter 4

# Leftover Hash Lemma for Joint Leakage

The classical leftover hash lemma [HILL99] has seen wide usage in cryptography, especially in leakage resilient cryptography. Here, inspired by the Universal Computational Extractors assumption [BHK13], we derive a variant of Leftover Hash Lemma for joint leakage.

## 4.1 Preliminaries

### 4.1.1 Universal Computational Extractors

Universal Computational Extractor (UCE), is a security assumption on hash functions, captured by the security game in Figure 4.1. Let

$$\mathsf{Adv}^{\mathrm{UCE}}_{H,\mathcal{S},\mathcal{A}}(\lambda) := 2\Pr\left[\mathsf{UCE}[H,\mathcal{S},\mathcal{A}] \Rightarrow 1\right] - 1.$$

With slight abuse of notation, we say that $\mathsf{H} \in \mathsf{UCE}[\mathcal{S}]$ if $\mathsf{Adv}^{\mathrm{UCE}}_{H,\mathcal{S},\mathcal{A}}$ is negligible for all efficient $\mathcal{A}$.

```
Game $\mathsf{UCE}[H, \mathcal{S}, \mathcal{A}]$
  $b \xleftarrow{\$} \{0,1\}$
  $\mathsf{hk} \xleftarrow{\$} H.\mathsf{Gen}()$
  $z \xleftarrow{\$} \mathcal{S}^{\mathsf{Hash}[\mathsf{hk},b]}()$
  $b' \xleftarrow{\$} \mathcal{D}(\mathsf{hk}, z)$
  return $(b = b')$
```

```
Proc $\mathsf{Hash}[\mathsf{hk}, b](x)$ :
  if $T[x] = \bot$ then
    if $b = 0$ then
      $T[x] \leftarrow H_{\mathsf{hk}}(x)$
    else
      $T[x] \xleftarrow{\$} \{0,1\}^{H.\mathsf{ol}}$
  return $T[x]$
```

Figure 4.1: Game defining UCE security for a hash family $H$, source $\mathcal{S}$, and distinguisher $\mathcal{D}$.

**UCE Sources** A UCE source, $\mathcal{S}$, is different compared to a source specified in Section. A UCE source, $\mathcal{S}$, is a randomized oracle algorithm, taking exactly one oracle, $\mathsf{Hash}$. $\mathcal{S}$ shall make some queries to $\mathsf{Hash}$ before returning some leakage. UCE is a class of assumptions. For certain UCE sources, the notion is not achievable. For instance, the UCE source, $\mathcal{S}$ could leak the query point, $x$, along with the hash value $y = \mathsf{H}(x)$. Then, with all by $\frac{1}{2^n}$ probability, an adversary can predict $b$ in the UCE game by checking if $\mathsf{H}(x) = y$. The study of UCE involves exploring the feasibility, construction, and applications of UCE for different class of sources. Several interesting restrictions of sources has been studied.

**Split UCE Sources** A split UCE source $\mathcal{S} = \mathsf{splt}[\mathcal{S}_0, \mathcal{S}_1]$ is one that is generated in two stages. First, $\mathcal{S}_0$ generates a list of input points $\mathbf{x}$ with leakage $L_0$. Then, the hash values $\mathbf{y} = \mathsf{Hash}(\mathbf{x})$ is given to $\mathcal{S}_1$, producing leakage $L_1$. The overall leakage is $L = (L_0, L_1)$. The pseudocode of $\mathsf{splt}[\mathcal{S}_0, \mathcal{S}_1]$ is given in Figure 4.2. The class of split sources is denoted $\mathcal{S}^{\mathsf{splt}}$.

**Statistical Unpredictable UCE Sources** A UCE source, $\mathcal{S}$ is statistically unpredictable if the query points $\mathbf{x}$ of $\mathcal{S}$ is statistically unpredictable given the output of $\mathcal{S}$. More

```
Source $\mathsf{splt}[\mathcal{S}_0, \mathcal{S}_1]^{\mathsf{Hash}}$ :
$(L_0, \mathbf{x}) \leftarrow \mathcal{S}_0()$
for $i = 0, \ldots, |\mathbf{x}|$ do
  $\mathbf{y}[i] \leftarrow \mathsf{Hash}(\mathbf{x}[i])$
$L_1 \overset{\$}{\leftarrow} \mathcal{S}_1(\mathbf{y})$
$L \leftarrow (L_0, L_1)$
return $L$
```

Figure 4.2: Pseudocode defining the split source $\mathcal{S} = \mathsf{splt}[\mathcal{S}_0, \mathcal{S}_1]$.

concretely, for $L \overset{\$}{\leftarrow} \mathcal{S}(1^\lambda)$ and $\mathbf{x}$ be the query points of $\mathcal{S}$, we require

$$H_\infty(\mathbf{x} \mid L)$$

to be a negligible function of $\lambda$. The class of statistically unpredictable sources is denoted $\mathcal{S}^{\mathrm{sup}}$.

**Other conventions** We can also classify UCE sources by the number of distinct queries that it makes to the hash function. We use $\mathcal{S}^q$ to denote the class of sources making exactly $q$ distinct queries.

### Connections to Extractors

By construction, the assumption $\mathsf{UCE}[\mathcal{S}^1, \mathcal{S}^{\mathsf{splt}}, \mathcal{S}^{\mathrm{sup}}]$ is essentially requiring a family of hash function to be "extractors".

## 4.2 Leftover Hash Lemma for Joint Leakage

Let us look up the assumption without the $S^{\mathsf{splt}}$ requirement, i.e. $\mathsf{UCE}[\mathcal{S}^1, \mathcal{S}^{\mathrm{sup}}]$. We name this case "joint leakage", for the leakage depends on both the input and output to the hash function. Since there is only one query, we let the query point be $x$. Let $u$

be uniform in the output space of Hash. Then, the assumption $\mathsf{UCE}[\mathcal{S}^1, \mathcal{S}^{\text{sup}}]$ translates to,

$$(s, L(x, h_s(x))) \approx_c (s, L(x, u)).$$

We now prove that universal hash functions is $\mathsf{UCE}[\mathcal{S}^1, \mathcal{S}^{\text{sup}}]$ for appropriate parameters.

