Diss. ETH No. 12187

Entropy Measures and Unconditional Security in Cryptography

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY
ZÜRICH

for the degree of
Doctor of Technical Sciences

presented by

CHRISTIAN CACHIN
Dipl. Informatik-Ing. ETH

born February 15, 1968
citizen of Cerniaz VD and Zürich

accepted on the recommendation of

Prof. Dr. U. Maurer, referee
Prof. Dr. J.L. Massey, co-referee

1997

Acknowledgments

First of all, I thank Prof. Ueli Maurer for his support, his advice, and his encouragement during our collaboration that led to this thesis. I am grateful to Prof. Jim Massey for his interest, for his careful reading of the text, and for his acceptance to be my co-referee.

It is a pleasure to thank my officemate Jan Camenisch for many discussions about cryptography, computer problems, and lots of other things. Many thanks go also to the other members of the cryptography and information security group, Daniel Bleichenbacher, Markus Stadler, Stefan Wolf, Martin Hirt, Ronald Cramer, and Reto Kohlas. To all the department members on the "theory floor" sharing some of my time there, I am grateful for creating the open, stimulating, and relaxed atmosphere.

I am indebted to my parents, for supporting me and make my education possible at all. And finally, I thank Irene—for everything!

Abstract

One of the most important properties of a cryptographic system is a proof of its security. In the present work, information-theoretic methods are used for proving the security of unconditionally secure cryptosystems. The security of such systems does not depend on unproven intractability assumptions.

A survey of entropy measures and their applications in cryptography is presented. A new information measure, smooth entropy, is introduced to quantify the number of almost uniform random bits that can be extracted from a source by probabilistic algorithms. Smooth entropy unifies previous work on privacy amplification in cryptography and on entropy smoothing in theoretical computer science. It enables a systematic investigation of the spoiling knowledge proof technique to obtain lower bounds on smooth entropy.

The Rényi entropy of order at least 2 of a random variable is a lower bound for its smooth entropy, whereas an assumption about Rényi entropy of order 1, which is equivalent to the Shannon entropy, is too weak to guarantee any non-trivial amount of smooth entropy. The gap between Rényi entropy of order 1 and 2 is closed by proving that Rényi entropy of order α between 1 and 2 is a lower bound for smooth entropy, up to a small parameter depending on α, the alphabet size, and the failure probability.

The operation of many unconditionally secure cryptosystems can be divided into the three phases advantage distillation, information reconciliation, and privacy amplification. The relation between privacy amplification and information reconciliation is investigated, in particular, the effect of side information, obtained by an adversary through an initial reconciliation step, on the size of the secret key that can be distilled safely by subsequent privacy amplification. It is shown that each bit of side information reduces the size of the key that can be generated by at

most one bit, except with negligible probability.

A private-key cryptosystem and a protocol for key agreement by public discussion are proposed that are unconditionally secure based on the sole assumption that an adversary's memory capacity is limited. The systems make use of a random bit string of length slightly larger than the adversary's memory capacity that can be received by all parties.

Zusammenfassung

Eine der wichtigsten Eigenschaften jedes kryptographischen Systems ist ein Beweis für seine Sicherheit. In dieser Arbeit werden informationstheoretische Methoden eingesetzt, um die Sicherheit von Kryptosystemen zu garantieren, die nicht auf komplexitätstheoretischen Annahmen beruhen.

Eine Übersicht über Entropiemasse und ihre Anwendungen in der Kryptographie wird präsentiert. Es wird ein neues Informationsmass eingeführt, die Smooth-Entropie, das die Anzahl beinahe uniform verteilter Bits beziffert, die aus einer Informationsquelle mit Hilfe von probabilistischen Algorithmen extrahiert werden können. Das Konzept der Smooth-Entropie vereinigt bestehende Arbeiten zu Privacy-Amplification in der Kryptographie und Entropy-Smoothing in der Theoretischen Informatik. Es erlaubt eine systematische Untersuchung der Spoiling-Knowledge-Beweistechnik zur Konstruktion von unteren Schranken für die Smooth-Entropie.

Die Rényi-Entropie der Ordnung mindestens 2 einer Zufallsvariable ist eine untere Schranke für ihre Smooth-Entropie. Andererseits reicht eine Annahme über die Rényi-Entropie der Ordnung 1, die äquivalent zur Shannon-Entropie ist, nicht aus, um eine substantielle Grösse der Smooth-Entropie zu garantieren. Die Lücke zwischen Rényi-Entropien der Ordnung 1 und 2 wird geschlossen, in dem gezeigt wird, dass Rényi-Entropie der Ordnung α zwischen 1 und 2 eine untere Schranke für die Smooth-Entropie ist, bis auf einen Fehlerterm, der von α, der Grösse des Wertebereichs und der Fehlerwahrscheinlichkeit abhängt.

Viele kryptographische Systeme mit informationstheoretisch beweisbarer Sicherheit operieren in den drei Phasen Advantage-Distillation, Information-Reconciliation und Privacy-Amplification. Die Beziehung zwischen Information-Reconciliation und Privacy-Amplification wird untersucht, insbesondere der Effekt von Zusatzinformation, die der Gegner

während Information-Reconciliation erhält, auf die Grösse des geheimen Schlüssels, der mit nachfolgender Privacy-Amplification erzeugt werden kann. Es wird gezeigt, dass jedes Bit Zusatzinformation die Schlüsselgrösse um höchstens ein Bit reduziert, ausser mit vernachlässigbar kleiner Wahrscheinlichkeit.

Ferner werden ein Verschlüsselungssystem beschrieben und ein Protokoll zur Vereinbarung eines geheimen Schlüssels durch öffentliche Diskussion vorgeschlagen, die beweisbar sicher sind unter der Annahme, dass ein Gegner beschränkte Speicherressourcen hat (jedoch ohne Einschränkung seiner Rechenressourcen). Diese Systeme basieren auf einem zufälligen Bitstring, nur wenig grösser als der Speicherplatz des Gegners, der von einem Sender an alle Teilnehmer übermittelt wird.

Contents

Chapter 1

Introduction

Secure information transmission and storage is a paramount requirement in the emerging information economy. Techniques for realizing secure information handling are provided by cryptography, which can be defined as the *study of communication in an adversarial environment* [Riv90].

The classical goals of cryptography are secrecy and authentication: to protect information against unauthorized disclosure and against unauthorized modification. In the recent years, research in cryptography has addressed a broad range of more advanced questions, ranging from the authorization of user access to computer systems over secure electronic voting schemes to the realization of untraceable electronic cash. Surveys of contemporary cryptography are available in a number of reference works [MvOV97, Sti95, Gol95, Sim91, Riv90].

The roots of cryptography can be traced back to the invention of writing. The Roman emperor Caesar devised a simple encryption scheme that still bears his name today [Kah67]. Cryptography as a science originates with the seminal work of Shannon that laid the foundations for information theory [Sha48] and treats cryptography as one of its first applications [Sha49]. All cryptographic systems or *ciphers* designed until 1976 are based on a secret key known to the communicating partners that is used to encrypt and to decrypt information. Such methods are called *symmetric*, *secret-key*, or *private-key* systems, in contrast to the methods of *public-key* cryptography, which first appeared in the work of Diffie and Hellman [DH76]. Their revolutionary idea was to use separate keys for encryption and decryption: a public key that can be used by anybody and a secret key known only to its owner. This asymmetry is fundamental for realizing the two basic primitives of modern cryptography, public-key cryptosystems and digital signature schemes.

A *public-key cryptosystem* allows two partners to generate a secret key and to exchange information secretly only by communicating over public channels that can be eavesdropped by an adversary. Every participant in a public-key setup has a public key that can be used by any other party to encrypt a message. The corresponding secret key is known only to the recipient and is needed to decrypt a message encrypted with his public key. Thus, a public-key system provides secret transmission of information.

Its dual, a *digital signature scheme*, provides authenticity and uses also a public key and a secret key for every user. A user can digitally sign a message with his secret key to protect the integrity of the message. Any participant can verify the signature with the public key of the signer. The RSA cryptosystem proposed by Rivest, Shamir, and Adleman in 1978 [RSA78] is one of the most widely known public-key systems today and can be employed both as a public-key cryptosystem and as a digital signature scheme.

The security of most currently used cryptosystems is based on the difficulty of an underlying *computational* problem, such as factoring large numbers or computing discrete logarithms for many public-key systems. Security proofs for these systems show that the ability of an adversary to defeat the cryptosystem with significant probability contradicts the assumed difficulty of the problem. This notion of security is always *conditional* to an unproven assumption. Although the hardness of these problems is unquestioned at the moment, it can be dangerous to base the security of the global information economy on a very small number of mathematical problems.

In contrast, the stronger notion of information-theoretic or *unconditional* security assumes no limits on an adversary's computational power and does not base the security on intractability assumptions. Shannon's information-theoretic definition of perfect secrecy led immediately to his famous pessimistic theorem [Sha49], which states, roughly, that the shared secret key in any perfectly secure cryptosystem must be at least as long as the plaintext to be encrypted. However, recent developments show that practical and provably secure cryptosystems become possible when some small modifications of Shannon's model are made.

This thesis contributes to the research on unconditionally secure cryptographic systems in a number of ways. In Chapter 2, the basic concepts of information theory are introduced, which are the main tools for reasoning about unconditional security. We describe privacy amplification, which is a building block of many unconditionally secure cryptosystems, and we introduce Rényi entropy, which plays an important role in connection with entropy smoothing and privacy amplification.

Chapter 3 contains a survey of several *information measures* and their cryptographic applications. The comparison demonstrates that other information measures than the standard Shannon entropy answer natural quantitative questions in various of cryptographic scenarios. A collection of bounds that link the presented information measures to each other is presented, and a discussion of the role of Shannon entropy in cryptographic applications completes the chapter.

In Chapter 4, we introduce the notion of *smooth entropy* that allows a unifying formulation of privacy amplification and entropy smoothing. Smooth entropy is a measure for the number of almost uniform random bits that can be extracted from a random source by probabilistic algorithms. We examine the *spoiling knowledge* proof technique to obtain lower bounds on smooth entropy and give a characterization of those kinds of spoiling knowledge that lead to better lower bounds. In addition, we establish a new connection between smooth entropy and *Rényi entropy* by proving a lower bound on smooth entropy in terms of Rényi entropy of order α for any $1 < \alpha < 2$. Previously, it was only known that the Rényi entropy of order at least 2 of a random variable is a lower bound for its smooth entropy and that no such statement can be made for its Shannon entropy (which is Rényi entropy of order 1).

In Chapter 5, we focus on the realization of *unconditionally secure cryptosystems*. These systems can usually be divided into the three phases advantage distillation, information reconciliation, and privacy amplification. In the first part of Chapter 5, we investigate the effect of *side information* that an adversary obtains during information reconciliation on privacy amplification. We show that, with high probability, each bit of side information reduces the size of the key that can be safely distilled by at most one bit.

In the second part of Chapter 5, we propose a private-key cryptosystem and a protocol for key agreement by public discussion that are unconditionally secure based on the sole assumption that an adversary's *memory capacity* is limited. The system makes use of a random bit string of length slightly larger than the adversary's memory capacity that can be received by all parties.

Chapter 2

Preliminaries

This chapter reviews the probability theory and information theory needed for the later chapters. The purpose is to introduce the notation, but not to give self-contained treatments of probability theory or information theory. References for these topics are the books by Feller [Fel68], Billingsley [Bil95], Blahut [Bla87], and Cover and Thomas [CT91]. Some material in this chapter is also based on [Mau95, Lub96, Rén61]. In Section 2.6, we introduce the notion of privacy amplification in cryptography, which is fundamental in the remaining chapters.

All logarithms are to the base two. Concatenation of symbols is denoted by $\circ$, concatenation of random variables by juxtaposition. The cardinality of a set $\mathcal{S}$ is denoted by $|\mathcal{S}|$. Vectors $\mathbf{v}$ are typeset in boldface.

2.1 Discrete Probability Theory

A *discrete probability space* consists of a finite or countably infinite set Ω, the *sample space*, together with a *probability measure* P. The elements of the sample space Ω are called *elementary events* and the subsets of Ω are called *events*. Each elementary event can be viewed as a possible outcome of an experiment. The *probability distribution* or *probability measure* P is a mapping from the set of events to the real numbers such that the following is satisfied:

1. $\mathrm{P}[\mathcal{A}] \geq 0$ for any event $\mathcal{A} \subseteq \Omega$.

2. $\mathrm{P}[\Omega] = 1$.

3. For every two events $\mathcal{A}, \mathcal{B} \subset \Omega$ with $\mathcal{A} \cap \mathcal{B} = \emptyset$,

$$\mathrm{P}[\mathcal{A} \cup \mathcal{B}] \;=\; \mathrm{P}[\mathcal{A}] + \mathrm{P}[\mathcal{B}].$$

$\mathrm{P}[\mathcal{A}]$ is called the *probability* of the event $\mathcal{A}$. For any two events $\mathcal{A}$ and $\mathcal{B}$, the probability of the union event $\mathcal{A} \cup \mathcal{B}$ is thus

$$\mathrm{P}[\mathcal{A} \cup \mathcal{B}] \;=\; \mathrm{P}[\mathcal{A}] + \mathrm{P}[\mathcal{B}] - \mathrm{P}[\mathcal{A} \cap \mathcal{B}].$$

The *union bound* is a simple consequence of this:

$$\mathrm{P}[\mathcal{A} \cup \mathcal{B}] \;\leq\; \mathrm{P}[\mathcal{A}] + \mathrm{P}[\mathcal{B}].$$

Two events are called *independent* if

$$\mathrm{P}[\mathcal{A} \cap \mathcal{B}] \;=\; \mathrm{P}[\mathcal{A}] \cdot \mathrm{P}[\mathcal{B}].$$

The *conditional probability* $\mathrm{P}[\mathcal{A}|\mathcal{B}]$ of an event $\mathcal{A}$ given that another event $\mathcal{B}$ occurs is defined as

$$\mathrm{P}[\mathcal{A}|\mathcal{B}] \;=\; \frac{\mathrm{P}[\mathcal{A} \cap \mathcal{B}]}{\mathrm{P}[\mathcal{B}]}$$

whenever $\mathrm{P}[\mathcal{B}]$ is positive.

A *discrete random variable* X is a mapping from the sample space Ω to an alphabet $\mathcal{X}$. X assigns a value $x \in \mathcal{X}$ to each elementary event in Ω and the probability distribution of X is the function

$$P_X \;:\; \mathcal{X} \to \mathbb{R} \;:\; x \mapsto P_X(x) \;=\; \mathrm{P}[X = x] \;=\; \sum_{\omega \in \Omega : X(\omega) = x} \mathrm{P}[\omega]$$

Random variables are always denoted by capital letters and sometimes written as $X \in \mathcal{X}$. If not stated otherwise, the alphabet of a random variable is denoted by the corresponding script letter. A sequence $X_1, \ldots, X_n$ of random variables with the same alphabet is denoted by X^n.

Multiple random variables can be defined over the same sample space. The *joint distribution* of the random variables X and Y is defined as the distribution of the single, vector-valued random variable XY with alphabet $\mathcal{X} \times \mathcal{Y}$. The distributions of X and Y are determined uniquely by P_{XY}.

The *conditional probability distribution* of a random variable X given an event $\mathcal{A}$ with positive probability is defined as

$$P_{X|\mathcal{A}}(x) \;=\; \mathrm{P}[X = x|\mathcal{A}].$$

If the conditioning event involves another random variable Y defined on the same sample space, the conditional probability distribution of X given that Y takes on a value y is

$$P_{X|Y=y}(x) \;=\; \frac{P_{XY}(x,y)}{P_Y(y)}$$

whenever $P_Y(y)$ is positive. Two random variables X and Y are called *independent* if for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$

$$P_{XY}(x,y) \;=\; P_X(x) \cdot P_Y(y).$$

Random variables taking on real values are of particular importance. The *expected value* of a discrete random variable X over $\mathcal{X} \subset \mathbb{R}$ is

$$\mathrm{E}[X] \;=\; \sum_{x \in \mathcal{X}} x P_X(x)$$

and its *variance* is

$$\mathrm{Var}[X] \;=\; \mathrm{E}[(X - \mathrm{E}[X])^2] = \mathrm{E}[X^2] - \mathrm{E}[X]^2.$$

The notation $\mathrm{E}_X[\cdot]$ is sometimes used to state explicitly that the random experiment over which the expectation is taken is the random experiment underlying the random variable X.

The *collision probability* $P_2(X)$ of a random variable X is defined as

$$P_2(X) \;=\; \sum_{x \in \mathcal{X}} P_X(x)^2$$

and denotes the probability that X takes on the same value twice in two independent experiments. For any X, the collision probability satisfies

$$\frac{1}{|\mathcal{X}|} \;\le\; P_2(X) \;\le\; 1$$

with equality on the left if and only if P_X is the uniform distribution over $\mathcal{X}$ and equality on the right if and only if $P_X(x) = 1$ for some $x \in \mathcal{X}$.

Distances between probability distributions are quantified using two related measures: The L_1 *distance* between two probability distributions P_X and P_Y with the same alphabet $\mathcal{X}$ is defined as

$$\|P_X - P_Y\|_1 \;=\; \sum_{x\in\mathcal{X}} |P_X(x) - P_Y(x)|$$

and the *variational distance* between P_X and P_Y is

$$\|P_X - P_Y\|_v \;=\; \max_{\mathcal{X}_0\subseteq\mathcal{X}} \left| \sum_{x\in\mathcal{X}_0} P_X(x) - P_Y(x) \right| \;=\; \frac{1}{2}\|P_X - P_Y\|_1.$$

The variational distance between the distribution of a random variable X and the uniform distribution P_U over $\mathcal{X}$ can be interpreted in the following way. Assume that $\|P_X - P_U\|_v \le \epsilon$. Then there is a refinement of the probability space underlying X in which an event $\mathcal{E}$ exists that has probability at least $1-\epsilon$ such that $\|P_{X|\mathcal{E}} - P_U\|_v = 0$, i.e. X behaves like a uniformly distributed random variable with probability at least $1-\epsilon$.

2.2 Some Inequalities

2.2.1 The Jensen Inequality

A function f is called[1] *convex*-$\cup$ or *convex* on the interval $[a,b]$ if and only if, for all $x_1, x_2 \in [a,b]$ and $0 \le \lambda \le 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \;\le\; \lambda f(x_1) + (1-\lambda) f(x_2). \tag{2.1}$$

f is called *strictly convex*-$\cup$ if inequality (2.1) is strict, whenever $0 < \lambda < 1$. A function g is called [strictly] *convex*-$\cap$ or *concave* if and only if $-g$ is [strictly] convex.

The *Jensen inequality* states that for a convex-$\cup$ function f and a random variable X

$$\mathrm{E}[f(X)] \;\ge\; f(\mathrm{E}[X]). \tag{2.2}$$

Moreover, if f is strictly convex-$\cup$, then equality in (2.2) implies that $X = \mathrm{E}[X]$ with probability 1.

[1] Following the descriptive terminology by Massey [Mas93].

2.2.2 The Moment, Markov, Chebychef, and other Inequalities

Many useful inequalities are based on the following observation: Let X be a real-valued random variable, let f be a function from $\mathbb{R}$ to $\mathbb{R}$, let I be some interval of the real numbers, and let c be a constant such that for all $x \notin I, f(x) \geq c$. Let $\chi_I(x)$ be the characteristic function of I, i.e., $\chi_I(x) = 1$ if $x \in I$ and $\chi_I(x) = 0$ if $x \notin I$. Then

$$c \cdot \mathrm{P}[X \notin I] + \mathrm{E}[f(X) \cdot \chi_I(X)] \ \leq \ \mathrm{E}[f(X)]. \tag{2.3}$$

An application of this is the *k-th moment inequality* for any integer $k > 0$ and any $t \in \mathbb{R}^+$,

$$P[|X| \geq t] \ \leq \ \frac{\mathrm{E}[|X|^k]}{t^k} \tag{2.4}$$

which follows from (2.3) with $f(x) = |x|^k$ and $I = [-t, +t]$ by noting that for all $x \notin I$, $f(x) \geq t^k$ and $\mathrm{E}[f(X) \cdot \chi_I(X)] \geq 0$.

The special case $k = 1$ of (2.4) is known as the *Markov inequality*: For any positive-valued random variable X and any $t \in \mathbb{R}^+$,

$$P[X \geq t] \ \leq \ \frac{\mathrm{E}[X]}{t}. \tag{2.5}$$

The special case $k = 2$ of (2.4) is known as the *Chebychef inequality*: For any real-valued random variable X and any $t \in \mathbb{R}^+$,

$$P[|X - \mathrm{E}[X]| \geq t] \ \leq \ \frac{\mathrm{Var}[X]}{t^2}. \tag{2.6}$$

The following inequality yields tight bounds in many cases: For any real-valued random variable X, any $t \in \mathbb{R}^+$, and any $r \in \mathbb{R}$,

$$\mathrm{P}[X \geq r] \ \leq \ \mathrm{E}[e^{(X-r)t}]. \tag{2.7}$$

It follows from (2.3) with $f(x) = e^{(X-r)t}$ and $I = (-\infty, r]$ by noting that for all $x \notin I, f(x) \geq 1$ and $\mathrm{E}[f(X) \cdot \chi_I(X)] \geq 0$.