**Theorem 6** (Leftover Hash Lemma for Joint Leakge)**.** *Let $k = H_\infty(x \mid L(x, u))$. Let $h$ be an universal hash function with output length $m$. Let $u$ be uniform in the output space of $h$. Let $s$ be a uniform random seed for $h$. Then,*

$$(s, L(x, u)) \approx_{s,\epsilon} (s, L(x, h_s(u))),$$

*where $\epsilon = \frac{1}{2}\sqrt{2^{4m-k}}$.*

*Proof.* By Lemma 5,

$$H_\infty(x \mid L(x, u), u) \geq k - m.$$

Notice that since $s$ is independent from $x$. Let $u'$ be an i.i.d copy of $u$. By the Leftover Hash Lemma (Lemma 1),

$$(s, L(x, u), u, h_s(x)) \approx_{s,\epsilon} (s, L(x, u), u, u'), \tag{4.1}$$

where

$$\epsilon = \frac{1}{2}\sqrt{2^{m-(k-m)}} = \frac{1}{2}\sqrt{2^{2m-k}}.$$

Let $A, B$ be the two distribution on left and right side of Equation 4.1, respectively. Let $\Omega$ be the sample space of $A$ and $B$. Note that $\Omega$ consists of a 4-tuples. Define event $E \subseteq \Omega$ to be event where the third and fourth component collide, i.e. $E = \{(a, b, c, d) \in \Omega \mid c = d\}$. Since $u$ is uniform and independent of $h(x)$, $\Pr[A \in E] = \frac{1}{2^m}$. Since $u$ is uniform and independent of $u'$, $\Pr[B \in E] = \frac{1}{2^m}$. By Lemma 9, we conclude that

$$A \mid E \approx_{s,\epsilon'} B \mid E,$$

where

$$\epsilon' = 2^m \frac{1}{2}\sqrt{2^{2m-k}} = \frac{1}{2}\sqrt{2^{4m-k}}.$$

□

**Lemma 9.** *Let $A, B$ be two distributions over $\Omega$. Let $E \subseteq \Omega$ be an event such that* $\Pr[A \in E] = \Pr[B \in E] = p$. *Then,*

$$\mathsf{SD}(A \mid E; B \mid E) \leq \frac{\mathsf{SD}(A; B)}{p}.$$

*Proof.* It is easily obtained by expanding the definition of statistical distance.

$$\begin{aligned}
\mathsf{SD}(A \mid E; B \mid E) &= \frac{1}{2}\sum_{\omega \in E}\left|\frac{P_A(\omega)}{P_A(E)} - \frac{P_B(\omega)}{P_B(E)}\right| \\
&= \frac{1}{p}\cdot\frac{1}{2}\sum_{\omega \in E}|P_A(\omega) - P_B(\omega)| \\
&\leq \frac{1}{p}\mathsf{SD}(A; B).
\end{aligned}$$

□

## 4.3 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\sup}]$ with Constant Output Length

In this section, we extend the result obtain from the previous section to multiple queries. This amounts to extending the joint-leakage LHL to $2q$-wise independent hash functions.

**Theorem 7** (Leftover Hash Lemma for Joint Leakge)**.** *Let $h$ be a $2q$-wise independent hash function. Let $\mathbf{u}$ be $q$ independently uniform samples from the output space of $h$. Let $s$ be a uniform random seed for $h$. Let $\mathbf{x}$ be $q$-ary distribution over the input space of $h$. Suppose that $H_\infty(\mathbf{x} \mid L(\mathbf{x}, \mathbf{u})) \geq k$. Then,*

$$(s, L(\mathbf{x}, \mathbf{u})) \approx_{s,\epsilon} (s, L(\mathbf{x}, h_s(\mathbf{x}))),$$

*where $\epsilon = \frac{1}{2}\sqrt{2^{4qm-k}}$.*

*Proof.* The proof is very similar to the proof of Theorem 6, hence we will be brief here. By Lemma 5, we have that $H_\infty(\mathbf{x} \mid L(\mathbf{x}, \mathbf{u}), \mathbf{u}) \geq k - qm$. We claim that,

$$(s, L(\mathbf{x}, \mathbf{u}), \mathbf{u}, h_s(\mathbf{x})) \approx_{s,\epsilon} (s, L(\mathbf{x}, \mathbf{u}), \mathbf{u}, \mathbf{u}'), \tag{4.2}$$

with $\epsilon = \frac{1}{2}\sqrt{2^{k-2qm}}$. This follows by computing that $\mathrm{CP}(h_s(\mathbf{x}) \mid s, L(\mathbf{x}, \mathbf{u}), \mathbf{u}) = \frac{1+2^{k-2qm}}{2^{qm}}$, which follows from the $2q$-wise independence of $h$. Lastly, we use 9 on the event $E = \{(a, b, c, d) \in \Omega \mid c = d\}$, where $\Omega$ is the sample space both distribution in Equation 4.2, to conclude the theorem. □

## 4.4 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\mathsf{sup}} \cap \mathcal{S}^{\mathsf{splt}}]$ with Polynomial Output Length

In this section, we attempt the construct multiple query UCE without compromising on the output length.

### 4.4.1 Extreme Lossy Functions (ELF)

**Definition 9.** *An extremely lossy function,* $\mathsf{ELF}$*, consists of an algorithm* $\mathsf{ELF.Gen}$*.* $\mathsf{ELF.Gen}$ *takes in two parameter $M$ (with $\log(M)$ implicitly being the security parameter) and $r \in [M]$, and outputs a description of a function $f : [M] \to [N]$, such that*

- *$f$ is efficiently computable (in $\log(M)$).*
- *If $r = M$, then $f$ is injective with overwhelming probability.*
- *If $r < M$, then $|f([M])| \leq r$ with overwhelming probability.*
- *For any polynomial $p$ and non-negligible $\epsilon$, there exists a polynomial $q$ such that the following holds: any adversary running at most time $p$ has less than $\epsilon$ advantage in distinguishing* $\mathsf{ELF.Gen}(M, M)$ *from* $\mathsf{ELF.Gen}(M, r)$ *for any $r \geq q(\log(M))$.*

### 4.4.2 Our construction

Let $h$ be an $2q$-wise independent function with output space $[M]$ and seed length $d_1$. Let $h'$ be an universal hash function with seed length $d_2$ and output length $\frac{1}{2}\log(M)$. Let $f$ be an ELF with input domain $[M]$. Consider the hash function construction in Figure 4.3.

**Theorem 8.** *Let $\mathcal{S}$ be a $q$-query split source with statistical unpredictability $k$. If $q \in O(k/\log(\lambda))$, the construction $H_1[q]$ is a secure* $\mathsf{UCE}[\mathcal{S}]$ *hash function.*

*Proof.* The proof proceeds by using hybrids, which are parameterized by $r$ that will by specified later. The following games will be based off of the UCE game, except that the games will return $b'$, the value returned by $\mathcal{D}$. Notice that, in this setup, $\Pr[G_0 \Rightarrow 1] - \Pr[G_4 \Rightarrow 1]$ is upper-bounded by the UCE advantage.

- $G_0$ is the UCE game with the hash function, i.e. $b = 0$ in game $\mathsf{UCE}[H_1[q], \mathcal{S}, \mathcal{D}]$.

```
Proc $H_1[q]$.KG() :
  $f \xleftarrow{\$} \mathsf{ELF.Gen}(M, M)$
  $s_1 \xleftarrow{\$} \{0,1\}^{d_1}$
  $s_2 \xleftarrow{\$} \{0,1\}^{d_2}$
  return $(\mathsf{fk}, s)$

Proc $H_1[q](\mathsf{hk}, x)$ :
  return $h'_{s_2}(f(h_{s_1}(x)))$
```

Figure 4.3: Construction $H_1[q]$. $h$ is a $2q$-wise independent hash function, $h'$ is a universal hash function.

- $G_1$, we switch $f$ to be lossy in the underlying construction of $H_1[q]$, i.e. $f \xleftarrow{\$} \mathsf{ELF.Gen}(M, r)$.
- $G_2$, we replace the $q$ inputs to $f$ to be uniformly chosen in the input space of $f$.
- $G_3$, we switch $f$ back to be injective in $H_1[q]$, i.e. $f \xleftarrow{\$} \mathsf{ELF.Gen}(M, M)$.
- $G_4$ is the UCE game with RO, i.e. $b = 1$ in the game $\mathsf{UCE}[H_1[q], \mathcal{S}, \mathcal{D}]$.

Suppose towards a contradiction that a distinguisher $\mathcal{D}$, with running time $t_{\mathcal{D}}$, together with source $\mathcal{S}$, with running time $t_{\mathcal{S}}$, has non-negligible advantage $\epsilon$ in winning the UCE game. We will reach contradiction by bounding the advantage through our sequence of hybrids above. Let $t = \mathsf{poly}(t_{\mathcal{D}}, t_{\mathcal{S}})$ to be the maximum running time of all five games $G_0, \ldots, G_4$. Let $\epsilon' = \epsilon/3$. By the ELF security, we can pick some $r \in \mathsf{poly}(\lambda)$ such that games $G_0, G_1$ and $G_2, G_3$ are close, i.e.

$$\Pr[G_0 \Rightarrow 1] - \Pr[G_1 \Rightarrow 1] < \epsilon', \tag{4.3}$$

$$\Pr[G_2 \Rightarrow 1] - \Pr[G_3 \Rightarrow 1] < \epsilon'. \tag{4.4}$$

It remains to bound the distance between $G_1, G_2$ and $G_3, G_4$. We observe that the distance between $G_1$ and $G_2$ can be bounded using the Generalized Leftover Hash Lemma 1, since $f$ is *independently* chosen here.

$$\Pr[G_1 \Rightarrow 1] - \Pr[G_2 \Rightarrow 1] \leq \mathsf{SD}((s, z, f(h_s(\mathbf{x}))); (s, z, f(\mathbf{u})) \leq \sqrt{\frac{r^q}{2^k}}.$$

This advantage is negligible for polynomial $r$ as long as $k - c \cdot q \log(\lambda) \in \omega(\log(\lambda))$. Which is achievable as long as $\frac{k}{\log(\lambda)} \in \omega(q)$. Observe that, in $G_3$, inputs to $h_2$ are independently random. Hence, we can simply apply the standard leftover hash lemma to obtain

$$\Pr[G_3 \Rightarrow 1] - \Pr[G_4 \Rightarrow 1] \leq \mathsf{SD}((s, z, h'(f(\mathbf{u}))); (s, z, \mathbf{u})) \leq 2^{-\frac{\log(M)}{4}}.$$

Now, we have that the UCE advantage of $\mathcal{A}$ is bounded by

$$\Pr[G_0 \Rightarrow 1] - \Pr[G_4 \Rightarrow 1] \leq \frac{2\epsilon}{3} + \sqrt{\frac{r^q}{2^k}} + 2^{-\log(M)/4},$$

which is asymptotically smaller than $\epsilon$. Hence, we reach a contradiction and $H_1[q]$ must be $\mathsf{UCE}[\mathcal{S}]$ secure. □

## 4.5 Construction of $\mathsf{UCE}[\mathcal{S}^q \cap \mathcal{S}^{\mathrm{sup}}]$ with Polynomial Output Length

To prove security against non-split sources, the main difficulty is dealing with joint leakage. We will use the same sequence of hybrid. The proof differ in bounding the distance of $G_2$ and $G_1$. To obtain the new bound, we observe the following variant of leftover hash lemma.