2.2.3 Chernoff-Hoeffding Bounds

Sharp bounds attributed to Chernoff and Hoeffding exist for the sum of *independent and identically distributed (i.i.d.)* random variables. Because independent sampling is used, the error probability in the approximation decreases exponentially with the number of sample points.

Let $X_1, \ldots, X_n$ be a sequence of i.i.d. real-valued random variables with distribution P_X, expected value $\mathrm{E}[X]$, and range in the interval $[a, b]$. Then for any $t \in \mathbb{R}^+$,

$$\mathrm{P}\Big[\Big|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathrm{E}[X]\Big| \geq t\Big] \;\leq\; 2e^{-\frac{2t^2}{(b-a)^2}\cdot n}. \tag{2.8}$$

Similar bounds hold also for sums of random variables with limited independence [SSS95]: Let $X_1, \ldots, X_n$ be k-wise independent random variables with distribution P_X, alphabet in the interval $[0, 1]$ and expectation $\mathrm{E}[X] = \mu$. Then

$$\mathrm{P}\Big[\Big|\sum_{i=1}^{n} X_i - n\mu\Big| \geq \sqrt{e^{1/3}kn\mu}\Big] \;\leq\; e^{-\lfloor k/2 \rfloor}. \tag{2.9}$$

2.3 Entropy and Information Theory

The *(Shannon) entropy* [Sha48] of a random variable X with probability distribution P_X and alphabet $\mathcal{X}$ is defined as

$$H(X) \;=\; -\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)$$

with the convention that $0 \log 0 = 0$, which is justified by the fact that $\lim_{p \to 0} p \log \frac{1}{p} = 0$. The *conditional entropy* of X conditioned on a random variable Y is

$$H(X|Y) \;=\; \sum_{y \in \mathcal{Y}} P_Y(y) H(X|Y = y)$$

where $H(X|Y = y)$ denotes the entropy computed from the conditional probability distribution $P_{X|Y=y}$. The definition of entropy can also be stated as an expected value,

$$H(X) \;=\; \mathrm{E}_X[-\log P_X(X)].$$

The entropy $H(X)$ of a random variable X is a measure of its *average uncertainty.* It is the minimum number of bits required on the average to describe the value x of the random variable X. Similarly, $H(X|Y)$ is the average number of bits required to describe X when Y is already known. In communication theory, where information theory originates, the entropy of a source gives the ultimately achievable error-free compression in terms of the average codeword length per source symbol, which is the first fundamental result of information theory.

Proposition 2.1. *Some important immediate properties of H are:*

1. *Entropy is positive: $0 \leq H(X)$ with equality if and only if $P_X(x) = 1$ for some $x \in \mathcal{X}$.*
2. *Entropy is bounded: $H(X) \leq \log|\mathcal{X}|$ with equality if and only if X is uniformly distributed over $\mathcal{X}$.*
3. *Conditioning on side information reduces entropy: $H(X|Y) \leq H(X)$ with equality if and only if X and Y are independent.*
4. *Chain rule: $H(XY) = H(X) + H(Y|X)$.*

The probability distribution of a binary random variable is completely characterized by the parameter $p = P_X(0)$. The *binary entropy function* is defined as the entropy of such X, i.e.

$$h(p) \;=\; -p\log p - (1-p)\log(1-p).$$

The *relative entropy* or *discrimination* between two probability distributions P_X and P_Y with the same alphabet $\mathcal{X}$ is defined as

$$D(P_X\|P_Y) = \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{P_Y(x)} \tag{2.10}$$

using the conventions $0\log\frac{0}{q} = 0$ and $p\log\frac{p}{0} = \infty$. The *conditional relative entropy* between P_X and P_Y conditioned on a random variable Z is defined as

$$D(P_{X|Z}\|P_{Y|Z}) = \sum_{z\in\mathcal{Z}} P_Z(z)\sum_{x\in\mathcal{X}} P_{X|Z=z}(x)\log\frac{P_{X|Z=z}(x)}{P_{Y|Z=z}(x)}. \tag{2.11}$$

Similar to entropy, relative entropy is always non-negative. It is zero if and only if $P_X(x) = P_Y(x)$ for all $x \in \mathcal{X}$. One can think of the relative entropy $D(P_X\|P_Y)$ as the increase of information about some experiment when the probability distribution corresponding to the knowledge about the experiment is changed from P_Y to P_X. The total "uncertainty" about X when the distribution is assumed to be P_Y is the relative entropy $D(P_X\|P_Y)$ plus $H(Y)$, the entropy of Y.

The following useful relation connects entropy, relative entropy, and the size of the alphabet: If P_U is the uniform distribution over $\mathcal{X}$, then

$$H(X) + D(P_X \| P_U) \;=\; \log |\mathcal{X}|. \tag{2.12}$$

In other words, if one only knows that a random variable X exists with alphabet $\mathcal{X}$, the "uncertainty" of X is $\log |\mathcal{X}|$. If one learns its distribution P_X, the information about X increases by $D(P_X \| P_U)$ and an uncertainty of $H(X)$ remains.

Relative entropy forms a basis for the definition of mutual information between two random variables X and Y: Let $P_X \times P_Y$ denote the product distribution of P_X and P_Y such that $P_X \times P_Y(x, y) = P_X(x) \cdot P_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. The *mutual information* $I(X;Y)$ of X and Y is the relative entropy between the joint distribution and the product distribution of X and Y, i.e.

$$I(X;Y) \;=\; D(P_{XY} \| P_X \times P_Y).$$

Equivalently, mutual information can be defined as the reduction of the uncertainty of X when Y is learned,

$$I(X;Y) \;=\; H(X) - H(X|Y).$$

$I(X;Y) = I(Y;X)$ follows by symmetry. Similarly, the *conditional mutual information* $I(X;Y|Z)$ of X and Y given a random variable Z is

$$I(X;Y|Z) \;=\; D(P_{XY|Z} \| P_{X|Z} \times P_{Y|Z}).$$

In information theory, communication channels are modeled as systems in which the output depends probabilistically on the input as characterized by a transition probability matrix. The *capacity* of a communication channel with input X and output Y is defined as

$$C \;=\; \max_{P_X} I(X;Y).$$

The second fundamental result of information theory shows that capacity is the maximum rate at which information can be sent over the channel such that it can be recovered at the output with a vanishing probability of error.

2.4 Rényi Entropy

Rényi entropy is perhaps best introduced using the concepts of *generalized probability distributions* and *generalized random variables*, which are extensions of the corresponding ordinary notions to random experiments that cannot always be observed. This presentation follows Rényi's original work [Rén61, Rén70].

Consider a discrete probability space over Ω and let $\Omega_1 \in 2^\Omega$ with $\mathrm{P}[\Omega_1] > 0$. Ω_1 and P define a *generalized discrete probability space* that differs from a probability space only by the fact that $\mathrm{P}[\Omega_1] < 1$ is possible. A random variable X_1 defined on a generalized discrete probability space is called a *generalized discrete random variable*. If $\mathrm{P}[\Omega_1] = 1$, then X_1 is a *complete (or ordinary) random variable*; if $0 < \mathrm{P}[\Omega_1] < 1$, then X_1 is an *incomplete random variable*. X_1 can be interpreted as a quantity resulting from a random experiment that is not always observable, but can be observed only with probability $\mathrm{P}[\Omega_1] < 1$.

The probability distribution P_X of a generalized random variable X is called a *generalized probability distribution*. The *weight* $W(X)$ of X is defined as

$$W(X) \ = \ \sum_{x \in \mathcal{X}} P_X(x).$$

It follows that $0 < W(X) \leq 1$ with equality if and only if X is an ordinary random variable.

Axiomatic characterizations of information measures for random experiments have been studied intensively in the mathematical community [AD75]. An information measure H in this sense associates with every random variable X a number $H(X)$ that corresponds to its information content. Rényi showed that the following five postulates for an information measure uniquely define Shannon entropy [Rén61].

Postulate 1: Reordering the correspondence between values $x \in \mathcal{X}$ and probabilities $P_X(x)$ does not change $H(X)$.

Postulate 2: If X denotes the singleton generalized random variable with $\mathcal{X} = \{x\}$ and $P_X(x) = p$, then $H(X)$ is a continuous function of p for p in the interval $0 < p < 1$.

Postulate 3: If B is a binary random variable with $\mathcal{B} = \{0, 1\}$ and $P_B(0) = P_B(1) = \frac{1}{2}$, then $H(B) = 1$.

Postulate 4: Let X and Y be generalized random variables and define $X \times Y$ as the generalized random variable with alphabet $\mathcal{X} \times \mathcal{Y}$ and distribution $P_{X \times Y}(x, y) = P_X(x) \cdot P_Y(y)$ for all $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. Then

$$H(X \times Y) \;=\; H(X) + H(Y).$$

Postulate 5: Let X and Y be generalized random variables with $W(X) + W(Y) \leq 1$ and $\mathcal{X} \cap \mathcal{Y} = \emptyset$ and define $X \cup Y$ as the generalized random variable with alphabet $\mathcal{X} \cup \mathcal{Y}$ such that $P_{X \cup Y}(x) = P_X(x)$ for $x \in \mathcal{X}$ and $P_{X \cup Y}(y) = P_Y(y)$ for $y \in \mathcal{Y}$. Then

$$H(X \cup Y) \;=\; \frac{W(X)H(X) + W(Y)H(Y)}{W(X) + W(Y)}.$$

Proposition 2.2. *Let H be an information measure for any generalized random variable X that satisfies Postulates 1–5. Then H is uniquely defined and given by*

$$H(X) \;=\; \frac{-\sum_{x \in \mathcal{X}} P_X(x) \log P_X(x)}{\sum_{x \in \mathcal{X}} P_X(x)}.$$

Postulate 5 imposes an arithmetic mean value on the information measure. The general form of a mean value for the numbers $a_1, \ldots, a_n$ with positive weights $w_1, \ldots, w_n$ that sum to 1 is

$$g^{-1}\Big(\sum_{i=1}^{n} w_i g(a_i)\Big)$$

for some monotonic and continuous function g. If the arithmetic mean value in Postulate 5 is replaced by the generalized mean value

$$H(X \cup Y) \;=\; g^{-1}\Big(\frac{W(X)g(H(X)) + W(Y)g(H(Y))}{W(X) + W(Y)}\Big),$$

it can be shown that the only admissible functions g in this context are linear functions $g(x) = ax + b$, which lead to Shannon entropy by Proposition 2.2, and exponential functions[2] $g(x) = 2^{(1-\alpha)x}$, which lead to Rényi entropy [Rén61, Rén65] by Proposition 2.3 below.

[2] Rényi erroneously states $g(x) = 2^{(\alpha-1)x}$ [Rén61], but it is easy to verify that the definition of Rényi entropy requires $g(x) = 2^{(1-\alpha)x}$.

Postulate 5': Let X and Y be generalized random variables such that $W(X) + W(Y) \leq 1$ and define $X \cup Y$ as the generalized random variable with alphabet $\mathcal{X} \cup \mathcal{Y}$ such that $P_{X\cup Y}(x) = P_X(x)$ for $x \in \mathcal{X}$ and $P_{X\cup Y}(y) = P_Y(y)$ for $y \in \mathcal{Y}$. For $\alpha > 0$ and $\alpha \neq 1$, let

$$g_\alpha(x) \;=\; 2^{(1-\alpha)x}.$$

Then

$$H(X \cup Y) \;=\; g_\alpha^{-1}\Big(\frac{W(X)g_\alpha(H(X)) + W(Y)g_\alpha(H(Y))}{W(X) + W(Y)}\Big).$$

For $\alpha > 0$ and $\alpha \neq 1$, the *Rényi entropy of order* α of a generalized random variable X with alphabet $\mathcal{X}$ is defined as

$$H_\alpha(X) \;=\; \frac{1}{1-\alpha} \log \frac{\sum_{x\in\mathcal{X}} P_X(x)^\alpha}{\sum_{x\in\mathcal{X}} P_X(x)}. \tag{2.13}$$

Proposition 2.3 ([Rén61]). *Let H be an information measure for any generalized random variable X that satisfies Postulates 1–4 and Postulate 5'. Then H is uniquely defined and equal to the Rényi entropy H_α.*

In the rest of this section, the properties of Rényi entropy are described only for complete random variables. It is easy to see that $\lim_{\alpha\to 1} H_\alpha(X) = H(X)$. This explains why $H(X)$ can be interpreted as Rényi entropy of order 1 and is sometimes written as $H_1(X)$. Similarly, the *min-entropy* of X, defined as

$$H_\infty(X) \;=\; -\log \max_{x\in\mathcal{X}} P_X(x),$$

results from $\lim_{\alpha\to\infty} H_\alpha(X) = H_\infty(X)$.

On the other end of the interval of admissible α, the Rényi entropy of order 0 can be defined as the logarithm of the alphabet size,

$$H_0(X) \;=\; \log |\mathcal{X}|$$

using the convention $0^0 = 1$. An important property of Rényi entropy is shown in the following proposition.

Proposition 2.4. *Rényi entropy $H_\alpha(X)$ for $\alpha \geq 0$ is a positive decreasing function of α, i.e., for $0 \leq \alpha < \beta$*

$$H_\alpha(X) \;\geq\; H_\beta(X) \tag{2.14}$$

with equality if and only if X is uniformly distributed over $\mathcal{X}$ when $\alpha = 0$ or X is uniformly distributed over a subset of $\mathcal{X}$ when $\alpha > 0$.

Proof. For $0 \leq \alpha < \beta$ with $\alpha \neq 1$ and $\beta \neq 1$,

$$\begin{aligned}
H_\alpha(X) &= \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X(x)^\alpha \\
&= -\log \mathrm{E}\Big[P_X(X)^{\alpha-1}\Big]^{\frac{1}{\alpha-1}} \\
&= -\log \mathrm{E}\Big[P_X(X)^{\alpha-1}\Big]^{\frac{\beta-1}{\alpha-1}\frac{1}{\beta-1}} \\
&\geq -\log \mathrm{E}\Big[P_X(X)^{(\alpha-1)\frac{\beta-1}{\alpha-1}}\Big]^{\frac{1}{\beta-1}} \\
&= -\log \mathrm{E}\Big[P_X(X)^{\beta-1}\Big]^{\frac{1}{\beta-1}} \\
&= \frac{1}{1-\beta} \log \sum_{x \in \mathcal{X}} P_X(x)^\beta \\
&= H_\beta(X).
\end{aligned}$$

We observe that x^c is convex-∪ for $c \geq 1$ or $c \leq 0$ and convex-∩ for $0 \leq c \leq 1$. The inequality in the above derivation follows from the Jensen inequality in the following cases:

$\beta > \alpha > 1 : \ c = \frac{\beta-1}{\alpha-1} > 1$, x^c is convex-∪ and $\frac{1}{\beta-1} > 0$;

$\beta > 1 > \alpha \geq 0 : \ c = \frac{\beta-1}{\alpha-1} < 0$, x^c is convex-∪ and $\frac{1}{\beta-1} > 0$;

$1 > \beta > \alpha \geq 0 : \ 1 > c = \frac{\beta-1}{\alpha-1} > 0$, x^c is convex-∩ and $\frac{1}{\beta-1} < 0$.

For $\alpha = 1$ or $\beta = 1$, the Jensen inequality can be applied directly. The conditions for equality in (2.14) follow directly from the Jensen inequality. □

We define *conditional Rényi entropy* $H_\alpha(X|Y)$ similar to conditional entropy as

$$H_\alpha(X|Y) \ = \ \sum_{y \in \mathcal{Y}} P_Y(y) H_\alpha(X|Y=y). \tag{2.15}$$

(There seems to be no agreement about a standard definition of conditional Rényi entropy.) In contrast to Shannon entropy, however, both the chain rule and the property that conditioning on side information reduces entropy (Proposition 2.1) do not hold for this definition of conditional Rényi entropy; $H_\alpha(X) > H_\alpha(X|Y)$ and $H_\alpha(XY) \neq H_\alpha(X) + H_\alpha(X|Y)$ are possible in general. Other definitions of conditional Rényi entropy are examined by Csiszár [Csi95b].

2.5 The Asymptotic Equipartition Property

The *Asymptotic Equipartition Property (AEP)* is a form of the law of large numbers used in information theory. The AEP states that the alphabet of a sequence $X^n = X_1, \ldots, X_n$ of n independent, identically distributed (i.i.d.) random variables with distribution P_X can be divided into two sets, a typical set and a non-typical set, and that the probability of the typical set goes to 1 as $n \to \infty$. Furthermore, all typical sequences are almost equally probable and the probability of a typical sequence is close to $2^{-nH(X)}$.

The material in this section is based on the presentation by Cover and Thomas [CT91] that uses the method of types, which is closely related to the method of strongly typical sequences.

Fix an alphabet $\mathcal{X}$. Let x^n be a sequence of n symbols from $\mathcal{X}$. The *type* or *empirical probability distribution* Q_{x^n} of x^n is the mapping that specifies the relative proportion of occurrences of each symbol a of $\mathcal{X}$, i.e., $Q_{x^n}(a) = \frac{N_a(x^n)}{n}$ for all $a \in \mathcal{X}$, where $N_a(x^n)$ is the number of times the symbol a occurs in the sequence x^n.

The type Q_{x^n} of a sequence x^n can be interpreted as an empirical probability distribution. The *set of types with denominator* n is denoted by $\mathcal{Q}_n$.

For a particular type $Q \in \mathcal{Q}_n$, the set of sequences of length n and type Q is called the *type class* of Q and is denoted by $T(Q)$:

$$T(Q) \;=\; \{x^n \in \mathcal{X}^n : Q_{x^n} = Q\}.$$

The following proposition summarizes the basic properties of types.

Proposition 2.5 ([CT91, CK81]). *Let $X^n = X_1, \ldots, X_n$ be a sequence of n i.i.d. random variables with distribution P_X and alphabet $\mathcal{X}$ and let $\mathcal{Q}_n$ be the set of types. Then*

1. *The number of types with denominator n is at most polynomial in n, more particularly*

$$|\mathcal{Q}_n| \;\leq\; (n+1)^{|\mathcal{X}|}. \tag{2.16}$$

2. *The probability of a sequence x^n depends only on its type and is given by*

$$P_{X^n}(x^n) \;=\; 2^{-n(H(Q_{x^n})+D(Q_{x^n}\|P_X))}. \tag{2.17}$$

3. *For any $Q \in \mathcal{Q}_n$, the size of the type class $T(Q)$ is approximately $2^{nH(Q)}$. More precisely,*
$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(Q)} \;\leq\; |T(Q)| \;\leq\; 2^{nH(Q)}. \tag{2.18}$$
4. *For any $Q \in \mathcal{Q}_n$, the probability of the type class $T(Q)$ is approximately $2^{-nD(Q\|P_X)}$. More precisely,*
$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-nD(Q\|P_X)} \;\leq\; \sum_{x^n \in T(Q)} P_{X^n}(x^n) \;\leq\; 2^{-nD(Q\|P_X)}. \tag{2.19}$$

An important property of types is that there are only a polynomial number of types but an exponential number of sequences of each type. Since the probability of each type class depends exponentially on the relative entropy distance between the type Q and the distribution P_X, type classes that are far from the true distribution P_X have exponentially small probability. Type classes close to P_X have high probability and form the typical set.

Given an $\epsilon > 0$, we define the *(strongly) typical set* S_ϵ^n of sequences of length n for the distribution P_X as
$$S_\epsilon^n \;=\; \left\{ x^n \in \mathcal{X}^n : \left| \frac{1}{n} N_a(x^n) - P_X(a) \right| \leq \frac{\epsilon}{|\mathcal{X}|}, \text{ for all } a \in \mathcal{X} \right\}$$
where $N_a(x^n)$ denotes the number of times the symbol a occurs in the sequence x^n. We are now ready to state the AEP for strongly typical sequences with a proof based on the work of Csiszár and Körner [CK81].