**Lemma 10** (Crooked Leftover Hash Lemma for Joint Leakage)**.** *Let $h$ be a $2q$-wise independent function with input space $[N]$. Let $\mathbf{X}$ be some distribution over $([N])^q$ and let $\mathbf{U}$ be uniform over $([N])^q$. Let $L$ be an arbitrary function such that $H_\infty(\mathbf{X} \mid L(\mathbf{X}, \mathbf{U})) \geq k$. Let $r$ be the bound on the image size of $f$. Then,*

$$\mathsf{SD}(s, L(\mathbf{X}, f(\mathbf{U})); s, L(\mathbf{X}, f(h_s(\mathbf{X})))) \leq (r + \frac{r^2}{1 - r\epsilon})\epsilon,$$

*where $\epsilon = \frac{1}{2}\sqrt{\frac{r^q}{2^k}}$. In particular, when $r$ is polynomial (in the security parameter), the distance is negligible.*

*Proof.* First we observe that since $s$ is independently picked, $H_\infty(X \mid s, L(X, f(\mathbf{u}))) \geq k$. By the leakage lemma, $H_\infty(X \mid s, L(X, f(\mathbf{u})), f(\mathbf{u})) \geq k - m$. Hence,

$$s, L(X, f(\mathbf{u})), f(\mathbf{u}), f(h(\mathbf{x})) \approx_{s,\epsilon} s, L(X, f(\mathbf{u})), f(\mathbf{u}), f(\mathbf{u}')$$

We define the event $E$ to be the event that the third component collides with the fourth component. The right distribution admits event $E$ with probability $\mathrm{CP}(f(\mathbf{u})) \geq \frac{1}{r}$. The left distribution admits event $E$ with probability within $\mathrm{CP}(f(\mathbf{u})) \pm \epsilon$. Observe that we need to compute $\mathsf{SD}(A \mid E; B \mid E)$. The result follows from the lemma below and the Generalized Leftover Hash Lemma 1. □

**Lemma 11** (Closeness Under Conditioning)**.** *Let $A, B$ be two distribution over $\Omega$. Let $E \subset \Omega$ be an event. Suppose that $\Pr[B \in E] \geq p \geq \mathsf{SD}(A; B)$, then*

$$\mathsf{SD}(A \mid E; B \mid E) \leq (\frac{1}{p} + \frac{1}{p(p - \epsilon)})\mathsf{SD}(A; B).$$

*Proof.* Let $\epsilon = \mathsf{SD}(A;\, B)$. Compute that

$$\begin{aligned}
\mathsf{SD}(A \mid E;\, B \mid E) &= \sum_{\substack{\omega \in E \\ P_{A|E}(\omega) > P_{B|E}(\omega)}} \frac{P_A(\omega)}{\Pr\left[A \in E\right]} - \frac{P_B(\omega)}{\Pr\left[B \in E\right]} \\
&\leq \sum_{\substack{\omega \in E \\ P_{A|E}(\omega) > P_{B|E}(\omega)}} \frac{P_A(\omega)}{\Pr\left[B \in E\right] - \epsilon} - \frac{P_B(\omega)}{\Pr\left[B \in E\right]} \\
&= \sum_{\substack{\omega \in E \\ P_{A|E}(\omega) > P_{B|E}(\omega)}} \frac{P_A(\omega)}{\Pr\left[B \in E\right]} \frac{\Pr\left[B \in E\right]}{\Pr\left[B \in E\right] - \epsilon} - \frac{P_B(\omega)}{\Pr\left[B \in E\right]} \\
&= \sum_{\substack{\omega \in E \\ P_{A|E}(\omega) > P_{B|E}(\omega)}} \frac{P_A(\omega)}{\Pr\left[B \in E\right]} (1 + \frac{\epsilon}{\Pr\left[B \in E\right] - \epsilon}) - \frac{P_B(\omega)}{\Pr\left[B \in E\right]} \\
&= \sum_{\substack{\omega \in E \\ P_{A|E}(\omega) > P_{B|E}(\omega)}} \frac{P_A(\omega)}{\Pr\left[B \in E\right]} (1 + \frac{\epsilon}{p - \epsilon}) - \frac{P_B(\omega)}{\Pr\left[B \in E\right]} \\
&\leq \frac{\epsilon}{\Pr\left[B \in E\right]} + \frac{\Pr\left[A \in E\right]}{\Pr\left[B \in E\right]} \cdot \frac{\epsilon}{p - \epsilon} \\
&\leq \frac{\epsilon}{p} + \frac{\epsilon}{p(p - \epsilon)}.
\end{aligned}$$

□

We now prove the following theorem, for the same construction $H_1[q]$.

**Theorem 9.** *Let $\mathcal{S}$ be a $q$-query UCE source with statistical unpredictability $k$. If $q \in O(k/\log(\lambda))$, the construction $H_1[q]$ is a secure $\mathsf{UCE}[\mathcal{S}]$ hash function.*

*Proof.* We use the same four hybrids. Notice that the only difference with considering non-split sources is the difference between $G_1$ and $G_2$. By Lemma 10, we need

$$(r + \frac{r^2}{1 - r\sqrt{r^q 2^{-k}}})\sqrt{r^q 2^{-k}} \tag{4.5}$$

be to negligible, where $r$ is not polynomial in $\lambda$. Suppose that $q \in O(k/\log(\lambda))$. Then,

$r^q 2^{-k}$ is negligible. Hence, $1 - r\sqrt{r^q 2^{-k}}$ is negligibly close to 1. Finally, we conclude that 4.5 is negligible. This, with the rest of the proof of Theorem 8, concludes the proof of Theorem 9. □

# Chapter 5

# Unified Lossy Primitive: Lossy Deterministic Encryption

## 5.1 Preliminaries

**Sources** Inspired by recent works [BHK13, BS16], we offer a parameterized security definitions based on distributions. A source, $\mathcal{D}$, is a (potentially) inefficient randomized algorithm, taking an input $q$, and generates $q$ output points together with leakage $z$. In this thesis, we will restrict to block-sources. $\mathcal{D}$ is a 1-block-source with min-entropy $k$ if for $(x_1, \ldots, x_q) \xleftarrow{\$} \mathcal{D}(q)$, $H_\infty(x_i \mid x_{<i}) \geq k$ for all $0 \leq i \leq q$.