Proposition 2.6 (AEP). *Let $X^n = X_1, \ldots, X_n$ be a sequence of n i.i.d. random variables with distribution P_X and let $\epsilon \leq \frac{1}{2}$. Then*

1. *The typical set has probability almost 1:*
$$\mathrm{P}\big[X^n \in S_\epsilon^n\big] \;>\; 1 - (n+1)^{|\mathcal{X}|} \cdot 2^{-\frac{n}{2\ln 2}\frac{\epsilon^2}{|\mathcal{X}|^2}}. \tag{2.20}$$
2. *The number of elements in the typical set is close to $2^{nH(X)}$, or*
$$H(X) + \frac{\epsilon}{n}\log\frac{\epsilon}{|\mathcal{X}|} - |\mathcal{X}|\frac{\log(n+1)}{n} \;\leq\; \frac{1}{n}\log|S_\epsilon^n| \;\leq$$
$$H(X) - \frac{\epsilon}{n}\log\frac{\epsilon}{|\mathcal{X}|} + |\mathcal{X}|\frac{\log(n+1)}{n} \tag{2.21}$$

3. *The probability of a typical sequence is upper bounded by an expression close to* $2^{-nH(X)}$, *i.e.*

$$\forall x^n \in S_\epsilon^n \ : \ P_{X^n}(x^n) \ \leq \ 2^{-n(H(X)+\epsilon \log \frac{\epsilon}{|\mathcal{X}|})}. \tag{2.22}$$

Proof. Because $Q_{x^n}(x) = \frac{1}{n} N_x(x^n)$, the type of any non-typical x^n satisfies $\|Q_{x^n} - P_X\|_1 \ > \ \frac{\epsilon}{|\mathcal{X}|}$ and

$$D(Q_{x^n}\|P_X) \ > \ \frac{1}{2\ln 2} \cdot \frac{\epsilon^2}{|\mathcal{X}|^2}$$

by Lemma 3.3 (see page 42). Using (2.16) and (2.19), we can bound the probability of all non-typical sequences in (2.20),

$$\begin{aligned} \mathrm{P}[X^n \notin S_\epsilon^n] \ &= \ \sum_{Q \in \mathcal{Q}_n : \|Q - P_X\|_1 > \frac{\epsilon}{|\mathcal{X}|}} \ \sum_{x^n : x^n \in T(Q)} P_{X^n}(x^n) \\ &< \ (n+1)^{|\mathcal{X}|} \cdot 2^{-n \frac{1}{2\ln 2} \frac{\epsilon^2}{|\mathcal{X}|^2}}. \end{aligned}$$

To prove the second statement of the theorem, note that all $x^n \in S_\epsilon^n$ satisfy $\|Q_{x^n} - P_X\|_1 \leq \epsilon$ and thus also

$$|H(Q_{x^n}) - H(X)| \ \leq \ -\epsilon \log \frac{\epsilon}{|\mathcal{X}|} \tag{2.23}$$

if $\epsilon \leq \frac{1}{2}$, by Lemma 3.5 (see page 43). This can be used, together with the third property of types (2.18), to bound the size of the type class $T(Q_{x^n})$ for any typical sequence x^n,

$$\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(X)+\epsilon \log \frac{\epsilon}{|\mathcal{X}|}} \ \leq \ |T(Q_{x^n})| \ \leq \ 2^{nH(X)-\epsilon \log \frac{\epsilon}{|\mathcal{X}|}}. \tag{2.24}$$

Clearly, $T(Q_{x^n}) \subseteq S_\epsilon^n$ if $x^n \in S_\epsilon^n$ which establishes the lower bound in 2.21. Similarly, S_ϵ^n is contained in the union of the type classes corresponding to all typical sequences. Thus, the upper bound in (2.24) can be extended to all $(n+1)^{|\mathcal{X}|}$ types by (2.16). The result is equivalent to (2.21).

Finally, it follows again from (2.23) by the non-negativity of relative entropy that for any typical sequence x^n

$$H(Q_{x^n}) + D(Q_{x^n}\|P_X) \ \geq \ H(X) + \epsilon \log \frac{\epsilon}{|\mathcal{X}|}.$$

The bound on the probability of a typical sequence (2.22) follows immediately from the second property of types (2.17). □

2.6 Universal Hashing and Privacy Amplification

Privacy amplification by universal hashing is a key component of many unconditionally secure cryptographic protocols and is a recurring theme of this work. This section presents an idealized cryptographic scenario to introduce privacy amplification. Applications, motivating discussions, and further references are provided in Chapters 4 and 5.

Privacy amplification is based on *universal hash functions*, introduced by Carter and Wegman [CW79, WC81]. Universal hash functions were first used for privacy amplification by Bennett, Brassard, and Robert [BBR86].

Definition 2.1. A *k-universal hash function* is a set $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$ such that for all distinct $x_1, \ldots, x_k \in \mathcal{X}$, there are at most $|\mathcal{G}|/|\mathcal{Y}|^{k-1}$ functions g in $\mathcal{G}$ such that $g(x_1) = g(x_2) = \cdots = g(x_k)$.

A k-universal hash function is l-universal for all $l < k$. The term *universal hash function* usually refers to a 2-universal hash function. There is a stronger notion of universal hash functions, which is closely related to k-wise independent random variables (see Section 5.4.2 and [LW95]).

Definition 2.2. A *strongly k-universal hash function* is a set $\mathcal{G}$ of functions $\mathcal{X} \to \mathcal{Y}$ such that for all distinct $x_1, \ldots, x_k \in \mathcal{X}$ and all (not necessarily distinct) $y_1, \ldots, y_k \in \mathcal{Y}$, exactly $|\mathcal{G}|/|\mathcal{Y}|^k$ functions from $\mathcal{G}$ take x_1 to y_1, x_2 to y_2, ..., x_k to y_k.

Assume Alice and Bob share a random variable W, while an eavesdropper Eve knows a correlated random variable V that summarizes her knowledge about W. The details of the distribution P_{WV} are unknown to Alice and Bob except that they know a lower bound on the Rényi entropy of $P_{W|V=v}$ for the particular value v of Eve's knowledge V about W. For example, Eve might have received some symbols of W or the result of some function of W, such as some parity bits. However, Alice and Bob do not know more about Eve's knowledge than the fact that it satisfies a lower bound on Rényi entropy of order 2.

Using an authenticated public channel, which is susceptible to eavesdropping but immune to tampering, Alice and Bob wish to agree on a function g such that Eve knows nearly nothing about $g(W)$. The following theorem by Bennett, Brassard, Crépeau, and Maurer [BBCM95] shows that if Alice and Bob choose g at random from a universal hash function $\mathcal{G} : \mathcal{W} \rightarrow \mathcal{Y}$ for suitable $\mathcal{Y}$, then Eve's information about $Y = g(W)$ is negligible.

Theorem 2.7 (Privacy Amplification Theorem). *Let X be a random variable over the alphabet $\mathcal{X}$ with probability distribution P_X and Rényi entropy $H_2(X)$, let G be the random variable corresponding to the random choice (with uniform distribution) of a member of a 2-universal hash function $\mathcal{G} : \mathcal{X} \rightarrow \mathcal{Y}$, and let $Y = G(X)$. Then*

$$H(Y|G) \;\geq\; H_2(Y|G) \;\geq\; \log|\mathcal{Y}| - \frac{2^{\log|\mathcal{Y}| - H_2(X)}}{\ln 2}. \tag{2.25}$$

In the statement above, G is a random variable and the quantity $H(Y|G) = H(G(X)|G)$ is an average over all choices of the function g. It is possible that $H(G(X)|G = g) = H(g(X))$ differs from $\log|\mathcal{Y}|$ by a non-negligible amount for some g, but such a g can occur only with negligible probability when $\log|\mathcal{Y}| < H_2(X)$. Thus the entropy of Y is almost maximal and the distribution of Y is close to uniform.

This theorem applies also to conditional probability distributions such as $P_{W|V=v}$ discussed above. If Eve's Rényi entropy $H_2(W|V = v)$ is known to be at least t and Alice and Bob choose an r-bit string $Y = G(W)$ as their secret key, then

$$H(Y|G, V = v) \;\geq\; r - 2^{r-t}/\ln 2.$$

The key Y is virtually secret because $H(Y|G, V = v)$ is arbitrarily close to the maximum r. More precisely, if $r < t$, then Eve's total information about S decreases exponentially in the excess compression $t - r$.

A previous version of this theorem developed by Bennett, Brassard, and Robert was restricted to *deterministic* eavesdropping functions $e(\cdot)$ that Eve might use to get $V = e(W)$ [BBR86, BBR88]. This result was then generalized to probabilistic eavesdropping strategies.

It should be pointed out that Theorem 2.7 cannot be generalized to Rényi entropy conditioned on a random variable, i.e., $H(Y|GV) \geq r - 2^{r-H_2(W|V)}/\ln 2$ is false in general [BBCM95].

Chapter 3

Information Measures in Cryptography

3.1 Introduction

Information measures are the main abstractions for modeling cryptographic scenarios with information-theoretic methods. This chapter presents a survey of several information measures and their significance for cryptographic systems that provide unconditional security.

The fundamental concepts of information theory are definitions of measures for the *uncertainty* of the outcome of a random experiment; information is measured as the *reduction* of uncertainty. An entropy measure is a mapping from probability distributions to the real numbers that associates a number with every probability distribution.

Entropy measures play two different roles in connection with unconditionally secure cryptographic systems. On one hand, positive results can be obtained in the form of information-theoretic security proofs for such systems. On the other hand, lower bounds on the required key sizes in some scenarios are negative results that follow from entropy-based arguments.

Two separate aspects of any entropy measure are its formal definition and its operational characterization(s). An entropy measure is usually defined *formally* in terms of the probability distribution, and its numerical value can be computed immediately for any given probability distribution. The justification for a definition is given by an *operational* *characterization* of the entropy measure, that is, an application scenario in which the entropy measure gives an answer to an important question arising from the context. In this chapter, we will focus on operational characterizations in cryptography for a number of entropy measures defined formally in terms of a probability distribution.

The reverse process is also possible. In Chapter 4, for instance, we give an operational definition of an entropy measure (called smooth entropy) that quantifies the number of uniform bits that can be extracted from a random source by probabilistic algorithms. We then search for bounds on its numerical value and its relation to formally defined entropy measures. Another example of an operationally defined information measure is the secrecy capacity of key agreement from common information by public discussion introduced by Maurer [Mau93, MW97].

The formal definitions of the information measures discussed in this chapter are summarized below.

Shannon Entropy. The *Shannon entropy* of a random variable X is

$$H(X) \;=\; -\sum_{x\in\mathcal{X}} P_X(x)\log P_X(x).$$

Relative Entropy. The *relative entropy* or *discrimination* between two probability distributions P_X and P_Y with the same alphabet $\mathcal{X}$ is

$$D(P_X\|P_Y) \;=\; \sum_{x\in\mathcal{X}} P_X(x)\log\frac{P_X(x)}{P_Y(x)}.$$

Rényi Entropy of Order α. For $\alpha \geq 0$ and $\alpha \neq 1$, the *Rényi entropy of order* α of a random variable X is

$$H_\alpha(X) \;=\; \frac{1}{1-\alpha}\log\sum_{x\in\mathcal{X}} P_X(x)^\alpha.$$

Min-Entropy. The *min-entropy* of a random variable X is

$$H_\infty(X) \;=\; -\log\max_{x\in\mathcal{X}} P_X(x).$$

"Guessing Entropy." For a random variable X with n values, denote the elements of its probability distribution P_X by $p_1,\ldots,p_n$ such that $p_1 \geq p_2 \geq \cdots \geq p_n$. The average number of guesses needed to determine X using an optimal strategy is

$$\mathrm{E}[G(X)] \;=\; \sum_{i=1}^{n} i p_i .$$

We call this quantity the *guessing entropy* of X.

Collision Probability. The *collision probability* of a random variable X is

$$P_2(X) \;=\; \sum_{x \in \mathcal{X}} P_X(x)^2 .$$

Variational Distance. The *variational distance* between two probability distributions P_X and P_Y with the same alphabet $\mathcal{X}$ is

$$\|P_X - P_Y\|_v \;=\; \max_{\mathcal{X}_0 \subseteq \mathcal{X}} \Big| \sum_{x \in \mathcal{X}_0} P_X(x) - P_Y(x) \Big| .$$

A notable omission from this list is *quantum information*, which forms the basis of quantum cryptography. However, there is no notion of "quantum entropy" in terms of a probability distribution. We refer to the surveys by Brassard and Crépeau [BC96] and Spiller [Spi96] and the references therein for treatments of quantum cryptography and its foundations.

The chapter is organized as follows. Cryptographic applications of several information measures are presented next in Section 3.2, which is the main part of the chapter. An overview of bounds for relating the different information measures is the subject of Section 3.3, and Section 3.4 examines converting lower bounds on the Shannon entropy of a random variable into cryptographically more relevant information measures.

3.2 Scenarios and Information Measures

This section contains a selective survey of cryptographic scenarios in which the information measures mentioned above play an important role. The classical dual goals of cryptography, secrecy and authentication, are covered first, together with their relation to the two fundamental concepts of information theory, entropy and relative entropy (or discrimination). The material in this section is based on the cited references.

3.2.1 Perfect Secrecy: Shannon Entropy

The notion of *perfect security* or *perfect secrecy* was introduced by Shannon and means that an adversary does not have any information at all about some secret, typically the secret plaintext to be transmitted in a cryptographic system. This is equivalent to saying that the random variable constituting the secret and the random variable modeling the adversary's knowledge are independent. Perfect security has been used among other applications for symmetric cryptosystems, secret sharing, and conference key distribution. In all three cases, lower bounds on the amount of information that a participant has to transmit or to store have been obtained, which suggest that perfectly secure systems are impractical. For every one of these three applications, there is a construction that can meet the lower bound with equality in many cases. Interestingly, these constructions are remarkably similar.

Secret-Key Cryptosystems

Shannon's model of a secret-key cryptosystem consists of a sender Alice, a receiver Bob, an eavesdropper Eve, and an open channel from Alice to Bob and to Eve [Sha49]. The secret key Z is known to Alice and to Bob only. Alice encrypts the plaintext X using Z according to the encryption rule of the system, resulting in the cryptogram Y that is sent to Bob and can also be received by Eve. Bob can recover X with his knowledge of Z.

A cipher in this model is called *perfect* if and only if the plaintext X and the cryptogram Y are independent random variables, i.e. if and only if $I(X;Y) = 0$ or, equivalently, $H(X|Y) = H(X)$. Bob must be able to recover X uniquely from Y and Z, i.e. $H(X|YZ) = 0$.

It follows that $H(Z) \geq H(X)$ for any cryptosystem with perfect security, because

$$\begin{aligned} H(X) = H(X|Y) \leq H(XZ|Y) &= H(Z|Y) + H(X|YZ) \\ &= H(Z|Y) \leq H(Z). \end{aligned} \tag{3.1}$$

This is Shannon's famous "impracticality result" that the entropy of the secret key must be at least as large as the entropy of the plaintext to be encrypted.

Vernam's *one-time pad* is the prime example of a perfectly secure cryptosystem. It uses a randomly and uniformly chosen n-bit secret key $Z \in \{0,1\}^n$ to encrypt (and decrypt) the n-bit plaintext $X \in \{0,1\}^n$ with a simple XOR-operation.

Secret Sharing

Secret sharing is an important and widely studied tool in cryptography and distributed computation [Sti92]. A *perfect secret sharing scheme* is a protocol in which a dealer distributes a secret S among a set of participants such that only specific subsets of them, defined by the *access structure*, can recover the secret at a later time and any non-qualified subset can obtain no information about the secret.

If the access structure allows any subset of k or more of the n participants to reconstruct the secret but no set of $k-1$ participants or less to do so, the secret sharing scheme is called a *threshold scheme*. It can be implemented with Shamir's construction [Sha79] based on polynomial interpolation.

Given a set of participants $\mathcal{P} = \{A, B, C, \ldots\}$ and a dealer $\mathcal{D}$, where $\mathcal{D} \notin \mathcal{P}$ is assumed, the *access structure* $\mathcal{A} \subseteq 2^{\mathcal{P}}$ is a family of subsets of $\mathcal{P}$ containing the sets of participants qualified to recover the secret. It is natural to require $\mathcal{A}$ to be monotone, that is, if $X \in \mathcal{A}$ and $X \subseteq X' \subseteq \mathcal{P}$, then $X' \in \mathcal{A}$. Let S be the secret to share and denote the random variables describing the shares given to a participant $P \in \mathcal{P}$ or to a group of participants $X \subseteq \mathcal{P}$ also by P or by X, respectively.

A secret sharing scheme is called *perfect* if any set of qualified participants $X \in \mathcal{A}$ can uniquely determine S, i.e. $H(S|X) = 0$, but any unqualified set $X \notin \mathcal{A}$ can obtain no information about S, i.e. $H(S|X) = H(S)$. By an argument similar to Shannon's lower bound for perfect security (3.1), it follows that $H(P) \geq H(S)$ for the share of any participant $P \in \mathcal{P}$.

For many access structures, it can be proved that some shares in a perfect scheme have to be considerably larger than the secret. Consider the set $\mathcal{P} = \{A, B, C, D\}$ with the access structure $\mathcal{A}$ equal to the monotone closure of $\{AB, BC, CD\}$. Using the definition of a perfect secret sharing scheme, Capocelli et al. [CSGV92] derive the lower bound $H(BC) \geq 3H(S)$, which implies $H(B) \geq 1.5H(S)$ or $H(C) \geq 1.5H(S)$. Thus at least one share, B or C, must be 1.5 times the length of the secret or longer.

A perfect secret sharing scheme for this example can be realized as follows [CSGV92]. Let S be an n-bit string of even length and denote its first half by S_a and its second half by S_b. Choose four $n/2$-bit strings $R_1, \ldots, R_4$ randomly with uniform distribution and let $A = [R_1, S_b \oplus R_3]$, $B = [R_1 \oplus S_a, R_3, R_4]$, $C = [R_1, R_2, R_4 \oplus S_b]$, and $D = [R_2 \oplus S_a, R_4]$ be the shares, where $\oplus$ denotes the XOR operation. This construction meets the lower bound with equality, since $H(B) = H(C) = 1.5H(S)$.

Recursive application of an argument similar to the lower bound above shows that there are access structures on n participants for infinitely many n where the size of some shares must be at least $n/\log n$ times the size of the secret [Csi95a]. The known general techniques for realizing perfect secret sharing schemes [BL90, ISN93] produce exponentially large shares, but there is still an open gap between lower and upper bounds for the size of a share in the general case.

Key Distribution for Dynamic Conferences

A *key distribution scheme* (KDS) for dynamic conferences is a method by which initially a dealer distributes private individual pieces of information to a set of users $\mathcal{P}$ [Blo85, BSH+93, Sti96]. Later, any qualified conference of users, contained in the *key structure* $\mathcal{A} \subseteq 2^{\mathcal{P}}$, is able to compute a common key without further interaction, i.e. each user needs only his private piece of information and the identities of the other conference members. The key of each conference is perfectly secure against all coalitions of users contained in the *forbidden structure* $\mathcal{F} \subseteq 2^{\mathcal{P}}$, in the sense that even if a set $X \in \mathcal{F}$ of malicious users pool their pieces together, they have no information about the key of any other conference Y that is disjoint from X (i.e. such that $X \cap Y = \emptyset$).

We adopt the same notation as above and denote users and the random variables representing their pieces of information interchangeably with capital letters. In this terminology, the key S_X of any conference $X \in \mathcal{A}$ can be computed by any user $A \in X$ without further interaction and therefore satisfies $H(S_X|A) = 0$. In addition, for every key S_X and any set $Y \in \mathcal{F}$ that is disjoint from X, it is required that $H(S_X|Y) = H(S_X)$.

Consider the important special case of a key distribution scheme for a set of n participants in which the key structure $\mathcal{A}$ consist of all subsets of a certain cardinality as also does the forbidden structure $\mathcal{F}$. A *k-secure non-interactive t-conference key distribution scheme* for some $k \leq n - t$ is defined as a distribution scheme for information pieces such that $\mathcal{A}$ consists of every set $X \subseteq \mathcal{P}$ of cardinality t and $\mathcal{F}$ contains all sets

$X \subseteq \mathcal{P}$ of cardinality at most k.

Assuming that the common keys of all groups of t users from $\mathcal{P}$ have the same entropy $H(S)$, Blundo et al. [BSH+93] show that in any k-secure non-interactive t-conference KDS

$$H(A) \;\geq\; \binom{k+t-1}{t-1} H(S) \tag{3.2}$$

for each user $A \in \mathcal{P}$. As an example, consider a KDS for the particular case of n users with $t = 2$ and $k = n - 2$. Every two users share a key, which is secure against a coalition of the remaining $n - 2$ users. Then, the size of the information that a user needs to store must be at least $n - 1$ times the size of a common key and the total information to be stored by all n users is at least $n(n-1)H(S)$. This is the well-known "n^2 problem" of establishing a secret key between every pair from a set of n users.

The following general construction meets the bound (3.2) [BSH+93]. Denote the set of users by $\mathcal{P} = \{1, \ldots, n\}$ and let $GF(q)$ be the finite field with q elements for some prime $q > n$. The dealer chooses a symmetric polynomial $p(x_1, \ldots, x_t)$ over $GF(q)$ of degree k uniformly at random. The symmetric polynomial satisfies $p(x_1, \ldots, x_t) = p(x_{\sigma(x_1)}, \ldots, x_{\sigma(x_t)})$ for all permutations $\sigma : \{1, \ldots, t\} \to \{1, \ldots, t\}$. The dealer sends to user i for $i = 1, \ldots, n$ the polynomial $f_i(x_2, \ldots, x_t) = p(i, x_2, \ldots, x_t)$, obtained by evaluating $p(\cdot)$ at $x_1 = i$. When the users in a set $\{i_1, \ldots, i_t\} \in \mathcal{A}$ want to set up their conference key, each user i_j evaluates f_{i_j} at $(x_2, \ldots, x_t) = (i_1, \ldots, i_{j-1}, i_{j+1}, \ldots, i_t)$ for $j = 1, \ldots, t$. The common key for the users $i_1, \ldots, i_t$ is equal to $S_{i_1,\ldots,i_t} = p(i_1, \ldots, i_t)$.

If interaction among the users is allowed after the dealer has finished distributing the pieces of information, more efficient schemes are possible. For example, a k-secure *interactive* t-conference key distribution scheme can be constructed from a $k + t - 2$-secure non-interactive 2-conference KDS [BSH+93]. The idea is that a designated member of a group of t users chooses the key at random and sends it to the other $t-1$ group members, encrypted using a one-time pad with the keys obtained in the $k+t-2$-secure 2-conference KDS. This protocol is secure against coalitions of up to k users because the random key is encrypted $t - 1$ times with different keys that are $k+t-2$-secure each (see also [BC94]).