## 5.2 Survey of Existing Results

### 5.2.1 Lossy Trapdoor Functions

A Lossy Trapdoor family of function is a family of trapdoor functions with two modes of key generation. Formally, it consists of three algorithms $(\mathsf{KG}, \mathrm{Eval}, \mathrm{Inv})$.

- In injective mode, $\mathsf{KG}(0)$ generates $(\mathsf{pk}, \mathsf{sk})$ with $\mathsf{pk}$ being the public evaluation key and $\mathsf{sk}$ being the secret trapdoor. We require that $\mathrm{Inv}(\mathsf{sk}, \mathrm{Eval}(\mathsf{pk}, \cdot))$ to be the identity function.

- In lossy mode, $\mathsf{KG}(1)$ generates $(\mathsf{lk}, \perp)$, with $\mathsf{lk}$ being the public evaluation key and no trapdoors. Additionally, $\mathrm{Eval}(\mathsf{lk}, \cdot)$ should have image size at most $2^{n-k}$.

- Furthermore, injective keys and lossy keys are computationally indistinguishable, i.e.
$$\{\mathsf{pk} : \mathsf{pk} \xleftarrow{\$} \mathsf{KG}(0)\} \approx_{\mathsf{c}} \{\mathsf{lk} : \mathsf{lk} \xleftarrow{\$} \mathsf{KG}(,1)\}.$$

### 5.2.2 Deterministic Encryption

A deterministic encryption scheme is a *public-key* encryption scheme that is deterministic. However, with the absence of encryption randomness, schemes cannot achieve the standard IND-CPA security anymore. In literature, handful of definitions are proposed, and they turn out to be equivalent. Below we present the PRIV security [BFOR08].

An deterministic encryption scheme $\mathsf{DE} = (\mathsf{KG}, \mathrm{Enc}, \mathrm{Dec})$ is said to be $\mathsf{PRIV}[\mathcal{D}]$-secure if the advantage, $\Pr[G_{\mathsf{PRIV}}[\mathsf{DE}, \mathcal{D}, \mathcal{A}] \Rightarrow 1]$, is negligible in $\lambda$ for any efficient $\mathcal{A}$.

Notice that here $\mathcal{D}$ generates a $q$ outputs, each with two component. The two component could be arbitrarily correlated. We will construct $\mathsf{DE}$ that is $\mathsf{PRIV}[\mathcal{D}]$-security for 1-block-source $\mathcal{D}$.

```
Game $G_{\mathsf{PRIV}}[\mathsf{DE}, \mathcal{D}, \mathcal{A}]()$:
    $b \xleftarrow{\$} \{0,1\}$
    $((\mathbf{x}_0, \mathbf{x}_1), z) \xleftarrow{\$} \mathcal{D}(q)$
    $\mathsf{pk} \xleftarrow{\$} \mathsf{DE.KG}$
    $c = \mathsf{DE}.\mathrm{Enc}(\mathsf{pk}, \mathbf{x}_b)$
    $b' \xleftarrow{\$} \mathcal{A}(c, z)$
    return $(b = b')$
```

Figure 5.1: Game defining the PRIV security of DE.

### 5.2.3 Lossy Encryption

Similar to LTDFs, Lossy Encryption also has two modes. However, the lossiness property implies something different about the encryption. Formally, a Lossy Encryption scheme is a collection of three algorithm $(\mathsf{KG}, \mathrm{Enc}, \mathrm{Dec})$ that has the following properties.

- In injective mode, $\mathsf{KG}(0)$ generates $(\mathsf{pk}, \mathsf{sk})$. We require that $\mathrm{Dec}(\mathsf{sk}, \mathrm{Enc}(\mathsf{pk}, m)) = m$ for all $m$ in the message space.

- In lossy mode, $\mathsf{KG}(1)$ generates $(\mathsf{lk}, \perp)$, and $\mathrm{Eval}(\mathsf{lk}, \cdot)$ defines a function whose image size is at most $2^{n-k}$.

- We require that the two modes are indistinguishable, i.e

$$\{\mathsf{pk} : \mathsf{pk} \xleftarrow{\$} \mathsf{KG}(0)\} \approx_c \{\mathsf{lk} : \mathsf{lk} \xleftarrow{\$} \mathsf{KG}(1)\}.$$

- We require that under the encryption randomness ($r$ sampled uniformly randomly), the ciphertext of *any* two plaintext, $x, x'$, are *statistically* close, i.e. it

```
**Proc** $\mathsf{LE.KG}()$:
    $(\mathsf{fpk}, \mathsf{fsk}) \xleftarrow{\$} \mathsf{LT.KG}()$
    $\mathsf{hk} \xleftarrow{\$} \{0,1\}^d$
    $\mathsf{pk} \leftarrow (\mathsf{fpk}, \mathsf{hk})$
    $\mathsf{sk} \leftarrow (\mathsf{fsk}, \mathsf{hk})$
    **return** $(\mathsf{pk}, \mathsf{sk})$

**Proc** $\mathsf{LE}.\mathrm{Enc}(\mathsf{pk}, m, r)$:
    $(\mathsf{fk}, \mathsf{hk}) \leftarrow \mathsf{pk}$
    $c_1 \leftarrow F(\mathsf{fk}, r)$
    $c_2 \leftarrow m \oplus h_{\mathsf{hk}}(r)$
    **return** $(c_1, c_2)$

**Proc** $\mathsf{LE}.\mathrm{Dec}(\mathsf{sk}, c)$:
    $(\mathsf{fsk}, \mathsf{hk}) \leftarrow \mathsf{sk}$
    $(c_1, c_2) \leftarrow c$
    $r \leftarrow \mathsf{LT}.\mathrm{Inv}(\mathsf{fsk}, c_1)$
    $m \leftarrow h(r) \oplus c_2$
    **return** $m$
```