3.2.2 Authentication: Relative Entropy

Message authentication techniques are the focus of *authentication theory*. Authentication provides assurance to the receiver of a message that it originates from the specified, legitimate sender, although an adversary may have intercepted, modified, or inserted the message. Unconditionally secure authentication assumes unlimited computational power for an adversary and is based on secret information shared by the sender and the receiver [WC81, Mas91]. This section is based on work of Maurer [Mau96], which shows that the problem of deciding whether a received message is authentic or not can be seen as a *hypothesis testing* problem. The receiver must decide which one of two hypotheses is true: either the message was generated by the legitimate sender in possession of the secret key or by an opponent without knowledge of the secret key. A central information measure for hypothesis testing is relative entropy or discrimination.

Hypothesis testing is the task of deciding which one of two hypotheses H_0 or H_1 is the true explanation for an observed measurement Q [Bla87]. In other words, there are two possible probability distributions, denoted by P_{Q_0} and P_{Q_1}, over the space $\mathcal{Q}$ of possible measurements. If H_0 is true, then Q was generated according to P_{Q_0}, and if H_1 is true, then Q was generated according to P_{Q_1}. A *decision rule* is a binary partitioning of $\mathcal{Q}$ that assigns one of the two hypotheses to each possible measurement $q \in \mathcal{Q}$. The two possible errors that can be made in a decision are called a *type I error* for accepting hypothesis H_1 when H_0 is actually true and a *type II error* for accepting H_0 when H_1 is true. The probability of a type I error is denoted by α, the probability of a type II error by β.

A method for finding the optimum decision rule is given by the Neyman-Pearson theorem. The decision rule is specified in terms of a threshold parameter T; α and β are then functions of T. For any given threshold $T \in \mathbb{R}$ and a given maximal tolerable probability β of type II error, α can be minimized by assuming hypothesis H_0 for an observation $q \in \mathcal{Q}$ if and only if

$$\log \frac{P_{Q_0}(q)}{P_{Q_1}(q)} \ \geq \ T. \tag{3.3}$$

To find the optimal decision rule, many values of T must be examined in general. The term on the left in (3.3) is called the *log-likelihood ratio*. The expected value of the log-likelihood ratio with respect to P_{Q_0} is equal to the relative entropy $D(P_{Q_0} \| P_{Q_1})$ between P_{Q_0} and P_{Q_1}, which makes relative entropy an important information measure for distinguishing probability distributions by hypothesis testing. Let $d(\alpha, \beta)$ be defined as

$$d(\alpha, \beta) \;=\; \alpha \log \frac{\alpha}{1-\beta} + (1-\alpha) \log \frac{1-\alpha}{\beta}.$$

A fundamental result of hypothesis testing states that the type I and type II error probabilities α and β satisfy

$$d(\alpha, \beta) \;\leq\; D(P_{Q_0} \| P_{Q_1}), \tag{3.4}$$

which implies for $\alpha = 0$ that $\beta \geq 2^{-D(P_{Q_0} \| P_{Q_1})}$.

A similar result holds also for a generalized hypothesis testing scenario where the distributions P_{Q_0} and P_{Q_1} depend on an additional random variable V. The decision rule, the probability distributions, and the error probabilities are now parameterized by V (e.g. $P_{Q_0|V=v}$ for $v \in \mathcal{V}$), and the average type I and type II errors are $\overline{\alpha} = \sum_{v \in \mathcal{V}} P_V(v)\alpha(v)$ and $\overline{\beta} = \sum_{v \in \mathcal{V}} P_V(v)\beta(v)$. The bound (3.4) becomes

$$d(\overline{\alpha}, \overline{\beta}) \;\leq\; D(P_{Q_0|Z} \| P_{Q_1|Z}). \tag{3.5}$$

Consider a scenario with a sender Alice who wants to send a sequence of plaintext messages $X_1, \ldots, X_n$ to a receiver Bob. Alice and Bob share a secret key Z that is used to authenticate each message X_i separately by encoding it as Y_i in a way depending on Z. We assume that Bob can determine X_i uniquely from $Y_1, \ldots, Y_{i-1}$ and Z for all $i = 1, \ldots, n$. Thus, Bob must decide about the authenticity of the i-th message Y_i based on $Y_1, \ldots, Y_{i-1}$ and Z.

An opponent Eve with reading and writing access to the communications channel from Alice to Bob can use two different strategies for cheating. In an *impersonation attack* at time i, Eve waits until she has observed (but not modified) $Y_1, \ldots, Y_{i-1}$ and then sends a forged message $\widetilde{Y}_i$ that she hopes to be accepted by Bob as Y_i. We denote Eve's impersonation success probability for the particular observed sequence $Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}$ by $p_i^I(y_1, \ldots, y_{i-1})$ and her overall impersonation success probability by $p_i^I = \mathrm{E}[p_i^I(Y_1, \ldots, Y_{i-1})]$, both computed assuming that Eve uses an optimum strategy, i.e. one that maximizes the probability of successful impersonation.

In a *substitution attack* at time i, Eve observes $Y_1, \ldots, Y_{i-1}$, intercepts Y_i, and replaces it with a different value $\widetilde{Y}_i$. The overall substitution success probability that Bob accepts $\widetilde{Y}_i$ as valid and decodes it to some $\widetilde{X}_i \neq X_i$ is denoted by p_i^S, and the substitution success probability for the particular observed sequence $Y_1 = y_1, \ldots, Y_i = y_i$ is denoted by $p_i^S(y_1, \ldots, y_i)$, again assuming that Eve uses an optimal strategy. Thus, $p_i^S = \mathrm{E}[p_i^S(Y_1, \ldots, Y_i)]$.

Consider an impersonation attack by Eve after the particular sequence $Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}$ has been transmitted in the first $i-1$ steps. Bob sees a message $\overline{Y}_i$ and has to decide, using his knowledge of the key Z, whether it is a correct message Y_i from Alice (hypothesis H_0) or a fraudulent message $\widetilde{Y}_i$ inserted by Eve (hypothesis H_1). If Bob rejects a valid message from Alice, a type I error results (probability α), and if Bob accepts a message from Eve, a type II error results (probability β). Thus, under hypothesis H_0, the distribution of $\overline{Y}_i Z$ is $P_{Y_i Z|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}}$, since Alice knows the secret key Z. However, because Eve has no a priori knowledge about Z, the distribution of $\overline{Y}_i Z$ under hypothesis H_1 is $P_{\widetilde{Y}_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}} \times P_{Z|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}}$ (where $P_A \times P_B$ denotes the product distribution of P_A and P_B as in Section 2.3). Eve is free to choose any distribution $P_{\widetilde{Y}_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}}$, in particular also the distribution $P_{Y_i|Y_1=y_1,\ldots,Y_{i-1}=y_{i-1}}$. In this case, it follows from (3.4) and from the definition of mutual information that

$$d(\alpha, p_i^I(y_1, \ldots, y_{i-1})) \;\leq\; I(Y_i; Z|Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}),$$

which is a lower bound for the impersonation probability at round i for the previous messages $Y_1 = y_1, \ldots, Y_{i-1} = y_{i-1}$ when Bob is constrained to reject valid messages from Alice with probability at most α. The corresponding bound on the overall impersonation success probability p_i^I follows from (3.5):

$$d(\overline{\alpha}, p_i^I) \;\leq\; I(Y_i; Z|Y_1 \cdots Y_{i-1}),$$

which reduces to

$$p_i^I \geq 2^{-I(Y_i;Z|Y_1\cdots Y_{i-1})} \tag{3.6}$$

for $\overline{\alpha} = 0$.

Using similar arguments, one can derive lower bounds on the substitution success probability $p_i^S(y_1, \ldots, y_i)$ for the particular sequence $Y_1 = y_1, \ldots, Y_i = y_i$ and the average substitution attack probability p_i^S for $\alpha = \overline{\alpha} = 0$ [Mau96]:

$$p_i^S(y_1, \ldots, y_i) \;\geq\; 2^{-H(Z|Y_1=y_1,\ldots,Y_i=y_i)}$$

and

$$p_i^S \geq 2^{-H(Z|Y_1 \cdots Y_i)}. \tag{3.7}$$

Combining the bounds (3.6) and (3.7) on the impersonation and substitution success probabilities yields

$$p_i^I \cdot p_i^S \geq 2^{-H(Z|Y_1 \cdots Y_{i-1})}$$

and

$$\max\{p_i^I, p_i^S\} \geq 2^{-H(Z|Y_1 \cdots Y_{i-1})/2}.$$

Thus, some part of the secret key is used up in the same way by every message for preventing impersonation and substitution.

3.2.3 Privacy Amplification: Rényi Entropy

Privacy amplification is a key component of many unconditionally secure cryptographic protocols (see Section 2.6, Chapter 5, and [BBCM95]). Assume Alice and Bob share a random variable W, while an eavesdropper Eve knows a correlated random variable V that summarizes her knowledge about W. The details of the distribution P_{WV} are unknown to Alice and Bob because Eve can choose her eavesdropping strategy secretly. By communication over a public channel, which is totally susceptible to eavesdropping by Eve, Alice and Bob wish to agree on a compression function g such that Eve knows nearly nothing about $K = g(W)$. If Eve has arbitrarily small information about K, this value can be used as a cryptographic key for unconditionally secure encryption and authentication. Rényi entropy of order $\alpha > 1$ is important for defining Eve's admissible eavesdropping strategies in privacy amplification.

The need for privacy amplification shows up typically towards the end of a protocol with unconditional security, when a highly secret key should be generated from a large body of partially secret information. Why some information has leaked can have many different reasons, but it occurs in almost all such protocols. In quantum key distribution [BBB$^+$92, BC96], for example, the key bits are encoded in nonorthogonal states of a quantum system and an eavesdropper is prevented from extracting the complete information by the uncertainty principle of quantum mechanics. However, she can obtain partial information by specific measurements that disturb the quantum states only slightly more than the noise generated in the sender's or the receiver's equipment and are therefore not detected.

Leaking information is present also in the unconditionally secure key agreement protocols proposed by Maurer [Mau93, Mau94]. These protocols are based on the output of a randomizer, which is transmitted to Alice, Bob, and Eve over partially independent noisy channels that insert errors with certain probabilities. Alice and Bob must apply error-correction protocols in order to obtain the same values with high probability. Because error correction is done by communicating over a public channel, some information leaks to Eve.

Consider possible kinds of partial information V that Eve might have about an n-bit string W shared by Alice and Bob. V could consist of t physical bits of W, whose positions Alice and Bob do not know. Alternatively, Eve may have obtained t parities of bits of W or even the output of an arbitrary function $e : \{0,1\}^n \rightarrow \{0,1\}^t$ of her choice, but unknown to Alice and Bob, which is applied to W.

For these kinds of Eve's information, Bennett et al. [BBR88] showed that Alice and Bob can indeed extract about $n-t$ virtually secret bits using a function g that is randomly chosen from a special set of functions, called a 2-universal hash function. (A 2-universal hash function is a set $\mathcal{G}$ of functions $\mathcal{X} \rightarrow \mathcal{Y}$ such that for all distinct $x_1, x_2 \in \mathcal{X}$ when g is chosen uniformly from $\mathcal{G}$, the probability that $g(x_1) = g(x_2)$ is at most $1/|\mathcal{Y}|$, see Definition 2.1.)

Subsequent work by Bennett et al. shows that the same security can also be achieved with universal hash functions in the more general case when Eve's particular knowledge $V = v$ leaves her with enough Rényi entropy of order 2 about W [BBCM95]. Assume that Eve has observed a value $V = v$ such that she has Rényi entropy $H_2(W|V=v)$ about W, let G be the random variable corresponding to the random choice (with uniform distribution) of a member of a 2-universal hash function $\mathcal{G} : \{0,1\}^n \rightarrow \{0,1\}^l$, and let $K = G(W)$. Then

$$H(K|G, V=v) \;\geq\; l - 2^{l-H_2(W|V=v)}/\ln 2.$$

This implies that Eve has arbitrarily small information about K because if W is compressed to an l-bit key K with $l = n - t - s$ for some $s > 0$, her knowledge V and G about K satisfies

$$I(K;GV) \;\leq\; 2^{-s}/\ln 2. \tag{3.8}$$

Recently, we further generalized the restrictions on Eve's knowledge under which privacy amplification works to Rényi entropy of order α

for any $\alpha > 1$ [Cac97] (see Section 4.5.1). Assume that Eve's particular value $V = v$ leaves her with Rényi entropy $H_\alpha(W|V = v)$ about W for some $\alpha > 1$. Let $r, q, s > 0$, let m be an integer such that $m - \log(m+1) > n + q$, and let K be an l-bit key computed as above using a 2-universal hash function $\mathcal{G}$ with

$$l \;=\; H_\alpha(W|V = v) - \log(m+1) - \frac{r}{\alpha - 1} - q - 2 - s.$$

Then, there is an event $\mathcal{E}$ that has probability at least $1 - 2^{-r} - 2^{-q}$ such that

$$H(K|G, V = v, \mathcal{E}) \;\geq\; l - 2^{-s}/\ln 2.$$

This implies the same security of K as (3.8) except with probability $2^{-r} + 2^{-q}$. For $\alpha \to 1$, the term $\frac{r}{\alpha-1}$ becomes dominating, preventing Alice and Bob from extracting a secret key. (An assumption in terms of Shannon entropy, which is Rényi entropy of order 1, guarantees only a trivial amount of secrecy in a privacy amplification scenario, see example 4.4.)

3.2.4 Guessing Keys: Min-Entropy and "Guessing Entropy"

Consider the problem of guessing the value of a random variable X by asking only questions of the form "is X equal to x?" until the correct value is found and the answer is "yes." The study of this problem was initiated by Massey [Mas94]. Such situations occur, for example, in the cryptanalysis of computationally secure ciphers. Assuming that a cryptosystem is secure in the way intended by its designers, the only attack for finding the secret key is trying all possible keys in sequence for a given plaintext-ciphertext pair. Key guessing attacks are especially relevant for secret-key algorithms such as DES or IDEA, where a specialized technique like linear or differential cryptanalysis can be used to reduce the number of keys that must be tried [MvOV97].

Cryptanalysis of public-key systems, on the other hand, usually requires sophisticated mathematical methods because public-key systems are based on intractable problems with rich mathematical structure (e.g. factoring integers) and direct key guessing attacks are prohibitively inefficient. Therefore, the focus of cryptanalysis lies much more on the mathematical aspects of the problem, although brute-force searching of some large space is also often needed [Pom90, MvOV97].

The probability that the correct value is guessed in the *first* trial is directly linked to the *min-entropy* of X and is equal to $2^{-H_\infty(X)} = \max_{x\in\mathcal{X}} P_X(x)$ under an optimal strategy. An upper bound on this probability in terms of Shannon entropy is provided by the well-known *Fano inequality*, which gives a lower bound on the error probability of guessing X from knowledge of a correlated random variable Y [CT91]. The estimate $\widehat{X}$ for X is therefore a function of Y. The Fano inequality states that the error probability $p_e = \mathrm{P}[\widehat{X} \neq X]$ satisfies

$$h(p_e) + p_e \log(|\mathcal{X}| - 1) \;\geq\; H(X|Y). \tag{3.9}$$

The optimal strategy for successive guessing until the value of X is found is obviously to try the possible values in order of decreasing probability. Denote the elements of the probability distribution P_X by $p_1, \ldots, p_n$ such that $p_1 \geq p_2 \geq \cdots \geq p_n$ with $n = |\mathcal{X}|$. For a fixed optimal guessing strategy, let a *guessing function* for X be a function $G : \mathcal{X} \to \mathbb{N}$ such that $G(x)$ denotes the number of guesses needed when $X = x$. The average number of guesses needed to determine X,

$$\mathrm{E}[G(X)] \;=\; \sum_{i=1}^{n} i p_i,$$

can be called the *guessing entropy* of X. In the case of guessing X with knowledge of a correlated random variable Y, let $G(X|Y)$ be a *guessing function* for X given Y when $G(X|y)$ is a guessing function for the probability distribution $P_{X|Y=y}$ for any fixed $y \in \mathcal{Y}$. Thus,

$$\mathrm{E}[G(X|Y)] \;=\; \sum_{y\in\mathcal{Y}} P_Y(y)\mathrm{E}[G(X|y)]$$

can be called the *conditional guessing entropy* of X given Y.

Massey obtained the following lower bound on $\mathrm{E}[G(X)]$ in terms of the Shannon entropy of X:

$$\mathrm{E}[G(X)] \;\geq\; 2^{H(X)-2} + 1 \tag{3.10}$$

for any random variable X with $H(X) \geq 2$.

McEliece and Yu [MY95] showed that the guessing entropy provides a (weak) lower bound on the Shannon entropy of X,

$$H(X) \;\geq\; \frac{2\log|\mathcal{X}|}{|\mathcal{X}|-1}\big(\mathrm{E}[G(X)] - 1\big).$$

A connection between the guessing entropy and Rényi entropy of order α with $0 < \alpha \leq 1$ is given by Arikan [Ari96]. Let

$$\widehat{H}_\alpha(X|Y) = \frac{\alpha}{1-\alpha} \log \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} P_{XY}(x,y)^\alpha \right)^{1/\alpha}.$$

$\widehat{H}_\alpha(X|Y)$ would be another possible definition for the *conditional Rényi entropy* of order α of X given Y (see the discussion on page 16).

It was first observed by Arimoto [Ari77] that $\widehat{H}_\alpha(X|Y)$ satisfies $0 \leq \widehat{H}_\alpha(X|Y) \leq H_\alpha(X)$ for any $\alpha \geq 0$. Arikan's result provides a lower bound on the ρ-th moment of the guessing function G. For any $\rho \geq 0$,

$$\begin{aligned} \mathrm{E}[G(X|Y)^\rho] &\geq (1 + \ln |\mathcal{X}|)^{-\rho} \sum_{y \in \mathcal{Y}} \left(\sum_{x \in \mathcal{X}} P_{XY}(x,y)^{\frac{1}{1+\rho}} \right)^{1+\rho} \\ &= (1 + \ln |\mathcal{X}|)^{-\rho}\, 2^{\rho \widehat{H}_{\frac{1}{1+\rho}}(X|Y)}. \end{aligned} \tag{3.11}$$

3.2.5 Hash Functions: Collision Probability

Cryptographic hash functions play a fundamental role in modern cryptography, in particular for ensuring data integrity and message authentication [MvOV97]. Hash functions take a message of arbitrary length as input and produce a fixed-length output, called the *hash value*. Because the number of inputs exceeds the number of outputs, *collisions* occur when different inputs are mapped to the same output. The basic idea of cryptographic hash functions is that a hash value serves as a compact representation of the input and can be used as if it were uniquely identifiable with the longer message. In this section, we discuss two applications of the collision probability to cryptographic hash functions and to one-way functions.

One-way functions are similar to cryptographic hash functions, but have fixed input and output sizes that are equal in most cases. A function f is called *one-way* if it takes an argument x and efficiently produces a value $f(x)$ such that it is computationally infeasible, given only $y = f(x)$, to find any x' (including $x' = x$) such that $f(x') = y$.

Following the terminology of Menezes et al. [MvOV97, p. 323], a *cryptographic hash function* $h : \mathcal{X} \to \mathcal{Y}$ is a function that can be computed efficiently for any input x and that maps an input x of arbitrary length to an output $h(x)$ of fixed length. In addition, h must have one or more of the following properties:

Preimage Resistance: h is one-way as described above.

2nd-preimage Resistance: Given an input x, it is computationally infeasible to find any second input $x' \neq x$ such that $h(x) = h(x')$.

Collision Resistance: It is computationally infeasible to find any two distinct inputs x, x' that hash to the same output, i.e. such that $h(x) = h(x')$.

A *one-way hash function* satisfies the conditions for preimage resistance and 2nd-preimage resistance. One-way hash functions are often used in connection with password-based user authentication in computer systems (e.g. by the Unix operating system). For each user, the password file contains the value of the one-way function applied to the password (or a longer passphrase), but not the password itself. In this way, the password file must only be write-protected instead of read-protected for ordinary users and the passwords are nevertheless kept secret. To grant access to a user, the system computes the one-way function of the entered password and compares it to the stored entry for that user.

Assume the password X of a particular user is stored as $Y = f(X)$ using the one-way hash function f. An adversary can usually sample the probability distribution of X, either by assuming X to be uniformly distributed in lack of further knowledge or by assuming that X is chosen from a dictionary of words (possibly with certain preferences). With this knowledge, the adversary can randomly choose an $\widehat{X}$ with distribution P_X and compare $\widehat{Y} = f(\widehat{X})$ to Y. Success leads to the impersonation of that user and occurs with probability equal to the *collision probability* $P_2(Y)$ of Y. Moreover, the adversary can repeat the choice of $\widehat{X}$ and succeed independently with probability $P_2(Y)$ every time. It is therefore crucial that one-way hash functions do not produce collisions with probability substantially larger than $1/|\mathcal{Y}|$. (Incidentally, universal hash functions achieve this probability for any pair of distinct inputs over the random choice of the *function*—but they are not necessarily one-way.)

The second application of collision probability to hash functions involves the security of collision resistant cryptographic hash functions. These must be secure against so-called *birthday attacks*. Assume h is a collision resistant cryptographic hash function such that $Y = h(X)$ is uniformly distributed over $\mathcal{Y}$ for uniformly chosen X.

Finding a collision of h is equivalent to the well-known *birthday problem* [Fel68]: What is the probability that at least two people out of m have the same date of birth within any year, ignoring the year itself? (Birthdays are assumed to be uniformly distributed over the year.) The solution, applied to n independent repetitions of the random variable Y, implies that a collision occurs with probability

$$1 - \frac{|\mathcal{Y}|!}{(|\mathcal{Y}| - n)!\,|\mathcal{Y}|^n}$$

which can be approximated [Fel68] by

$$1 - e^{-\frac{n(n-1)}{2|\mathcal{Y}|}}.$$

Only about $\sqrt{|\mathcal{Y}|}$ applications of h are therefore needed to break the hash function by producing a collision. This implies that such hash functions must be designed with output size at least $2t$ bits if 2^t operations are considered infeasible.

However, an imperfect hash function may induce a nonuniform distribution of Y for uniformly distributed inputs. In this case, the probability of a collision in n independent repetitions of Y increases and is approximately

$$1 - \frac{|\mathcal{Y}|!}{(|\mathcal{Y}| - n)!\,|\mathcal{Y}|^n} + \frac{n!}{2|\mathcal{Y}|^{n-2}}\binom{|\mathcal{Y}|-2}{n-2}\left(P_2(Y) - \frac{1}{|\mathcal{Y}|}\right).$$

This connection to $P_2(Y)$ follows from the approximation of the birthday problem for nonuniform distributions given by Nunnikhoven [Nun92]. However, the effect of a slightly nonuniform output is rather weak.

3.2.6 Probabilistic Bounds: Variational Distance

In this section we compare different probabilistic statements to express that one "knows nothing" about the outcome of a random experiment. This situation is not restricted to cryptography, where such statements can be used for the uncertainty of an adversary about a secret. The variational distance between the distribution of the random variable and the uniform distribution over the same range is the adequate information measures for eliminating the probabilities from such bounds.

The L_1 *distance* between P_X and P_Y, defined as

$$\|P_X - P_Y\|_1 = \sum_{x \in \mathcal{X}} |P_X(x) - P_Y(x)|,$$

differs from the variational distance always by a factor of two; for any two distributions P_X and P_Y over the same alphabet $\mathcal{X}$, we have

$$\|P_X - P_Y\|_v \;=\; \frac{1}{2}\|P_X - P_Y\|_1. \tag{3.12}$$

In our abstract setting, we model the experiment by a random variable X with alphabet $\mathcal{X}$ and investigate bounds in terms of Shannon entropy and variational distance. The strongest notion of ignorance about X is to demand that the distribution of X is the uniform distribution P_U over $\mathcal{X}$, which coincides with the notion of ignorance underlying perfect secrecy (see Section 3.2.1). In this case, the entropy of X satisfies $H(X) = \log|\mathcal{X}|$ and the distance between P_X and P_U is zero, $\|P_X - P_U\|_v = 0$.

Weaker bounds allow some small a priori knowledge about X and are of the form $H(X) \geq \log|\mathcal{X}| - \delta$ or $\|P_X - P_U\|_v \leq \delta$ for some small parameter δ. (How a statement of the first kind can be converted into one of the second kind is described in Section 3.4.)

Even weaker bounds are probabilistic in the sense that an event $\mathcal{E}$ is assumed to exist, which occurs with high probability, and the uniformity bounds hold only when $\mathcal{E}$ occurs. For some small ϵ, assume that $\mathrm{P}[\mathcal{E}] \geq 1 - \epsilon$ and that $H(X|\mathcal{E}) \geq \log|\mathcal{X}| - \delta$ in terms of entropy or $\|P_{X|\mathcal{E}} - P_U\|_v \leq \delta$ in terms of variational distance. Such statements are of particular importance because results of this kind can be derived in many situations. However, when these probabilistic bounds are converted into bounds that hold with probability 1, the two variants behave quite differently.

Let the event $\mathcal{E}$ be induced by an error indicator random variable Z with alphabet $\{0,1\}$ such that $\mathcal{E}$ corresponds to $Z = 1$. From the bound in terms of entropy, we can only conclude that

$$\begin{aligned} H(X) &\geq H(X|Z) \\ &= P_Z(0)H(X|Z=0) + P_Z(1)H(X|Z=1) \\ &\geq (1-\epsilon)(\log|\mathcal{X}| - \delta). \end{aligned} \tag{3.13}$$

Consider the example of a $100'000$-bit string X, error probability $\epsilon = 0.001$, and uniformity $\delta = 0.1$. According to (3.13), $H(X) \gtrapprox 99900$, which leaves open the possibility that 100 bits of X are known in any case. This is clearly a weaker statement of the ignorance about X than the one given in the specification of the example.

For variational distance, however, $||P_{X|\mathcal{E}} - P_U||_v \leq \delta$ and $\mathrm{P}[\mathcal{E}] \geq 1-\epsilon$ together imply that $||P_X - P_U||_v \leq \epsilon + \delta$ with probability 1, as explicitly proved in Lemma 3.1. Using the numbers of the example, when $||P_{X|\mathcal{E}} - P_U||_v \leq 0.1$ and $\mathcal{E}$ has probability at least 0.999, it follows that $||P_X - P_U||_v \leq 0.101$.

The next lemma shows that in contrast to bounds in terms of entropy, probabilistic bounds in terms of the L_1 and variational distances can be converted easily to bounds not involving a probability without much security loss.

Lemma 3.1. *Let X and Y be independent random variables over the same alphabet $\mathcal{X}$, and let $\mathcal{E}$ be an arbitrary event. Assume that*

$$\mathrm{P}[\mathcal{E}] \;\geq\; 1-\epsilon \qquad \textit{and} \qquad ||P_{X|\mathcal{E}} - P_{Y|\mathcal{E}}||_v \;\leq\; \delta.$$

Then

$$||P_X - P_Y||_v \;\leq\; \epsilon + \delta.$$

Proof. Let Z be the error indicator random variable as defined above.

$$\begin{aligned}
||P_X - P_Y||_v \;&=\; \max_{\mathcal{S}\subseteq\mathcal{X}} \Big| \sum_{x\in\mathcal{S}} P_X(x) - \sum_{x\in\mathcal{S}} P_Y(x) \Big| \\
&=\; \max_{\mathcal{S}\subseteq\mathcal{X}} \Big| \sum_{\substack{x\in\mathcal{S}\\ z\in\{0,1\}}} P_{XZ}(x,z) - \sum_{\substack{x\in\mathcal{S}\\ z\in\{0,1\}}} P_{YZ}(x,z) \Big| \\
&=\; \max_{\mathcal{S}\subseteq\mathcal{X}} \Big| \sum_{x\in\mathcal{S}} P_{XZ}(x,0) + \sum_{x\in\mathcal{S}} P_{XZ}(x,1) - \\
&\qquad \sum_{x\in\mathcal{S}} P_{YZ}(x,0) - \sum_{x\in\mathcal{S}} P_{YZ}(x,1) \Big| \\
&\leq\; \max_{\mathcal{S}_0\subseteq\mathcal{X}} \Big| \sum_{x\in\mathcal{S}_0} P_{XZ}(x,0) - \sum_{x\in\mathcal{S}_0} P_{YZ}(x,1) \Big| \;+ \\
&\qquad \max_{\mathcal{S}_{\mathcal{E}}\subseteq\mathcal{X}} \Big| \sum_{x\in\mathcal{S}_{\mathcal{E}}} P_{XZ}(x,0) - \sum_{x\in\mathcal{S}_{\mathcal{E}}} P_{YZ}(x,1) \Big| \\
&=\; P_Z(0)\cdot ||P_{X|Z=0} - P_{Y|Z=0}||_v \;+ \\
&\qquad P_Z(1)\cdot ||P_{X|Z=1} - P_{Y|Z=1}||_v \\
&\leq\; \epsilon \;+\; \delta
\end{aligned}$$

The first inequality follows from the triangle inequality and the second inequality from the assumption of the lemma and the fact that the variational distance between two distributions is at most 1. □

Remark. The statement of the lemma is equivalent to the following. Let X and Y be random variables over the same alphabet $\mathcal{X}$. Assume that a random variable Z exists that takes on a value z for which $\|P_{X|Z=z} - P_{Y|Z=z}\|_v \leq \delta$ with probability at least $1 - \epsilon$. Then $\|P_X - P_Y\|_v \leq \epsilon + \delta$.

3.3 Some Relations Between Information Measures

In this section we present a collection of inequalities that relate the entropy measures discussed in the preceding section. In all statements of this section, X and Y denote random variables with the same alphabet $\mathcal{X}$ and P_U denotes the uniform distribution over $\mathcal{X}$.

Table 3.1 (page 44) provides a systematic overview of relations between the entropy measures. Bounds in terms of the distance measures, $D(P_X\|P_Y)$ and L_1 distance, are included in the table for the distances between P_X and P_U. However, some of the results are more general and hold for the distances between arbitrary distributions P_X and P_Y. Variational distance is not included because it is equivalent to one-half of the L_1 distance.

Lemma 3.2. *The Rényi entropy $H_\alpha(X)$ for any $\alpha > 1$, the Shannon entropy $H(X)$, the Rényi entropy $H_\beta(X)$ for any β with $0 \leq \beta < 1$, and the min-entropy $H_\infty(X)$ satisfy*

$$0 \overset{(a)}{\leq} H_\infty(X) \overset{(b)}{\leq} H_\alpha(X) \overset{(c)}{\leq} H(X) \overset{(d)}{\leq} H_\beta(X) \overset{(e)}{\leq} \log|\mathcal{X}|. \qquad (3.14)$$

Equality in (a)–(d) holds if and only if $P_X(x) = 1$ for some $x \in \mathcal{X}$; equality in (b)–(e) holds if and only if $P_X(x) = 1/|\mathcal{X}|$ for all $x \in \mathcal{X}$; equality in (b)–(d) holds if and only if for some set $\mathcal{X}_0 \subseteq \mathcal{X}$, $P_X(x) = 1/|\mathcal{X}|_0$ for all $x \in \mathcal{X}_0$.

Proof. The lemma follows from Propositions 2.1 and 2.4. □

Lemma 3.3 ([CT91]).

$$D(P_X\|P_Y) \;\geq\; \frac{1}{2\ln 2}\|P_X - P_Y\|_1{}^2$$

Lemma 3.4.

$$H(X) \;\leq\; \log|\mathcal{X}| - \frac{1}{2\ln 2}\|P_X - P_U\|_1{}^2$$

Proof. Combine (2.12) and Lemma 3.3. □

Lemma 3.5 ([CT91]). *If* $||P_X - P_Y||_1 \le \frac{1}{2}$, *then*

$$|H(X) - H(Y)| \;\le\; -||P_X - P_Y||_1 \log \frac{||P_X - P_Y||_1}{|\mathcal{X}|}.$$

Lemma 3.6. *If* $||P_X - P_U||_1 \le \frac{1}{2}$, *then*

$$H(X) \;\ge\; \log|\mathcal{X}| + ||P_X - P_U||_1 \log \frac{||P_X - P_U||_1}{|\mathcal{X}|}.$$

Proof. Combine (2.12) and Lemma 3.5. □

The L_2 *distance* between P_X and P_Y is defined as

$$||P_X - P_Y||_2 \;=\; \sqrt{\sum_{x\in\mathcal{X}} \big(P_X(x) - P_Y(x)\big)^2}.$$

Lemma 3.7.

$$H_2(X) \;=\; -\log\Big(\frac{1}{|\mathcal{X}|} + ||P_X - P_U||_2{}^2\Big)$$

Proof. For all $x \in \mathcal{X}$, let $\Delta_x = P_X(x) - 1/|\mathcal{X}|$.

$$\begin{aligned} H_2(X) &= -\log \sum_{x\in\mathcal{X}} P_X(x)^2 \\ &= -\log \sum_{x\in\mathcal{X}} \Big(\frac{1}{|\mathcal{X}|} + \Delta_x\Big)^2 \\ &= -\log\Big(\frac{1}{|\mathcal{X}|} + \sum_{x\in\mathcal{X}} \Delta_x^2\Big) \end{aligned}$$

The Lemma follows by noting that $\sum_{x\in\mathcal{X}} \Delta_x^2 = ||P_X - P_U||_2{}^2$. □

	$D(\cdot)$	H_α	H_∞	$\mathrm{E}[G]$	$\lVert\cdot\rVert_1$
H	(2.12)	L3.2	(3.9),L3.2	(3.10)	L3.4,L3.5,L3.6
$D(\cdot)$	–	(2.12)–L3.2	(2.12)–L3.2	(2.12)–(3.10)	L3.3,L3.5
H_α	–	–	L3.2,L4.16	(3.11)	L3.7
H_∞	–	–	–	L3.2–(3.10)	–

Table 3.1. Overview of inequalities linking different entropy measures of a random variable. The entries for the distance measures ($D(\cdot)$ and $\lVert\cdot\rVert_1$) refer to the distance to the uniform distribution. "L" stands for Lemma, a comma for alternatives, and a dash as in "(1)–(2)" for first applying the bound (1) and then (2).

3.4 Shannon Entropy and Almost Uniform Distributions

Although Shannon entropy is the central concept of information theory for expressing the uncertainty about a random variable, other uncertainty measures are appropriate in some specific cryptographic scenarios, as Section 3.2 demonstrates. Nevertheless, statements in terms of Shannon entropy are the conventional way to guarantee the security of unconditionally secure cryptosystems. Such statements are typically of the form that Eve's entropy about a secret key or about the communicated plaintext is negligibly close to maximal and thus, the corresponding probability distribution is close to uniform. The purpose of this section is to justify this convention by investigating the connections between Shannon entropy and cryptographically relevant uncertainty measures for distributions that are close to uniform.

Basically, an entropy measure for a random variable X quantifies some aspect of a guessing game to determine some property of X, e.g. the number of questions that must be asked until the value of X is known. We can distinguish entropy measures by the restrictions for the kind of guessing questions allowed, by the property to be determined, and by the definition of success:

- The Shannon entropy $H(X)$ quantifies the minimum number of binary questions required on the average to determine the value of X, where the average is taken over independent repetitions of X.

- The min-entropy $H_\infty(X)$ is the negative logarithm of the probability that X is determined correctly with only one guess of the form "is X equal to x?" with an optimal strategy.

- The guessing entropy $\mathrm{E}[G(X)]$ quantifies the average number of questions of the form "is X equal to x?" until the value of X is determined, where the average is taken over X.

Other cryptographically relevant definitions for success of the guessing game might allow that the correct value of X must be determined only with non-negligible probability. The guessing scenarios of Shannon entropy or guessing entropy can be modified accordingly. However, we are not aware of any proposals for formalizations of the corresponding entropy measures.

We now examine the difficulty of the guessing games mentioned above for distributions that are close to uniform. Assume that $H(X) \geq \log|\mathcal{X}| - \delta$ for some small positive δ. Following the interpretation of Shannon entropy, almost $\log|\mathcal{X}|$ arbitrary binary questions are needed on the average to determine the value of X.

Let $p_{\max} = \max_{x\in\mathcal{X}} P_X(x)$. The probability that X is determined on the first guess with an optimal strategy is equal to $2^{-H_\infty(X)} = p_{\max}$ and can be upper bounded using the Fano inequality (3.9), which states that $p_{\max}$ must satisfy

$$h(1-p_{\max}) + (1-p_{\max})\log(|\mathcal{X}|-1) \;\geq\; \log|\mathcal{X}| - \delta.$$

It can be verified easily that $p_{\max}$ must be close to $1/|\mathcal{X}|$ for small δ and that only $p_{\max} = 1/|\mathcal{X}|$ is possible with $\delta = 0$.

For the expected number of questions in the sense of guessing entropy, Massey's result (3.10) implies

$$\mathrm{E}[G(X)] \;\geq\; 2^{\log|\mathcal{X}|-2-\delta} + 1 \;\geq\; \frac{1}{4}2^{-\delta}|\mathcal{X}| + 1,$$

which is only a factor of two below $\frac{1}{2}|\mathcal{X}|$, the maximum average number of questions that are needed when P_X is uniform.

The distribution of a random variable X with Shannon entropy almost maximal is also close to the uniform distribution P_U over $\mathcal{X}$ in terms of variational distance. Namely, it follows from $H(X) \geq \log|\mathcal{X}| - \delta$ by Lemma 3.4 that

$$\|P_X - P_U\|_v \;\leq\; 2\ln 2\,\sqrt{\delta}.$$

In summary, the preceding discussion shows that statements of the form $H(X) \geq \log|\mathcal{X}| - \delta$ for some small δ are sufficient to guarantee the security of a cryptographic key X because they can be transformed into tight lower bounds in terms of other entropy measures, which are more relevant in cryptographic scenarios.

Chapter 4

Smooth Entropy

The subject of this chapter is "entropy smoothing," which is the process of converting an arbitrary random source into a source with smaller alphabet and almost uniform distribution. We introduce a formal definition of smooth entropy to quantify the number of uniform bits that can be extracted from a random source by probabilistic algorithms. The main question is: Given an arbitrary random source, how many uniformly random bits can be extracted? We allow for an arbitrarily small deviation of the output bits from perfectly uniform random bits that may include a small correlation with the random bits used for smoothing. The inclusion of randomized extraction functions is the main difference between entropy smoothing and "pure" random number generation in information theory, where no additional random sources are available. However, entropy smoothing does not consider the auxiliary random bits as a resource, unlike extractors used in theoretical computer science.

Entropy smoothing has many applications in theoretical computer science and in cryptography. The fact that different parties involved in some scenario have different knowledge motivates the distinction between multiple random sources. In many applications, the smoothing algorithm should not depend on the distribution of the source and must work for all sources with a certain structural property. Therefore, our smoothing algorithms should not depend on the probability distribution of a source.

The outline of the chapter is as follows. After some introductory considerations and motivating examples, a general framework for entropy smoothing and the definition of smooth entropy are introduced in Section 4.2. The relation to previous work and other concepts such as privacy amplification, data compression, random number generation, and extractors are discussed in Section 4.3. Among other things, we observe that smooth entropy is lower bounded by Rényi entropy of order 2 and that average smooth entropy corresponds to Shannon entropy. We then turn to bounds on smooth entropy using the "spoiling knowledge" proof technique that is investigated in Section 4.4. In Section 4.5, we prove lower bounds on smooth entropy in terms of Rényi entropy and the profile of the distribution. In particular, we show that smooth entropy is lower bounded by Rényi entropy of order α for any $\alpha > 1$. Finally, in Section 4.6, we prove a generalized version of the main theorem of entropy smoothing that extends to the case of smoothing a random variable with unknown distribution. This chapter is partially based on [CM97b, Cac97].

4.1 Introduction

Consider a random variable X with alphabet $\mathcal{X}$ and distribution P_X. We want to apply a *smoothing function* $f : \mathcal{X} \to \mathcal{Y}$ to X such that $Y = f(X)$ is uniformly distributed over its range $\mathcal{Y}$. The size of the largest $\mathcal{Y}$ such that Y is still sufficiently uniform is a measure for the amount of *smooth entropy* inherent in X or extractable from X, relative to the allowed deviation from perfect uniformity. In an alternative view, this process eliminates almost all partial information available about X and concentrates as much as possible of the uncertainty of X in Y. For these reasons, f is also called an *extraction function* or *concentration function*.

Before we proceed, we need a way to quantify the remaining "nonrandomness" in Y, that is, its divergence from a perfectly uniform distribution. We do not fix any particular information measure—rather, any information measure, such as those discussed in Chapter 3, can be used as long as it is a measure of nonuniformity. For the moment, we will use the relative entropy between X and the uniform distribution P_U over $\mathcal{X}$,

$$D(P_X \| P_U) = \log |\mathcal{X}| - H(X)$$

and the L_1 distance from the uniform distribution,

$$\|P_X - P_U\|_1 = \sum_{x \in \mathcal{X}} |P_X(x) - P_U(x)|.$$

s	y	1	2	3	4	5	6
6	$P_{Y_6}(y)$	0.21	0.18	0.17	0.16	0.14	0.14
	$f_6^{-1}(y)$	a	b	c	d	e	f, g, h, i
5	$P_{Y_5}(y)$	0.22	0.21	0.19	0.19	0.19	
	$f_5^{-1}(y)$	e, f	a	b, i	c, h	d, g	
4	$P_{Y_4}(y)$	0.3	0.25	0.24	0.21		
	$f_4^{-1}(y)$	d, e	c, f	a, g	b, h, i		
3	$P_{Y_3}(y)$	0.34	0.33	0.33			
	$f_3^{-1}(y)$	b, e, h	c, d	a, f, g, i			
2	$P_{Y_2}(y)$	0.5	0.5				
	$f_2^{-1}(y)$	a, b, f, g	c, d, e, h, i				

Table 4.1. The table shows how the random variable X with nine values in Example 4.1 can be converted to more uniform random variables $Y_s \in \{1, \ldots, s\}$ with $s = 6, \ldots, 2$ values by the deterministic functions $f_6, \ldots, f_2$ (shown only implicitly).