Figure 5.2: Construction of Lossy Encryption: $\mathsf{LE}[\mathsf{LT}]$. $h$ is a universal hash function with seed length $d$ and output length $m$. [BHY09]

```
**Proc** $\mathsf{DE.KG}()$:
    $(\mathsf{fpk}, \mathsf{fsk}) \xleftarrow{\$} \mathsf{LT.KG}(0)$
    $(\mathsf{ppk}, \mathsf{ssk}) \xleftarrow{\$} \mathsf{PKE.KG}()$
    $\mathsf{hk} \xleftarrow{\$} \{0,1\}^d$
    $\mathsf{pk} \leftarrow (\mathsf{fpk}, \mathsf{ppk}, \mathsf{hk})$
    $\mathsf{sk} \leftarrow (\mathsf{fsk}, \mathsf{ssk}, \mathsf{hk})$
    **return** $(\mathsf{pk}, \mathsf{sk})$

**Proc** $\mathsf{DE}.\mathrm{Enc}(\mathsf{pk}, m)$:
    $(\mathsf{fpk}, \mathsf{ppk}, \mathsf{hk}) \leftarrow \mathsf{pk}$
    $y = \mathsf{LT}.\mathrm{Eval}(\mathsf{fpk}, m)$
    $r = h(\mathsf{hk}, m)$
    $c \leftarrow \mathsf{PKE}.\mathrm{Enc}(\mathsf{ppk}, y, r)$
    **return** $c$

**Proc** $\mathsf{DE}.\mathrm{Dec}(\mathsf{sk}, c)$:
    $(\mathsf{fsk}, \mathsf{ssk}, \mathsf{hk}) \leftarrow \mathsf{sk}$
    $c' \leftarrow \mathsf{PKE}.\mathrm{Dec}(\mathsf{ssk}, c)$
    $m \leftarrow \mathsf{LT}.\mathrm{Inv}(\mathsf{fsk}, c')$
    **return** $m$
```

Figure 5.3: The Encrypt-with-Hardcore Construction of Deterministic Encryption, $\mathsf{DE}[\mathsf{PKE}]$. PKE is a regular public-key encryption scheme. $h$ is a universal hash function with seed length $d$ and output length equal to the length of randomness of PKE

holds with negligible $\epsilon$ that

$$\mathrm{Enc}(\mathsf{lk}, x, r) \approx_{s,\epsilon} \mathrm{Enc}(\mathsf{lk}, x', r).$$

### 5.2.4 Constructions Using LTDF

LTDF has been used to construct Deterministic Encryption and Lossy encryption. One common method is to use a hardcore for LTDF to achieve security in Lossy and Deterministic Encryption. We highlight two existing constructions known in literature in Figure 5.2.4 and 5.2.4.

## 5.3 Unifying Definition: Lossy Deterministic Encryption

We show that a statistically-lossy version of Deterministic Encryption unifies the definition of deterministic encryption, lossy encryption, and hedged encryption.

**Definition 10.** *Let $\mathcal{D}$ be a source, $\epsilon : \mathbb{N} \to \mathbb{R}$ be a function. We say that an encryption scheme* $\mathsf{LDE} = (\mathsf{KG}, \mathrm{Enc}, \mathrm{Dec})$ *is a $(\mathcal{D}, \epsilon)$-Lossy Deterministic Encryption if the following holds.*

- *The following two distributions are computationally indistinguishable.*

$$\{\mathsf{pk} : \mathsf{pk} \overset{\$}{\leftarrow} \mathsf{KG}(0)\} \approx_{\mathsf{c}} \{\mathsf{lk} : \mathsf{lk} \overset{\$}{\leftarrow} \mathsf{KG}(1)\}.$$

- *Let $l$ some polynomial of security parameter $\lambda$. For any $m_0, \ldots, m_l \overset{\$}{\leftarrow} \mathcal{D}(l)$, it holds with overwhelming probability (over sampling* $\mathsf{lk} \overset{\$}{\leftarrow} \mathsf{KG}(1)$*) that*

$$\mathrm{Enc}_{\mathsf{lk}}(m_1), \ldots, \mathrm{Enc}_{\mathsf{lk}}(m_l) \approx_{s,\epsilon(l)} \mathrm{Enc}_{\mathsf{lk}}(u_1), \ldots, \mathrm{Enc}_{\mathsf{lk}}(u_l),$$

  *where $u_1, \ldots, u_l$ are independently uniformly random messages over the plaintext space.*

*In this thesis, we focus on 1-block-source $\mathcal{D}$. If $\mathcal{D}$ is a 1-block-source and an encryption scheme is $(\mathcal{D}, \epsilon)$-LDE, then we have $\epsilon(n) \leq n\epsilon(1)$. Hence, alternatively, we write $(k, \epsilon)$-LDE to mean $(\mathcal{D}, n\epsilon)$ for a 1-block-source $\mathcal{D}$ with component-wise entropy $k$.*

### 5.3.1 Unifying Construction from LTDFs

It is known that construction of LTDF from DDH and LWE are already deterministic encryptions schemes, since the LTDF is a strong extractor under the randomness of $\mathsf{lk}$. However, constructions of LTDF from QR, DCR are not directly extracting [Wee12].

```
Proc $\mathsf{ELT.KG}(b)$:
  $\mathsf{fk} \xleftarrow{\$} \mathsf{LT.KG}(b)$
  $\mathsf{hk} \xleftarrow{\$} \{0,1\}^d$
  return $(\mathsf{fk}, \mathsf{hk})$

Proc $\mathsf{ELT.pk}(m)$:
  $(\mathsf{fk}, \mathsf{hk}) \leftarrow \mathsf{pk}$
  return $F_{\mathsf{fk}}(H_{\mathsf{hk}}(m))$

Proc $\mathsf{ELT}^{-1}_{\mathsf{pk}}(c)$:
  $m \leftarrow H^{-1}_{\mathsf{hk}}(F^{-1}_{\mathsf{fk}}(c))$
  return $m$
```

Figure 5.4: Construction of Extracting LTDF

```
Proc $\mathsf{LDE.KG}(b)$:
  $(\mathsf{pk}, \mathsf{sk}) \xleftarrow{\$} \mathsf{ELT.KG}(b)$
  return $(\mathsf{pk}, \mathsf{sk})$

Proc $\mathsf{LDE.Enc}_{\mathsf{pk}}(m, r)$:
  return $F_{\mathsf{pk}}(r)$

Proc $\mathsf{LDE.Dec}_{\mathsf{pk}}(c)$:
  $m \leftarrow [(F^{-1}_{\mathsf{pk}}(c)]^{l}_{1}$
  return $m$
```

Figure 5.5: Construction of Lossy Deterministic Encryption from any Extracting LTDF.