Example 4.1. Suppose we want to produce an almost uniform Y from a random variable X with nine values and the following distribution

x	a	b	c	d	e	f	g	h	i
$P_X(x)$	0.21	0.18	0.17	0.16	0.14	0.08	0.03	0.02	0.01

The best possible concentration functions f_s for the output sizes $s = 6, \ldots, 2$ and the resulting outputs $Y_s = f_s(X)$ are shown in Table 4.1. For all output sizes, the optimal functions turn out to be the same with respect to both nonuniformity measures. The output deviations from a perfectly uniform distribution P_U over the corresponding alphabet in terms of relative entropy and L_1 distance are

s	6	5	4	3	2
$D(P_{Y_s} \| P_U)$	0.0150	0.00285	0.0119	0.000144	0
$\|P_{Y_s} - P_U\|_1$	0.12	0.06	0.1	0.0133	0

We observe that, in general, the smaller the alphabet of the output, the more uniform is its distribution. But note also that Y_4 is an exception to this rule because $D(P_{Y_4}||P_U) > D(P_{Y_5}||P_U)$ and $||P_{Y_4} - P_U||_1 > ||P_{Y_5} - P_U||_1$. A perfectly uniform 1-bit random variable or a 0.1-uniform 2-bit random variable (in L_1 distance) can be produced from X. ○

In general, we expect that the output can be made more uniform by decreasing its size. But this trade-off between the size of the output and its uniformity is not strict, as Y_4 in Example 4.1 shows.

The example hints also at the difficulty of finding the best smoothing function for a given distribution. In fact, it is easy to see that Bin Packing, which is **NP**-complete [Pap94], can be reduced to our problem. Bin Packing is the problem of deciding whether n positive integers (or items) can be partitioned into b subsets (or bins) such that the sum of every subset is at most c.

So far, only fixed-length outputs have been considered. Could a function with variable-length output that depends on the input achieve better uniformity?

Example 4.2. Consider the random variable X with distribution

x	a	b	c	d	e
$P_X(x)$	$1/3$	$1/6$	$1/6$	$1/6$	$1/6$

It is possible to generate always one perfectly random bit by grouping together a, b and c, d, e. On the other hand, one can produce two uniformly random bits with probability $\frac{2}{3}$ if $X = a$ is ignored (mapped to an empty output) and the other X are mapped one-to-one to the output. In this manner, $1\frac{1}{3}$ uniformly random bits result on the average. ○

Although more random bits can be extracted in the variable-length case, we do not pursue this further. The reason is that in the intended applications, the exact source distributions are generally not known to the party that chooses the extraction function. Instead, randomized smoothing algorithms are allowed and sometimes needed for entropy smoothing. However, the additional randomness used by the smoothing function must be independent of X and must be taken into account for calculating the uniformity of the output. Any suitable randomized smoothing function can be modeled as a family of deterministic functions, one of which is selected at random and then applied to X.

Example 4.3. Consider a random variable X distributed as in Example 4.2 except that the correspondence between values and probabilities is not known. For every deterministic function that outputs two bits, it is possible that one of the output values has probability $\frac{1}{2}$ (when $\frac{1}{3}$ and $\frac{1}{6}$

are mapped together). The resulting output Y has $D(P_Y \| P_U) = 0.208$. But consider the ten functions that map two inputs together and pass the rest one-to-one to the output. If one of them is selected with uniform distribution, the expected relative entropy distance to the uniform distribution drops to 0.132. ○

We do not count the amount of additional randomness used to choose one function from the family in our model. This is the primary difference between entropy smoothing and extractors, which are used in complexity theory (see Section 4.3.6).

As Example 4.3 shows, the smoothing problem is more involved when the distribution is not available. The focus of our work will be on smoothing sources with certain structural characteristics. We need universal smoothing algorithms that work for any member of a family of random variables sharing some property, such as a bound on the maximal probability of one value.

An algorithm of this kind is *universal hashing*. Originally introduced for storing keys in a table by Carter and Wegman [CW79], universal hash functions have been used for entropy smoothing in cryptography [BBR86] and in theoretical computer science [ILL89].

Many applications of entropy smoothing have since shown up in such diverse areas as unconditionally secure cryptographic protocols [Mau93], quantum cryptography [BBB$^+$92], pseudorandom generation [HILL91, Lub96], derandomization of algorithms [LW95], computational learning theory [KV94], computing with weak random sources [Zuc91], and numerous other areas of complexity theory dealing with probabilistic computations [Nis96]. Some of these applications will be discussed in Section 4.3 after the formalization of smooth entropy.

4.2 A General Formulation

The two main questions to be addressed by a formal notion of smooth entropy are:

1. How many random bits with a certain uniformity can be extracted with high probability from a given random variable?

2. Is there a structural characterization of all random variables from which a certain amount of smooth entropy can be extracted with high probability?

The formalization should respect all the aspects mentioned in the previous section such as

- different measures of uniformity for the output,
- producing more uniform outputs by reducing the output size,
- probabilistic smoothing functions, and
- a small probability of failure.

To measure the uniformity of the output, we need a way to quantify the distance between a random variable X and the uniform distribution over the corresponding alphabet. A *nonuniformity measure* in this sense should reflect some intuitive notion of distance between P_X and the uniform distribution over $\mathcal{X}$. Let M be a nonuniformity measure that associates with every random variable X a positive number $M(X)$ that is 0 if and only if P_X is the uniform distribution over $\mathcal{X}$. Further, $M(X')$ should be strictly greater than $M(X)$ for every random variable X' that is strictly less uniform than X in the sense that $P_{X'}(x) > P_X(x) \geq \frac{1}{|\mathcal{X}|}$ or $P_{X'}(x) < P_X(x) \leq \frac{1}{|\mathcal{X}|}$ holds for all $x \in \mathcal{X}$.

Given a nonuniformity measure $M(X)$, we define the expected nonuniformity of X given a random variable Y as

$$M(X|Y) \;=\; \sum_{y \in \mathcal{Y}} P_Y(y) M(X|Y=y)$$

where $M(X|Y=y)$ denotes the nonuniformity of the random variable with probability distribution $P_{X|Y=y}$.

Two nonuniformity measures that will be used in the sequel are relative entropy and L_1 distance to the uniform distribution. Other choices for M could be the maximum probability of any value of X, expressed using min-entropy as $\log|\mathcal{X}| - H_\infty(X)$, the variational distance, or the d-bar distance [GNS75]. However, these measures will not be studied systematically.

The smoothing algorithm should be able to produce outputs that achieve some desired uniformity. As demonstrated in Example 4.1, more uniform outputs can usually be obtained by reducing the output size. However, for a fixed input distribution, it is not possible to produce arbitrarily uniform outputs. We introduce the parameter s to control the trade-off between the uniformity of the output and the amount of entropy lost in the smoothing process.

We formalize probabilistic concentration functions by extending the input of f with an additional random variable T that models the random choices of f (any randomized algorithm can be converted to this form). However, T must be independent of X and its value must be taken into account when computing the uniformity of Y. One can think of T as a catalyst random input that enables the extraction of uniform entropy from a distribution.

It can be tolerated that the uniformity bound for an extraction process fails if an error event $\mathcal{E}$ occurs. $\mathcal{E}$ should have small probability, denoted by ϵ, and may depend on X. The uniformity is calculated only in the case that the complementary event $\overline{\mathcal{E}}$ occurs.

We explicitly ignore the size of the public random input T in our considerations because we assume that additional randomness is available in abundance.

We are now ready for a definition of smooth entropy that quantifies the number of almost uniform random bits in a specific random variable. It incorporates all of the above-mentioned aspects, except for the universality of the smoothing process.

Definition 4.1. Let M be a nonuniformity measure and let $\Delta : \mathbb{R} \to \mathbb{R}$ be a decreasing non-negative function. A random variable X with alphabet $\mathcal{X}$ has *smooth entropy* $\Psi(X)$ *within* $\Delta(s)$ *[in terms of* M*] with probability* $1 - \epsilon$ if $\Psi(X)$ is the maximum of all ψ such that for any security parameter $s \geq 0$, there exist a random variable T and a function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ with $|\mathcal{Y}| = \lfloor 2^{\psi - s} \rfloor$ such that there is a failure event $\mathcal{E}$ that has probability at most ϵ, and the expected value over T of the nonuniformity M of $Y = f(X, T)$, given T and $\overline{\mathcal{E}}$, is at most $\Delta(s)$. Formally,

$$\begin{aligned} \Psi(X) \;=\; \max_{\psi} \Big\{ \psi \Big| & \forall s \geq 0 \,:\, \exists T, f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}, |\mathcal{Y}| = \lfloor 2^{\psi - s} \rfloor \,: \\ & Y = f(X,T)\,,\, \exists \mathcal{E} \,:\, \mathrm{P}[\mathcal{E}] \leq \epsilon\,,\, M(Y|T\overline{\mathcal{E}}) \,\leq\, \Delta(s) \Big\}. \end{aligned} \tag{4.1}$$

The definition is illustrated in Figure 4.1. Some remarks on technical aspects of the definition are due:

1. The definition is stated as an average over the choices of T, which allows $M(Y|T = t)$ for some t to exceed $\Delta(s)$. But the probability that such t occur can be controlled by increasing s and such t can therefore be eliminated at the rate at which $\Delta(s)$ converges to 0.

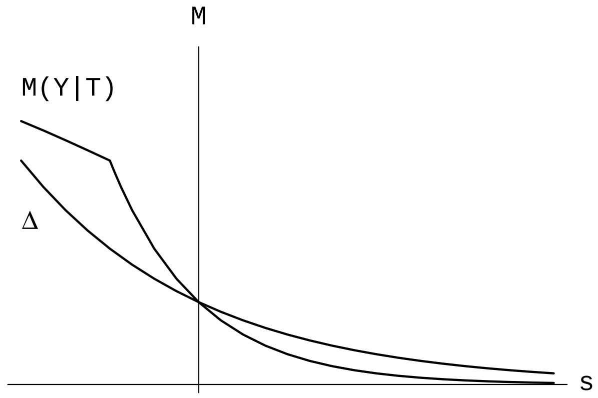

Figure 4.1. $M(Y|T)$ is the expected nonuniformity M of the output Y, given T, with an alphabet of $\lfloor 2^{\psi - s} \rfloor$ values. The definition of smooth entropy requires that $M(Y|T)$ is bounded from above by $\Delta(s)$ for all $s \geq 0$, but not for $s < 0$.

2. As noted above, the definition of smooth entropy ignores the size of T and the amount of public randomness needed. For practical applications, the smoothing function has to be polynomial-time computable and this limits the size of $|\mathcal{T}|$. However, this is no restriction in the case of smoothing with universal hash functions (see the proof of Corollary 4.1).

3. The notation $\Psi(X)$ is not explicit because it omits specification of the information measure M, the smoothness function $\Delta(s)$, and the failure probability ϵ. These parameters have to be mentioned separately if they are not clear from the context (which is usually the case).

The above definition does not have the desired universality of allowing the same smoothing function for different distributions. In many applications, it is known only that the random variable X has some property that is shared by many others. The smooth entropy of a family of random variables $\mathbb{X}$ as defined next requires that the same smoothing function works for all random variables in the family.

Definition 4.2. Let M be a nonuniformity measure and let $\Delta : \mathbb{R} \rightarrow \mathbb{R}$ be a decreasing non-negative function. A family $\mathbb{X}$ of random variables with alphabet $\mathcal{X}$ has *smooth entropy* $\Psi(\mathbb{X})$ *within* $\Delta(s)$ *[in terms of* M*] with probability* $1-\epsilon$ if $\Psi(\mathbb{X})$ is the maximum of all ψ such that for any security parameter $s \geq 0$, there exist a random variable T and a function $f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ with $|\mathcal{Y}| = \lfloor 2^{\psi - s} \rfloor$ such that for all $X \in \mathbb{X}$ there is a failure event $\mathcal{E}$ that has probability at most ϵ, and the expected value over T of the nonuniformity M of $Y = f(X,T)$, given T and $\overline{\mathcal{E}}$, is at most $\Delta(s)$. Formally,

$$\begin{aligned}\Psi(\mathbb{X}) \;=\; \max_{\psi}\Big\{\psi\,\Big|\,&\forall s \geq 0 \;:\; \exists T, f : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}, |\mathcal{Y}| = \lfloor 2^{\psi - s} \rfloor \;: \\ &\forall X \in \mathbb{X} \;:\; Y = f(X,T)\,,\, \exists \mathcal{E} \;:\; \mathrm{P}[\mathcal{E}] \leq \epsilon\,,\; M(Y|T\overline{\mathcal{E}}) \;\leq\; \Delta(s)\Big\}.\end{aligned}$$

For singleton sets $\{X\}$, we can use $\Psi(X)$ instead of $\Psi(\{X\})$ equivalently. The distance bound in the definition of smooth entropy is calculated under the condition that $\mathcal{E}$ does not occur, which has probability at least $1-\epsilon$. This probability cannot be eliminated for general information measures M because Y is only required to be close to an output which always satisfies $M(Y|T) \leq \Delta(s)$ in (4.1).

However, this deviation can be integrated into the uniformity parameter $\Delta(s)$ for certain nonuniformity measures, such as L_1 distance or variational distance. It is easy to see that a random variable X with smooth entropy $\Psi(X)$ within $\Delta(s)$ in terms of variational distance with probability $1-\epsilon$ has smooth entropy $\Psi(X)$ within $\Delta(s)+\epsilon$ in terms of variational distance with probability 1 (see Lemma 3.1). A similar relation holds for L_1 distance.

The distinction between nonuniformity measures to quantify the deviation is not central to our treatment. We will mainly use relative entropy distance. By Lemmas 3.3 and 3.5, statements about smooth entropy in terms of relative entropy can easily be converted to and from L_1 distance or variational distance. This shows that the properties of smooth entropy are—at least to some extent—independent of the nonuniformity measure used.

4.3 Previous Work and Related Concepts

We next review some of the history and applications of entropy smoothing in cryptography and in theoretical computer science. We point out the relation to other concepts from information theory, such as data compression and intrinsic randomness, and to extractors as used in the context of derandomization and computing with weak random sources.

Privacy amplification and entropy smoothing have been introduced independently and have been known for some time. Both techniques build on the fact that uniform entropy can be extracted with *universal hash functions* (see Section 2.6). Universal hash functions were first used in the present context by Bennett, Brassard, and Robert [BBR86] and by Impagliazzo, Levin, and Luby [ILL89], who also recognized the importance of Rényi entropy for entropy smoothing. An overview of some applications is given in the following subsections.

4.3.1 Privacy Amplification in Cryptography

Privacy amplification is a key component of many unconditionally secure cryptographic protocols and is introduced in Section 2.6. An overview of some unconditionally secure cryptographic systems is given in Chapter 5, where references are provided.

The Privacy Amplification Theorem (Theorem 2.7) has found many applications in cryptography. Often, privacy amplification forms the final step in a protocol to generate a secret key. This eliminates information that has leaked to an opponent during the protocol. Examples of this are quantum key distribution [BBB+92, BC96] and key agreement from common information [Mau93, Mau94]. Privacy amplification is also used to implement the cryptographic primitive oblivious transfer with unconditional security [BC97, Cré97].

Theorem 2.7 implies that the Rényi entropy of order 2 of a random variable X is a lower bound for its smooth entropy. It is crucial that smoothing by universal hashing does not require knowledge of the distribution—X may be any random variable with sufficient Rényi entropy. The same smoothing algorithm can be applied to any X from a family $\mathbb{X}$ of random variables and produce an output of the desired size and uniformity. The random variable T is used to select a member of a universal hash function with uniform distribution. This is stated as a corollary to the Privacy Amplification Theorem.

Corollary 4.1. *The smooth entropy of a family $\mathbb{X}$ of random variables within $2^{-s}/\ln 2$ in terms of relative entropy with probability 1 is at least the minimum Rényi entropy of order 2 of any $X \in \mathbb{X}$:*

$$\Psi(\mathbb{X}) \;\geq\; \min_{X \in \mathbb{X}} H_2(X).$$

Proof. For any $\mathcal{X}$ and for any $\mathcal{Y}$, the following set of functions is a universal hash function. Choose a prime $p \geq |\mathcal{X}|$. There exists a universal hash function from $\mathbb{Z}_p = \{0, \ldots, p-1\}$ to $\{0, \ldots, |\mathcal{Y}|-1\}$ that consists of the functions

$$h_{a,b}(x) = (ax + b \bmod p) \bmod |\mathcal{Y}|$$

for $a, b \in \mathbb{Z}_p$ with $a \neq 0$ [MR95]. Because there are sufficiently many primes, p can be chosen such that $p = O(|\mathcal{X}|)$ and a member of the universal family can be selected with $O(\log|\mathcal{X}|)$ random bits.

Thus, provided that $\psi \leq H_2(X)$, there exists for every s a suitable universal hash function $\mathcal{G}$ which is selected by $T = G \in \mathcal{G}$ with uniform distribution. Inserting $\log|\mathcal{Y}| = \psi - s$ in (2.25) shows that $H(Y|T) \geq \log|\mathcal{Y}| - 2^{-s}/\ln 2$ for any X with $H_2(X) \geq \psi$. □

Another well-known universal hash function that operates on binary strings is described in Section 5.4.2 on page 114. In addition, the set of all surjective functions from $\mathcal{X}$ to $\mathcal{Y}$ is 2-universal [Sar80] for any $\mathcal{X}$ and any $\mathcal{Y}$.

The role of smooth entropy in privacy amplification is to characterize the probability distributions that an adversary Eve may know. If it can be argued or assumed that Eve's knowledge about a value W known to Alice and Bob is a random variable X with smooth entropy at least $\Psi(X)$, then Alice and Bob can, only by public discussion, extract $\Psi(X)$ bits from W that are guaranteed to be almost completely hidden from Eve.

4.3.2 Entropy Smoothing in Pseudorandom Generation

Random bits are a ubiquitous and valuable resource in computer science [Lub96, MR95]. A *pseudorandom generator* is a deterministic polynomial-time computable algorithm A that, upon input of a short random seed X, produces a much longer string $A(X)$ that looks random to any polynomial-time bounded observer of $A(X)$. Thus, a pseudorandom generator can effectively convert a small amount of true randomness into a much larger amount of pseudorandomness that cannot be distinguished from a truly random string by any polynomial-time adversary.

Cryptographically secure pseudorandom generators are based on the concept of a one-way function, one of the most important concepts of modern cryptography. Such a function takes an argument x and efficiently produces a value $f(x)$ such that it is computationally infeasible for an adversary, given $y = f(x)$, to find any x' (including $x' = x$) such that $f(x') = y$.

Håstad, Impagliazzo, Levin, and Luby [HILL91] show how to construct a pseudorandom generator from any one-way function f. Their construction uses one iteration of f that generates somewhat pseudorandom, but not uniformly distributed bits. These bits are then converted into almost uniform random bits using a universal hash function. The following theorem, which is very similar to Theorem 2.7 and first appeared in [ILL89], guarantees that the output is almost uniform [HILL91].

Theorem 4.2. *Let m be a positive integer and let X be a random variable with alphabet $\{0,1\}^n$ such that $H_2(X) \geq m$. Let $e > 0$ be a positive integer parameter, let G be the random variable corresponding to the random choice (with uniform distribution) of a member of a universal hash function $\mathcal{G} : \{0,1\}^n \rightarrow \{0,1\}^{m-2e}$, let $Y = G(X)$, and let U be uniformly distributed over $\{0,1\}^{m-2e}$. Then*

$$||P_{YG} - P_{UG}||_1 \ \leq \ 2^{-e}.$$

We obtain the following lower bound on smooth entropy. The proof is similar to Corollary 4.1.

Corollary 4.3. *The smooth entropy of a family $\mathbb{X}$ of random variables within $2^{-s/2}$ in terms of L_1 distance with probability 1 is at least the minimum Rényi entropy of order 2 of any $X \in \mathbb{X}$:*

$$\Psi(\mathbb{X}) \ \geq \ \min_{X \in \mathbb{X}} H_2(X).$$

We note that Corollary 4.3 follows from Corollary 4.1 within a factor of $\sqrt{2}$, that is, for the smooth entropy within $2^{-s/2}\sqrt{2}$, by Lemma 3.3: $2^{-s}/\ln 2 \geq D(P_Y\|P_U) \geq \frac{1}{2\ln 2}\|P_Y - P_U\|_1{}^2$ and therefore $2^{-s/2}\sqrt{2} \geq \|P_Y - P_U\|_1$.

4.3.3 Relation to Entropy

Information theory demonstrates that a fundamental measure for randomness is entropy. A random variable X is about as random as a uniformly distributed bit string of length $H(X)$. This is shown by constructing an optimal prefix-free code for X that achieves average length between $H(X)$ and $H(X)+1$, where the average is taken over independent repetitions of X. Such a code is optimal because no prefix-free code can have expected length less than $H(X)$, where the expectation is over the choice of X.

So if the smooth entropy $\Psi(X)$ denotes the number of almost uniform random bits in X, what is the difference between $H(X)$ and $\Psi(X)$? The important distinction is that entropy denotes the average length of the optimal prefix-free code per instance of X (which is a variable-length code for X or a block code for a very large number of independent versions of X) whereas smooth entropy corresponds to the length of a fixed-length code for the single random variable X.

Average values are observed in general when a random experiment is repeated many times independently (by the effects of the law of large numbers). Therefore one expects intuitively that the average smooth entropy corresponds to entropy in the sense that

- entropy is an upper bound for smooth entropy and
- entropy is a lower bound for average smooth entropy.

These two bounds are presented formally in the remainder of this section. (A theorem similar to the lower bound is used in pseudorandom generation [Lub96, Shannon-to-Rényi-Theorem].)

The upper bound on smooth entropy is a simple consequence of the information-theoretic principle that processing cannot increase entropy.

Theorem 4.4. *Let $\Delta(s)$ be any non-negative function of s. Then, for any random variable X, the smooth entropy $\Psi(X)$ within $\Delta(s)$ in terms of relative entropy with probability 1 is upper bounded by the entropy of X in the sense that*

$$\Psi(X) \;\leq\; H(X) + \Delta(0) + 1$$

or, more precisely,

$$\log\lfloor 2^{\Psi(X)}\rfloor \;\leq\; H(X) + \Delta(0).$$

Proof. Let $s = 0$. Then, there exists a function f_0 such that $Y_0 = f_0(X,T)$ with $|\mathcal{Y}_0| = \lfloor 2^{\Psi(X)}\rfloor$. By (2.12) and by the definition of smooth entropy in terms of relative entropy

$$\log|\mathcal{Y}_0| - H(Y_0|T) \;=\; D(P_{Y_0|T}\|P_{U|T}) \;\leq\; \Delta(0). \tag{4.2}$$

Then

$$\begin{aligned} H(X) &\geq I(X;Y_0|T) \\ &= H(Y_0|T) - H(Y_0|XT) \\ &\geq \log|\mathcal{Y}_0| - \Delta(0) \\ &= \log\lfloor 2^{\Psi(X)} \rfloor - \Delta(0) \end{aligned}$$

where the last inequality follows from (4.2) and from $H(Y_0|XT) = 0$ because Y_0 is a deterministic function of X and T. □

The formal statement of the lower bound is based on the *Asymptotic Equipartition Property (AEP)* of information theory (see Section 2.5): The set of values of a sequence of n independent, identically distributed (i.i.d.) random variables with distribution P_X can be divided into two sets, a typical set and a non-typical set. The AEP states that an observed sequence lies in the typical set with high probability and that the probability of any typical sequence is close to $2^{-nH(X)}$. Therefore, almost all occurring sequences lie in the typical set and all sequences in the typical set are almost equally probable. Because Rényi entropy is equal to entropy for the uniform distribution, the average smooth entropy per repetition of X is close to the entropy $H(X)$.

Theorem 4.5. *Let $X^n = X_1, \ldots, X_n$ be a sequence of n i.i.d. random variables with distribution P_X and alphabet $\mathcal{X}$ and let $\epsilon > 0$. The average smooth entropy of a random variable X within $2^{-s}/\ln 2$ in terms of relative entropy is not smaller than a value close to the entropy of X with high probability. More precisely,*

$$\frac{1}{n}\Psi(X^n) \;\geq\; H(X) + \epsilon \log \frac{\epsilon}{|\mathcal{X}|}$$

with probability at least $1 - (n+1)^{|\mathcal{X}|} \cdot 2^{-\frac{n}{2\ln 2}\frac{\epsilon^2}{|\mathcal{X}|^2}}$.

Proof. The AEP (Proposition 2.6) states that X^n is in the typical set S_ϵ^n with probability at least

$$1 - (n+1)^{|\mathcal{X}|} \cdot 2^{-\frac{n}{2\ln 2}\frac{\epsilon^2}{|\mathcal{X}|^2}}.$$

All typical sequences satisfy $P_{X^n}(x^n) \leq 2^{-n(H(X)+\epsilon\log\frac{\epsilon}{|\mathcal{X}|})}$ and therefore,

$$H_2(X^n) \;\geq\; H_\infty(X^n) \;\geq\; n\Big(H(X) + \epsilon \log \frac{\epsilon}{|\mathcal{X}|}\Big)$$

by Proposition 2.4. The theorem follows from Corollary 4.1. □

Entropy cannot be used as a lower bound for smooth entropy. This was observed by Bennett et al. [BBCM95] and is illustrated next.

Example 4.4. Suppose that everything we know about a random variable X is $H(X) \geq t$. Then P_X could be such that $P_X(x_0) = p$ for some $x_0 \in \mathcal{X}$ with $p = 1 - t/\log(|\mathcal{X}| - 1)$ and $P_X(x) = (1-p)/(|\mathcal{X}| - 1)$ for all $x \neq x_0$. X satisfies $H(X) = h(p) + (1-p)\log(|\mathcal{X}| - 1) \geq t$. But $X = x_0$ occurs with probability p, and no matter how small a Y is extracted from X, its value can be predicted with probability p. Thus, with knowledge of a lower bound on $H(X)$ alone, the probability with which X can be guessed is not reducible and only a small part of the randomness in X can be converted to uniform bits. Therefore, the entropy of a random variable is not an adequate measure of its smooth entropy. In other words, there are random variables with arbitrarily large entropy and almost no smooth entropy. ○

4.3.4 Relation to Intrinsic Randomness

Intrinsic randomness was introduced by Vembu and Verdú in their work on generating random bits from an arbitrary source [VV95]. For short, intrinsic randomness differs from smooth entropy in that only deterministic extraction functions are considered whereas smooth entropy allows probabilistic extraction functions. But the goal of extracting random bits with a small deviation from the uniform distribution in terms of several distance measures is the same. In addition, Vembu and Verdú investigate fixed-length and variable-length outputs separately and define intrinsic randomness asymptotically, focusing on the achievable rate of intrinsic randomness as $n \to \infty$ for a general random source $X = \{P_{X^n}\}_{n=0}^{\infty}$.

The intrinsic randomness rate $U_M(X)$ of a source X is defined as the largest asymptotic rate at which almost independent equiprobable bits can be generated by a deterministic transformation of the source, with respect to the distance measure M. The distance measures M used by Vembu and Verdú are variational distance, d-bar distance, and normalized relative entropy. They show that the intrinsic randomness rate is the same for all of these measures and is equal to $\underline{\mathrm{H}}(X)$, the inf-entropy rate of X, in the fixed-length case and equal to $\liminf_{n\to\infty} \frac{1}{n} H(X^n)$ in the variable-length case (see [HV93, VV95] for definitions).

For stationary ergodic sources, the fixed-length and the variable-length intrinsic randomness rates are the same and are equal to the entropy rate of the source, $\lim_{n\to\infty} \frac{1}{n} H(X^n)$. Thus, the number of deterministically extractable uniform bits corresponds to the entropy of the source and also corresponds to the number of probabilistically extractable bits for a sequence of independently repeated experiments (see the results on smooth entropy in the previous Section).

For arbitrary sources, the fixed-length intrinsic randomness rate is equal to the inf-entropy rate, which corresponds to the min-entropy $H_\infty(X)$ in the finite case. Intuitively, this means that a random source X is at least as random as the uniform distribution on $2^{H_\infty(X)}$ values because the maximum probability of X is bounded by $2^{-H_\infty(X)}$ and its intrinsic randomness is $H_\infty(X)$. Precisely this amount of randomness, but not more, can be extracted asymptotically by deterministic functions. An example of a source with inf-entropy rate δ is the so-called δ-source [Zuc91], defined as any n-bit source X^n such that $\max_{x^n} P_{X^n}(x^n) \leq 2^{-\delta n}$.

Compared to smooth entropy, which allows probabilistic extraction functions, the fact that deterministic functions are required for intrinsic randomness makes a difference in producing almost uniformly random bits. Bearing in mind that it is a comparison between asymptotic and finite concepts with completely different formalizations, we see nevertheless from Corollary 4.1 that the smooth entropy of a source X is at least $H_2(X)$, the Rényi entropy of order 2, while the fixed-length intrinsic randomness is asymptotically equal to the min-entropy, $H_\infty(X)$. (It follows from our results in Section 4.5.1 that smooth entropy is lower bounded by Rényi entropy of order α for any $\alpha > 1$.)

4.3.5 An Application in Learning Theory

Computational learning theory provides a formalization of learning with Valiant's *PAC (probably approximately correct)* model [Val84, KV94]. A concept to be learned is simply a subset of an n-dimensional example space, containing all positive examples of the concept. The goal of the learning algorithm $\mathcal{A}$ is to identify a target concept c from the set $\mathcal{C}$ of all concepts ($\mathcal{C}$ is known to the algorithm). During learning, examples are sampled using the target distribution of the learning problem and presented to $\mathcal{A}$, together with the classification of the example by c (positive or negative). The target distribution does not depend on c.

The learning algorithm outputs with probability at least $1 - \delta$ a hypothesis $h \in \mathcal{C}$ whose error must not exceed ϵ. The error of h, and therefore the error of the learning algorithm, is defined as the probability that a random sample under the target distribution is mis-classified by h, i.e. that the classifications by c and by h differ.

An efficient PAC learning algorithm is one that takes at most polynomial time in n, $\frac{1}{\epsilon}$, and $\frac{1}{\delta}$. (The actual definition also takes into account the question of representing examples and concepts, which we ignore here.)

Research in computational learning theory focuses on constructing efficient learning algorithms and on characterizing problems for which no efficient learning algorithms exist. An important result in the latter direction was shown by Kharitonov [Kha93] and states that, under the assumption that factoring Blum integers is computationally hard, boolean formulas are not efficiently PAC learnable for very general target distributions. More precisely, the result holds if the target distribution has Rényi entropy of order 2 greater than $O(\log n)$. The main step in the proof uses entropy smoothing to concentrate the hard instances of the problem sufficiently to prevent efficient learning. (By our results of Section 4.5.1, this can be extended to Rényi entropy of order α for any $\alpha > 1$.)

4.3.6 Extractors, Weak Random Sources, and Derandomization

An *extractor* is a tool developed for running randomized algorithms with sources of non-perfect randomness instead of uniform random bits. Such sources are called *weak random sources* [CG85]. The focus of the research on extractors in complexity theory lies on characterizing the random sources needed for polynomial-time probabilistic computations. The principal difference between extractors and the entropy smoothing framework is that, for extractors, the number of auxiliary random bits is counted as a resource and should be small. If there are only a logarithmic number of them, a deterministic procedure can try all the possibilities in polynomial time and thereby *derandomize* the original algorithm.

Formally, an (m, ϵ)-extractor is a function $E : \mathcal{X} \times \mathcal{T} \to \mathcal{Y}$ if, for any random variable X with min-entropy $H_\infty(X) \geq m$, the output $Y = E(X, T)$ of the extractor satisfies $\|P_Y - P_U\|_v \leq \epsilon$ when T is chosen uniformly random from $\mathcal{T}$ and where P_U denotes the uniform distribution over $\mathcal{Y}$ [TS96].

Using $n = \log|\mathcal{X}|$, $t = \log|\mathcal{T}|$, and $s = \log|\mathcal{Y}|$, the best currently known constructions of extractors can produce $s = \Omega(m)$ almost uniform bits using $t = O(log(n/\epsilon))$ truly random bits, from an n-bit source with min-entropy $m = \Omega(n)$ [Zuc96a]. Another construction achieves $s = m$ with $t = poly(\log(n/\epsilon))$, extracting all min-entropy [TS96], but requiring a larger auxiliary random input.

At the core, all extractor constructions known are based on entropy smoothing by variations of universal hashing (in the sense of Theorems 2.7 and 4.2). But since it requires at least n uniformly random bits to smooth an n-bit source by universal hashing, several steps are taken to expand the truly random input. A recursive construction, initialized by T, is used that adds some part of the input randomness in every step for smoothing other parts in later steps. We refer to the survey by Nisan [Nis96] for further information about how extractors work.

Randomized algorithms are in widespread use today because of their speed, space efficiency, simplicity, or other desirable properties [MR95]. In order to implement them, sources of true random bits are needed but are usually not available and replaced by pseudorandom generators. Alternatively, physical sources of noise could be employed, such as Zener Diodes. The problem with using physical sources is, however, that they do not output truly unbiased and independent bits, hence the name "weak random sources."

A general model for such a source is the δ-source, which is any n-bit source with min-entropy at least δn [Zuc91]. An extractor can be used to compute any given randomized algorithm using a δ-source by simulating the possible choices of T. The current extractor constructions allow **RP** to be simulated in polynomial time with sources X that have $H_\infty(X) \geq n^c$ for any constant $c > 0$. **BPP** can be simulated in polynomial time as long as $H_\infty(X) = \Omega(n)$.

Other applications of extractors include decreasing the error probability of randomized algorithms (deterministic amplification, oblivious sampling) and the construction of graphs with "random" properties (superconcentrators, expanders). More information about applications of extractors can be found in the work of Zuckerman [Zuc91, SZ94, WZ95, NZ95, Zuc96a, Zuc96b] and Nisan [Nis96].

4.4 Spoiling Knowledge

4.4.1 Introduction

We now turn to further characterizations of smooth entropy and to lower bounds in terms of Rényi entropy. Corollary 4.1 shows that Rényi entropy of order 2 is a lower bound for the smooth entropy of a distribution. As mentioned in Section 2.4, a counter-intuitive property of conditional Rényi entropy of order $\alpha > 1$ is that it can increase even on the average when conditioned on a random variable that provides side information. Suppose side information that increases Rényi entropy is made available by a conceptual oracle. This increase can be exploited to prove lower bounds on smooth entropy that are much better than Rényi entropy of order 2. Side information of this kind was introduced by Bennett et al. [BBCM95] and is called *spoiling knowledge* because it leads to less information about the output of the smoothing process.

In this section we first distinguish between two kinds of spoiling knowledge, namely those that yield bounds which either hold with probability 1 or include some small failure probability. We then characterize spoiling knowledge of the first type in Section 4.4.2 and examine the second type in Section 4.4.3.

The results of this section and the next are stated for the smooth entropy of a random variable X in terms of relative entropy. However, they hold similarly for the smooth entropy of families of random variables and in terms of L_1 distance and other nonuniformity measures as discussed in Section 4.2.

An oracle giving spoiling knowledge is used only as a thought experiment for proofs and not actually for smoothing a random variable. The bounds provided by this method hold because such an oracle could exist in every actual scenario involving smoothing. The oracle knows the distribution of the source and can prepare the side information depending on the particular distribution.

This proof technique of using auxiliary random variables as spoiling knowledge was introduced by Bennett et al. [BBCM95] in the privacy amplification scenario.

Spoiling knowledge is modeled by a random variable U provided by the oracle. The side information increases the Rényi entropy of X such that $H_2(X|U = u)$ exceeds $H_2(X)$ for some u with a certain probability.

Let P_{XU} be an arbitrary distribution such that the marginal distribution for X coincides with P_X. By maximizing over choices of U, Theorem 2.7 gives [BBCM95]

$$H(Y|G) \;\geq\; \log|\mathcal{Y}| - \sum_{u\in\mathcal{U}} P_U(u)\cdot\min\left\{\log|\mathcal{Y}|, \frac{2^{\log|\mathcal{Y}|-H_2(X|U=u)}}{\ln 2}\right\}. \tag{4.3}$$

The discussion in [BBCM95] suggests that the maximization of the expected conditional Rényi entropy $H_2(X|U) = \sum_{u\in\mathcal{U}} P_U(u)H_2(X|U=u)$ corresponds to the maximization of the right-hand side in (4.3). However, this is not the case because

$$\begin{aligned}
&\sum_{u\in\mathcal{U}} P_U(u)\cdot\min\left\{\log|\mathcal{Y}|, \frac{2^{\log|\mathcal{Y}|-H_2(X|U=u)}}{\ln 2}\right\} \\
&= \sum_{u\in\mathcal{U}} P_U(u)\cdot\min\left\{\log|\mathcal{Y}|, \frac{|\mathcal{Y}|}{\ln 2}\sum_{x\in\mathcal{X}} P_{X|U=u}(x)^2\right\} \\
&= \frac{|\mathcal{Y}|}{\ln 2}\sum_{u\in\mathcal{U}} P_U(u)\cdot\min\left\{\frac{\ln 2\log|\mathcal{Y}|}{|\mathcal{Y}|}, \sum_{x\in\mathcal{X}} P_{X|U=u}(x)^2\right\}.
\end{aligned}$$

From this, we see that the right-hand side of equation (4.3) is maximal if the following expression is minimal for choices of U such that P_{XU} is consistent with P_X:

$$\sum_{u\in\mathcal{U}} P_U(u)\cdot\min\left\{\frac{\ln|\mathcal{Y}|}{|\mathcal{Y}|}, \sum_{x\in\mathcal{X}} P_{X|U=u}(x)^2\right\}. \tag{4.4}$$

Example 4.5. Consider a random variable X over the alphabet $\mathcal{X} = \{x_1,\ldots,x_{181}\}$ with distribution

x	x_1	$x_2,\ldots,x_{21}$	$x_{22},\ldots,x_{101}$	$x_{102},\ldots x_{181}$
$P_X(x)$	0.25	0.0075	0.005	0.0025

Suppose X is hashed to $Y = G(X)$ with $|\mathcal{Y}| = 2$ using the universal hash function $\mathcal{G}$ described on page 57. A calculation of the uniformity of the output shows that $H(Y|G) = 0.945$.

The lower bound from Theorem 2.7 that uses Rényi entropy of order 2 shows $H(Y|G) \geq 0.809$. Rényi entropy yields a weak bound because one value of X has a very high probability and hence the Rényi entropy of X lies significantly below its entropy: $H(X) = 6.35$ and $H_2(X) = 3.92$.

This bound can be improved by introducing an auxiliary random variable $U \in \{0,1\}$ that reveals whether $X = x_1$ with probability

$p/P_X(x_1)$ for some p. The joint distribution P_{XU} is $P_{XU}(x_1, 1) = P_U(1) = p$, $P_{XU}(x_1, 0) = 0.25 - p$ and $P_{XU}(x_i, 0) = P_X(x_i)$ for $i > 1$.

How large should p be? Based on the discussion in the preceding paragraph, p could be chosen either to maximize the expected conditional Rényi entropy $H_2(X|U)$ or to minimize (4.4), but these two goals are not achieved simultaneously:

Goal	p	$H_2(X\|U)$	$H(Y\|G)$ in (4.3)
Maximize cond. Rényi entropy	0.2320	5.55	0.753
Minimize expression (4.4)	0.0692	4.26	0.818
No side info. (Thm. 2.7)	0	3.92	0.809

(The last row is repeated for comparison.) No bound comes close to the true value $H(Y|G) = 0.945$.

If we accept that the desired uniformity is not achieved with some small probability, we can state a (probably sharper) conditional bound on the uniformity. For example, let $p = 0.1$. Then, with probability 0.9, $U = 0$ and $H_2(X|U = 0) \geq 4.95$, from which $H(Y|G, U = 0) \geq 0.907$ follows. Setting $p = 0.2$, it holds with probability 0.8 that $H_2(X|U = 0) \geq 6.71$ and $H(Y|G, U = 0) \geq 0.972$. ○

The example shows that a spoiling knowledge argument involving minimization of (4.4) (which is equivalent to maximizing the right-hand side of (4.3)) can yield better lower bounds on the smooth entropy than Rényi entropy of order 2. The effect of spoiling knowledge is based on the fact that Rényi entropy can increase under conditioning, but the optimal side information cannot be found by simply maximizing the conditional Rényi entropy of order 2. A characterization of optimal spoiling knowledge in the sense of minimizing (4.4) is given below in Section 4.4.2.

The alternative way to derive better lower bounds on the smooth entropy is to accept a failure event with some small probability ϵ, as in Example 4.5. This enables more general side information to be exploited for characterizing smooth entropy with probability $1 - \epsilon$. Such side information is further investigated in Section 4.4.3. Lower bounds on the smooth entropy using this kind of spoiling knowledge argument are obtained in Section 4.5.

4.4.2 Spoiling Knowledge for Increasing Smooth Entropy with Probability 1

The goal of this subsection is to investigate suitable choices of an auxiliary random variable U that can always (i.e. with probability 1) increase the lower bound on the smooth entropy of X. We restrict our attention to relative entropy distance as the nonuniformity measure and to smoothing via universal hash functions. Recall that the definition of smooth entropy with probability 1 requires, for every $s \geq 0$, the existence of a hash function f such that $Y = f(X,T)$ with $|\mathcal{Y}| = \lfloor 2^{\psi - s} \rfloor$ satisfies $M(Y|T) \leq \Delta(s)$ for some measure of nonuniformity M.

Consider again the optimization problem in (4.3) and (4.4): To maximize the smooth entropy by finding suitable side information, $\log |\mathcal{Y}|$ must be maximized and (4.4) must be minimized jointly, which makes the optimization more involved than it looks at first glance. An explicit lower bound on the smooth entropy of a random variable X can be obtained if we use the trivial bound $|\mathcal{Y}| \leq |\mathcal{X}|$. The oracle knows X and can prepare the side information depending on the distribution of X. This result is summarized in the next theorem.

Theorem 4.6. *The smooth entropy $\Psi(X)$ within $2^{-s}/\ln 2$ with probability 1 of a random variable X is lower bounded by the following expression involving a minimization over an arbitrary random variable U such that the joint distribution P_{XU} is consistent with P_X:*

$$\Psi(X) \;\geq\; -\log \min_{P_U} \left(\sum_{u \in \mathcal{U}} P_U(u) \cdot \min\left\{ \frac{\ln |\mathcal{X}|}{|\mathcal{X}|}, \sum_{x \in \mathcal{X}} P_{X|U=u}(x)^2 \right\} \right). \tag{4.5}$$

In the remainder of this section, we will use the following notation. Let $\mathcal{X} = \{1, \ldots, n\}$ and $p_i = P_X(i)$ such that

$$p_1 \geq \cdots \geq p_n. \tag{4.6}$$

Let $d = \frac{\ln |\mathcal{Y}|}{|\mathcal{Y}|}$ be a constant determined by the size of the output alphabet $\mathcal{Y}$ that satisfies

$$d \;\leq\; \frac{\ln |\mathcal{X}|}{|\mathcal{X}|}.$$

We first consider only binary auxiliary random variables $U \in \{0,1\}$ and denote the joint distribution P_{XU} by

$$P_{XU}(x,0) = \gamma_i p_i \quad \text{and} \quad P_{XU}(i,1) = (1-\gamma_i) p_i$$

for $i = 1, \ldots, n$. The next lemma shows a simple and useful fact about conditional Rényi entropy.