**Definition 11** (Extracting LTDF). *A LTDF,* $\mathsf{LT}$*, is* $(k', \epsilon)$*-extracting if for* $(\mathsf{pk}, \perp) \xleftarrow{\$} \mathsf{LT.KG}(1)$*, and for any distribution of messages, X over the input space* $\mathcal{M}$*, such that* $H_\infty(X) \geq k'$*, we have that*

$$\mathsf{LT.Eval}(\mathsf{pk}, X) \approx_{s,\epsilon} \mathsf{LT.Eval}(\mathsf{pk}, U_{\mathcal{M}}).$$

**Extracting LTDF from any LTDF** We note that the Generalized Leftover Hash Lemma (Theorem 1) gives rise to an extracting LTDF. Let $H$ be a pairwise independent permutation over $\{0,1\}^n$ with seed length $d$, and let $F$ be a $(n,k)$-LTDF, whose lossy image size is at most $2^{n-k}$. The construction in Figure 5.3.1 is an $(k', \sqrt{2^{n-k-k'}})$-extracting $(n,k)$-LTDF.

**Lossy Deterministic Encryption from Extracting LTDFs** Assuming an extracting LTDF, we use the pad-and-deterministic (PtD) technique [BBN$^+$09]. Let $F$ be an $(k', \epsilon)$-extracting $(n,k)$-LTDF. Then the below construction is a secure $(k', \epsilon)$-lossy deterministic encryption scheme.

| | Message length | Randomness length |
|---|---|---|
| EwH | $n-k-2\log(1/\epsilon)$ | $n$ |
| Construction from LDE | $n-k-2\log(1/\epsilon)$ | $k+2\log(1/\epsilon)$ |

Table 5.1: Table comparing EwH and LE from LDE, using the same $(n,k)$-LTDF.

**Comparision of Parameters** Table 5.3.1 compares Encryption-with-Hardcore (EwH) and the construction from LDE, using the same $(n,k)$-lossy trapdoor family. Construction of LE from LDE uses less randomness to encrypt the same message space with the same closeness parameter $\epsilon$.

## 5.4 Equivalence of Lossy Deterministic Encryption and Lossy Trapdoor Functions

In [HO13], it was observed that Lossy Encryption implies LTDF if the length of the randomness is shorter than the message length. In particular, it was shown that, with overwhelming probability, the function $\mathsf{LE.Enc}(\mathsf{lk},\cdot,h_s(\cdot))$ is lossy. This result leaves open the question of whether there is an equivalence between all aforementioned lossy primitives.

In the non-lossy case, we know that there does not exists black-box construction of Trapdoor functions from Public-Key Encryption schemes [GMR01]. Hence, finding such an equivalence in the statistical case is interesting since we know such equivalence does not hold in the computational case.

Unlike Lossy Encryption, Lossy Deterministic Encryption has a distributional assumption that is stronger. The intuition here is that the Lossy Deterministic Encryption assumption is strong enough to directly imply LTDF. We present a simple lemma to support our claim.

**Lemma 12.** *Let $f : \mathcal{M} \to \mathcal{N}$ be any function. Suppose that $|\mathcal{M}| = 2^n$. Suppose that for any distribution, $X$, on $\mathcal{M}$, such that $H_\infty(X) \geq k$, we have $f(X) \approx_{s,\epsilon} f(U_\mathcal{M})$. Then, $|f(U)| \leq (2^{k-n} + \epsilon) \cdot |\mathcal{M}|$, i.e. $f$ is $(n, \log(2^{k-n} + \epsilon)n)$-lossy.*

*Proof.* Pick a flat distribution, $X$, with minimum entropy $k$ (uniform over $2^k$ elements). Notice that $|f(X)| \leq 2^k$. By hypothesis, we have that $f(X) \approx_{s,\epsilon} f(U_\mathcal{M})$. Define $B = \{\omega \in \mathcal{N} \mid \Pr[f(X) = \omega] < \Pr[f(U_\mathcal{M}) = \omega]\}$. Notice that $|f(U_\mathcal{M})| \leq |f(X)| + |B|$ and that $\sum_{\omega \in B} \Pr[f(U_\mathcal{M}) = \omega] - \Pr[f(X) = \omega] \leq \epsilon$. We compute,

$$\begin{aligned}
\epsilon &\geq \sum_{\omega \in B} \Pr[f(U_\mathcal{M}) = \omega] - \Pr[f(X) = \omega] \\
&\geq \sum_{\omega \in B} \Pr[f(U_\mathcal{M}) = \omega] \\
&\geq \Pr\left[U_\mathcal{M} \in f^{-1}(B)\right].
\end{aligned}$$

Hence, $|B| \leq |f^{-1}(B)| \leq |\mathcal{M}|\epsilon$, since $U_\mathcal{M}$ is uniform. Hence, $|f(U_\mathcal{M})| \leq |f(K)| + B \leq 2^k + \epsilon 2^n = (2^{k-n} + \epsilon)|\mathcal{M}|$. □

Observe that LDE and LTDF differs only in how "lossiness" is defined. Using the above lemma to relate the two notions of lossiness, the following theorem follows naturally.

**Theorem 10.** *Any Lossy Deterministic Encryption Scheme with message length $n$, that is secure for messages with min-entropy at least $k$ and closeness $\epsilon$, is also a $(n, \log(2^{k-n} + \epsilon)n)$-Lossy Trapdoor Function.*

# Bibliography

[BBN$^+$09] Mihir Bellare, Zvika Brakerski, Moni Naor, Thomas Ristenpart, Gil Segev, Hovav Shacham, and Scott Yilek. Hedged public-key encryption: How to protect against bad randomness. In *Advances in Cryptology–ASIACRYPT 2009*, pages 232–249. Springer, 2009.

[BD00] Rajendra Bhatia and Chandler Davis. A better bound on the variance. *The American Mathematical Monthly*, 107(4):353–357, 2000.

[BFOR08] Mihir Bellare, Marc Fischlin, Adam ONeill, and Thomas Ristenpart. Deterministic encryption: Definitional equivalences and constructions without random oracles. In *Advances in Cryptology–CRYPTO 2008*, pages 360–378. Springer, 2008.

[BHK13] Mihir Bellare, Viet Tung Hoang, and Sriram Keelveedhi. Instantiating random oracles via uces. In *Advances in Cryptology–CRYPTO 2013*, pages 398–415. Springer, 2013.

[BHY09] Mihir Bellare, Dennis Hofheinz, and Scott Yilek. Possibility and impossibility results for encryption and commitment secure under selective opening. In *Advances in Cryptology-EUROCRYPT 2009*, pages 1–35. Springer, 2009.

[BR08] Mihir Bellare and Phillip Rogaway. Code-based game-playing proofs and the security of triple encryption. 2008.

[BS16] Mihir Bellare and Igors Stepanovs. Point-function obfuscation: a framework and generic constructions. In *Theory of Cryptography Conference*, pages 565–594. Springer, 2016.

[CPS14] Benoit Cogliati, Jacques Patarin, and Yannick Seurin. Security amplification for the composition of block ciphers: simpler proofs and new results. In *Selected Areas in Cryptography–SAC 2014*, pages 129–146. Springer, 2014.

[CV08] Kai-Min Chung and Salil Vadhan. Tight bounds for hashing block sources. In *Approximation, Randomization and Combinatorial Optimization. Algorithms and Techniques*, pages 357–370. Springer, 2008.

[DORS08] Yevgeniy Dodis, Rafail Ostrovsky, Leonid Reyzin, and Adam Smith. Fuzzy extractors: How to generate strong keys from biometrics and other noisy data. *SIAM journal on computing*, 38(1):97–139, 2008.

[DS05] Yevgeniy Dodis and Adam Smith. Correcting errors without leaking partial information. In *Proceedings of the thirty-seventh annual ACM symposium on Theory of computing*, pages 654–663. ACM, 2005.

[FB14] Serge Fehr and Stefan Berens. On the conditional rényi entropy. *Information Theory, IEEE Transactions on*, 60(11):6801–6810, 2014.

[FOR15] Benjamin Fuller, Adam ONeill, and Leonid Reyzin. A unified approach to deterministic encryption: New constructions and a connection to computational entropy. *Journal of Cryptology*, 28(3):671–717, 2015.

[GMR01] Yael Gertner, Tal Malkin, and Omer Reingold. On the impossibility of basing trapdoor functions on trapdoor predicates. In *focs*, volume 1, pages 126–135. Citeseer, 2001.

[GS02] Alison L Gibbs and Francis Edward Su. On choosing and bounding probability metrics. *International statistical review*, 70(3):419–435, 2002.

[HILL99] Johan Håstad, Russell Impagliazzo, Leonid A Levin, and Michael Luby. A pseudorandom generator from any one-way function. *SIAM Journal on Computing*, 28(4):1364–1396, 1999.

[HMR12] Viet Tung Hoang, Ben Morris, and Phillip Rogaway. An enciphering scheme based on a card shuffle. In *Advances in Cryptology–CRYPTO 2012*, pages 1–13. Springer, 2012.

[HO13] Brett Hemenway and Rafail Ostrovsky. Building lossy trapdoor functions from lossy encryption. In *Advances in Cryptology-ASIACRYPT 2013*, pages 241–260. Springer, 2013.

[LR88] Michael Luby and Charles Rackoff. How to construct pseudorandom permutations from pseudorandom functions. *SIAM Journal on Computing*, 17(2):373–386, 1988.

[Mau02] Ueli Maurer. Indistinguishability of random systems. In *Advances in CryptologyEUROCRYPT 2002*, pages 110–132. Springer, 2002.

[MP04] Ueli Maurer and Krzysztof Pietrzak. Composition of random systems: When two weak make one strong. In *Theory of Cryptography Conference*, pages 410–427. Springer, 2004.

[MPR07] Ueli Maurer, Krzysztof Pietrzak, and Renato Renner. Indistinguishability amplification. In *Advances in Cryptology-CRYPTO 2007*, pages 130–149. Springer, 2007.

[PVW08] Chris Peikert, Vinod Vaikuntanathan, and Brent Waters. A framework for efficient and composable oblivious transfer. In *Advances in Cryptology–CRYPTO 2008*, pages 554–571. Springer, 2008.

[PW11] Chris Peikert and Brent Waters. Lossy trapdoor functions and their applications. *SIAM Journal on Computing*, 40(6):1803–1844, 2011.

[$R^+61$] Alfréd Rényi et al. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California, 1961.

[Tes14] Stefano Tessaro. Optimally secure block ciphers from ideal primitives. In *International Conference on the Theory and Application of Cryptology and Information Security*, pages 437–462. Springer, 2014.

[$V^+$] Salil P Vadhan et al. *Pseudorandomness*, volume 56.

[Wee12] Hoeteck Wee. Dual projective hashing and its applicationslossy trapdoor functions and more. In *Advances in Cryptology–EUROCRYPT 2012*, pages 246–262. Springer, 2012.

[Zha16] Mark Zhandry. The magic of elfs. In *Proceedings of CRYPTO 2016*, 2016.