Lemma 4.7. *Let X be an arbitrary random variable and let U be a binary random variable. Then*

$$P_U(0) \cdot 2^{-H_2(X|U=0)} + P_U(1) \cdot 2^{-H_2(X|U=1)} \;\geq\; 2^{-H_2(X)}$$

and

$$\max\Big\{\sum_x P_{X|U=0}(x)^2, \sum_x P_{X|U=1}(x)^2\Big\} \;\geq\; \sum_x P_X(x)^2$$

with equality in both statements if and only if $P_U(0) = 0$ or $P_U(1) = 0$.

Proof. The first statement of the lemma is equivalent to

$$P_U(0) \sum_x P_{X|U=0}(x)^2 + P_U(1) \sum_x P_{X|U=1}(x)^2 \;\geq\; \sum_x P_X(x)^2 \qquad (4.7)$$

from which the second statement follows immediately. Using the notation introduced above, this is equivalent to

$$\frac{\sum_i \gamma_i^2 p_i^2}{\sum_i \gamma_i p_i} + \frac{\sum_i (1-\gamma_i)^2 p_i^2}{\sum_i (1-\gamma_i) p_i} \;\geq\; \sum_i p_i^2. \qquad (4.8)$$

Equality for trivial side information with $P_U(0) = 0$ or $P_U(1) = 0$ is obvious. We now show that the inequality in (4.8) is strict for non-trivial side information. Subtracting $\sum_i p_i^2$ and multiplying by $(\sum_i \gamma_i p_i)(\sum_i (1 - \gamma_i) p_i)$, we get

$$\begin{aligned}
&\Big(\sum_i \gamma_i^2 p_i^2\Big)\Big(\sum_i (1-\gamma_i) p_i\Big) + \Big(\sum_i (1-\gamma_i)^2 p_i^2\Big)\Big(\sum_i \gamma_i p_i\Big) \\
&\qquad\qquad - \Big(\sum_i \gamma_i p_i\Big)\Big(\sum_i (1-\gamma_i) p_i\Big)\Big(\sum_i p_i^2\Big) \\[1em]
&= \sum_i \gamma_i^2 p_i^2 - \Big(\sum_i \gamma_i^2 p_i^2\Big)\Big(\sum_i \gamma_i p_i\Big) + \Big(\sum_i (1-\gamma_i)^2 p_i^2\Big)\Big(\sum_i \gamma_i p_i\Big) \\
&\qquad\qquad - \Big(\sum_i \gamma_i p_i - \Big(\sum_i \gamma_i p_i\Big)^2\Big)\Big(\sum_i p_i^2\Big)
\end{aligned}$$

$$\begin{aligned} &= \sum_i \gamma_i^2 p_i^2 - \sum_{i,j} \gamma_i^2 p_i^2 \gamma_j p_j + \Big(\sum_i (1 - 2\gamma_i + \gamma_i^2) p_i^2\Big) \sum_i \gamma_i p_i \\ &\qquad - \Big(\sum_i \gamma_i p_i\Big) \sum_i p_i^2 + \Big(\sum_i \gamma_i p_i\Big)^2 \sum_i p_i^2 \end{aligned}$$

$$\begin{aligned} &= \sum_i \gamma_i^2 p_i^2 - \underline{\sum_{i,j} \gamma_i^2 p_i^2 \gamma_j p_j} + \underline{\sum_{i,j} p_i^2 \gamma_j p_j} - 2 \sum_{i,j} \gamma_i p_i^2 \gamma_j p_j \\ &\qquad + \underline{\sum_{i,j} \gamma_i^2 p_i^2 \gamma_j p_j} - \underline{\sum_{i,j} \gamma_i p_i p_j^2} + \Big(\sum_j \gamma_j p_j\Big)^2 \sum_i p_i^2. \end{aligned}$$

The underlined sums cancel out. Isolating p_i^2 from the remaining terms, we get

$$\sum_i p_i^2 \left(\gamma_i^2 - 2\gamma_i \sum_j \gamma_j p_j + \Big(\sum_j \gamma_j p_j\Big)^2 \right) = \sum_i p_i^2 \Big(\gamma_i - \sum_j \gamma_j p_j\Big)^2$$

which is strictly positive. □

To investigate the effect of binary spoiling knowledge, we assume w.l.o.g. that $H_2(X|U = 0) > H_2(X|U = 1)$, i.e.

$$\frac{\sum_i \gamma_i^2 p_i^2}{(\sum_i \gamma_i p_i)^2} \;<\; \frac{\sum_i (1-\gamma_i)^2 p_i^2}{(\sum_i (1-\gamma_i) p_i)^2}. \tag{4.9}$$

The minimization in Theorem 4.6 becomes

$$\begin{aligned} &\sum_i \gamma_i p_i \cdot \min\left\{ d, \frac{\sum_i \gamma_i^2 p_i^2}{(\sum_i \gamma_i p_i)^2} \right\} + \\ &\qquad \sum_i (1-\gamma_i) p_i \cdot \min\left\{ d, \frac{\sum_i (1-\gamma_i)^2 p_i^2}{(\sum_i (1-\gamma_i) p_i)^2} \right\} \end{aligned} \tag{4.10}$$

for some constant $d \le \frac{\ln |\mathcal{X}|}{|\mathcal{X}|}$. Depending on the selections in the minimum operators, four different cases have to be considered in the minimization. But only one of them is interesting, as the next result shows.

Lemma 4.8. *Let $d \leq \frac{\ln|\mathcal{X}|}{|\mathcal{X}|}$. Binary side information U can increase the lower bound on smooth entropy $\Psi(X)$ with probability 1 in the sense of Theorem 4.6 only if*

$$\frac{\sum_i \gamma_i^2 p_i^2}{(\sum_i \gamma_i p_i)^2} \ < \ d \quad \text{and} \quad d \ < \ \frac{\sum_i (1-\gamma_i)^2 p_i^2}{(\sum_i (1-\gamma_i) p_i)^2}.$$

Proof. If side information U is to increase the smooth entropy of X, expression (4.10) has to be smaller than $\sum_i p_i^2$. Because of (4.9), there are three possible cases for the minimum operators: In the first case, if $d < \sum_i \gamma_i^2 p_i^2 / (\sum_i \gamma_i p_i)^2$, (4.10) reduces (4.5) to the trivial bound that entropy is non-negative. The second case is the one mentioned in the lemma. For the third case $d > \sum_i (1-\gamma_i)^2 p_i^2 / (\sum_i (1-\gamma_i) p_i)^2$, the bound on smooth entropy can be increased by side information U only if

$$\frac{\sum_i \gamma_i^2 p_i^2}{\sum_i \gamma_i p_i} + \frac{\sum_i (1-\gamma_i)^2 p_i^2}{\sum_i (1-\gamma_i) p_i} \ < \ \sum_i p_i^2. \tag{4.11}$$

But this is not possible according to Lemma 4.7. No increase of the lower bound is therefore possible in the third case. □

Binary spoiling knowledge can therefore only be useful if observing one of its values leaves small Rényi entropy about X, $H_2(X|U=1) < 2^{-d}$, where d is determined by the size of the hashing output. This property extends to the general case: An auxiliary random variable U with arbitrary alphabet $\mathcal{U}$ can sharpen the lower bound on smooth entropy only if for all $u \in \mathcal{U}$ except for one, $H_2(X|U=u) < 2^{-d}$. W.l.o.g. we assume $\mathcal{U} = \{0, \ldots, m\}$ and

$$H_2(X|U=0) \ \geq \ \ldots \ \geq \ H_2(X|U=m).$$

Theorem 4.9. *Let $d \leq \frac{\ln|\mathcal{X}|}{|\mathcal{X}|}$. Side information U can increase the lower bound on smooth entropy $\Psi(X)$ with probability 1 in the sense of Theorem 4.6 only if*

$$\begin{aligned} H_2(X|U=0) \ &> \ 2^{-d} \\ H_2(X|U=j) \ &\leq \ 2^{-d} \quad \textit{for } j = 1, \ldots, m. \end{aligned}$$

Proof. The proof is based on Lemma 4.7. Suppose that there is an auxiliary random variable U minimizing (4.4) and satisfying, for some $k \neq l$

$$H_2(X|U=k) > 2^{-d} \quad \text{and} \quad H_2(X|U=l) > 2^{-d}.$$

Consider the random variable X' with alphabet $\{1,\ldots,n\}$ and distribution $P_{X'} = P_{X|U\in\{k,l\}}$ and the binary auxiliary random variable $U' \in \{0,1\}$ with joint distribution

$$P_{X'U'}(i,0) = \frac{P_{XU}(i,k)}{P_{XU}(i,k)+P_{XU}(i,l)} P_X(i).$$

It follows from Lemma 4.7 that for non-trivial side information

$$\begin{aligned}\sum_i P_{X|U=k}(i)^2 + \sum_i P_{X|U=l}(i)^2 \\ = \sum_i P_{X'|U'=0}(i)^2 + \sum_i P_{X'|U'=1}(i)^2 \\ > \sum_i P_{X'}(i)^2 = \sum_i P_{X|U\in\{k,l\}}(i)^2.\end{aligned}$$

This contradicts the minimality of U in (4.4). □

The theorem shows that, for optimal spoiling knowledge, $H_2(X|U = u) > 2^{-d}$ only for one u, and that the optimal distribution can be found by a minimization in only n variables. The minimal value of (4.4) can be achieved when all other values of U indicate one value of X in a one-to-one correspondence.

Corollary 4.10. *The optimal auxiliary random variable U that gives a lower bound on the smooth entropy $\Psi(X)$ with probability 1 in the sense of Theorem 4.6 can be found by solving the following optimization problem in the n variables $\gamma_1,\ldots,\gamma_n$:*

minimize	$\frac{\sum_i \gamma_i^2 p_i^2}{\sum_i \gamma_i p_i} + d(1-\sum_i \gamma_i p_i)$
subject to	$0 \le \gamma_i \le 1$ for all $i = 1,\ldots,n$

Proof. Suppose $\gamma_1,\ldots,\gamma_n$ are solutions of the minimization above. Let $\{i_1,\ldots,i_m\} = \{i|\gamma_i < 1\}$ be the indices of values of X affected by the side information. Define the random variable $U \in \{0,\ldots,m\}$ jointly distributed with X according to

$$P_{XU}(i,j) = \begin{cases} \gamma_i p_i & \text{if } j = 0 \\ (1-\gamma_{i_j})p_i & \text{if } i = i_j \\ 0 & \text{otherwise.} \end{cases}$$

U satisfies $H_2(X|U=j) = 0 < 2^{-d}$ for all $j > 0$. As a consequence of Theorem 4.9, no other cases of the minimum operators in (4.4) are relevant. Therefore, it is easy to see that no auxiliary random variable achieving the minimum in (4.4) with a distribution different from P_U can yield a better lower bound. □

The distribution of optimal spoiling knowledge can be found by solving the optimization problem numerically. It contains only $|\mathcal{X}|$ variables. What can be said about its solution? The first term in the minimization corresponds to the probability $\sum_i P_{X|U=0}(i)^2$, to which large probabilities contribute an undesirably large amount (see Example 4.5, page 66). As the next theorem shows, the best strategies for smooth entropy-increasing side information assign smaller γ to larger probabilities of X. In this way, the larger values of P_X are prevented from increasing $\sum_i P_{X|U=0}(i)^2$ by too much.

Theorem 4.11. *The optimal auxiliary random variable U that gives a lower bound on the smooth entropy $\Psi(X)$ with probability 1 in the sense of Theorem 4.6 satisfies*

$$\gamma_1 \leq \cdots \leq \gamma_n.$$

Proof. The minimization in Corollary 4.10 is equivalent to finding the minimum of

$$\Theta = \frac{\sum_i \gamma_i p_i \Big(\gamma_i p_i + d\big(1 - \sum_j \gamma_j p_j\big)\Big)}{\sum_i \gamma_i p_i}.$$

Assume Θ is minimal and there exist indices $a < b$ such that $\gamma_a > \gamma_b$ (remember that $p_1 \geq \cdots \geq p_n$). We can construct $\Theta' < \Theta$ as follows: Choose any $\gamma'_b > \gamma_b$ and let $\gamma'_a = \gamma_a + \frac{p_b}{p_a}(\gamma_b - \gamma'_b) < \gamma_a$. Let $\gamma'_i = \gamma_i$ for $i \neq a$ and $i \neq b$. We observe that $\sum_i \gamma'_i p_i = \sum_i \gamma_i p_i$. Then

$$\begin{aligned}
&\Theta - \Theta' \\
&= \frac{\sum_i \gamma_i p_i \Big(\gamma_i p_i + d\big(1 - \sum_j \gamma_j p_j\big)\Big)}{\sum_i \gamma_i p_i} - \frac{\sum_i \gamma'_i p_i \Big(\gamma'_i p_i + d\big(1 - \sum_j \gamma'_j p_j\big)\Big)}{\sum_i \gamma'_i p_i} \\
&\qquad \geq 0,
\end{aligned}$$

together with $\delta = d(1 - \sum_j \gamma_j p_j) > 0$, reduces to

$$\begin{aligned}
0 &\leq \gamma_a p_a(\gamma_a p_a + \delta) + \gamma_b p_b(\gamma_b p_b + \delta) \\
&\quad -\gamma'_a p_a(\gamma'_a p_a + \delta) - \gamma'_b p_b(\gamma'_b p_b + \delta) \\
&= \delta\underbrace{(\gamma_a p_a + \gamma_b p_b - \gamma'_a p_a - \gamma'_b p_b)}_{=0} + \gamma_a^2 p_a^2 + \gamma_b^2 p_b^2 - {\gamma'_a}^2 p_a^2 - {\gamma'_b}^2 p_b^2 \\
&= \gamma_a^2 p_a^2 + \gamma_b^2 p_b^2 - \left(\gamma_a + \frac{p_b}{p_a}(\gamma_b - \gamma'_b)\right)^2 p_a^2 - {\gamma'_b}^2 p_b^2 \\
&= \underbrace{(-2p_b)}_{A}{\gamma'_b}^2 + \underbrace{(2\gamma_a p_a p_b + 2\gamma_b p_b^2)}_{B}\gamma'_b \underbrace{-2\gamma_a\gamma_b p_a p_b}_{C},
\end{aligned}$$

where a quadratic polynomial in γ'_b results in the last step after expanding and simplifying the previous expression. It follows from basic calculus that this quadratic polynomial is lower bounded by

$$\begin{aligned}
\frac{4AC - B^2}{4A} &= \frac{16\gamma_a\gamma_b p_a p_b^2 - 4\gamma_a^2 p_a^2 p_b^2 - 8\gamma_a\gamma_b p_a p_b^3 - 4\gamma_b^2 p_b^4}{-8p_b^2} \\
&= -2\gamma_a\gamma_b p_a p_b + \frac{1}{2}\gamma_a^2 p_a^2 + \frac{1}{2}\gamma_b^2 p_b^2 + \gamma_a\gamma_b p_a p_b \\
&= \frac{1}{2}\left(\gamma_a^2\gamma_b^2 - 2\gamma_a\gamma_b p_a p_b + \gamma_b^2 p_b^2\right) \\
&= \frac{1}{2}\left(\gamma_a p_a - \gamma_b p_b\right)^2
\end{aligned}$$

which is always positive under our assumptions $p_a \geq p_b$ and $\gamma_a > \gamma_b$. □

The results of this subsection characterize side information that increases the lower bounds on the smooth entropy with probability 1. It turns out that suitable side information reduces large values of the probability distribution by the constant lower bound d in the minimum operator in Theorem 4.6. Hence, only spoiling knowledge that "cuts off" the most probable values of a random variable can increase the lower bounds on smooth entropy. The distribution of the optimal side information can be found by numerical optimization methods.

4.4.3 Spoiling Knowledge for Increasing Smooth Entropy with Probabilistic Bounds

As the results of the preceding section show, auxiliary random variables provided by a conceptual oracle can be used to obtain better bounds on smooth entropy. The side information investigated above was confined not to change the probability with which the resulting bound holds. If we relax this constraint and allow a failure event with probability ϵ, a broader range of side information is applicable. We refer to Example 4.5 (page 66) for an illustration.

Smooth entropy with probability $1-\epsilon < 1$ is defined in terms of a failure event $\mathcal{E}$ such that $\mathrm{P}[\mathcal{E}] \leq \epsilon$. For characterizing spoiling knowledge to increase the corresponding lower bounds, the statement of Theorem 4.6 has to be conditioned on $\overline{\mathcal{E}}$. As Theorem 4.9 shows for smooth entropy with probability 1, spoiling knowledge U can increase the bound of Theorem 4.6 only if $H_2(X|U=0)$ is large (where we have adopted the terminology of Theorem 4.9). We therefore focus only on conditioning $P_{X|U=0}$ on $\overline{\mathcal{E}}$, which leads to a further increase of $H_2(X|U=0)$ and the bound in Theorem 4.6. The following theorem summarizes this formally.

Theorem 4.12. *Given a random variable X, let U be side information with alphabet $\{0,\ldots,m\}$ as used in Theorem 4.9. A lower bound on the smooth entropy $\Psi(X)$ within $2^{-s}/\ln 2$ with probability $1-\epsilon$ of a random variable X can be found by maximizing the conditional Rényi entropy $H_2(X|U=0,S=0)$ over the choice of binary side information S such that $P_{XS|U=0}$ is consistent with $P_{X|U=0}$ and $P_{S|U=0}(0) \geq 1-\epsilon$:*

$$\Psi(X) \;\geq\; -\log \min_{P_U,\; P_S \,:\, P_{S|U=0}(0)\geq 1-\epsilon} \Bigg(P_U(0) \sum_{x\in\mathcal{X}} P_{X|U=0,S=0}(x)^2 + \; (1-P_U(0)) \frac{\ln|\mathcal{X}|}{|\mathcal{X}|} \Bigg). \quad (4.12)$$

The main result of this section shows that optimal spoiling knowledge in the sense of Theorem 4.12 is provided a random variable S that "cuts off" the values of the probability distribution $P_{X|U=0}$ above some level σ such that the total probability mass above σ is ϵ.

The notation is similar to the preceding section, using $\mathcal{X}=\{1,\ldots,n\}$ and $p_i = P_{X|U=0}(i)$ such that $p_1 \geq \cdots \geq p_n$. Summations over i are implicit from 1 to n if not restricted further. The side information that induces $\mathcal{E}$ is denoted by the random variable S with alphabet $\{0,1\}$.

Consider the following algorithm (in C-like pseudocode) that takes `eps` $=\epsilon, n$, and `p[1..n]` $=[p_1,\ldots,p_n]$ as inputs and returns `e[1..n]` $=[\epsilon_1,\ldots,\epsilon_n]$ and `k` $=k$. We will show that $\epsilon_1,\ldots,\epsilon_n$ determined by the algorithm correspond to the distribution of side information S that provides the maximal increase of conditional Rényi entropy with probability

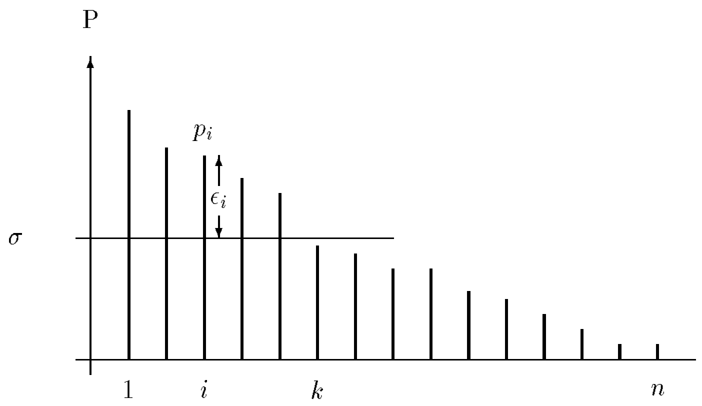

Figure 4.2. The probability distribution $p_1, \ldots, p_n$ of X and the optimal spoiling knowledge U of Theorem 4.13. k is the index of the largest probability p_i equal to or less than σ.

$1 - \epsilon$.

```
spoil(float eps, int n, float p[1..n])
{
  int i, k; float d, e[1..n];

  for (i = 1; i <= n; i++) e[i] = 0;
  k = 1;
  while (eps > 0 && k < n) {
    d = min(eps/k, p[k]-p[k+1]);
    for (i = 1; i <= k; i++) e[i] += d;
    eps -= k*d;
    k++;
  }
  return (e[1..n], k);
}
```

Theorem 4.13. *The optimal side information S in the sense of Theorem 4.12 is given by the joint distribution $P_{XS|U=0}(i, 0) = p_i - \epsilon_i$ and $P_{XS|U=0}(i, 1) = \epsilon_i$ for $i = 1, \ldots, n$, where $\epsilon_1, \ldots, \epsilon_n$ are determined by the algorithm described above.*

Furthermore, if $\epsilon < 1 - np_n$, then $\epsilon_1, \ldots, \epsilon_n$ are nonnegative numbers *such that*

$$\sum_i \epsilon_i \;=\; \epsilon \quad \textit{and} \quad p_i - \epsilon_i = \sigma \;\; \textit{for all } i \textit{ with } \epsilon_i > 0 \tag{4.13}$$

for some constant σ determined by ϵ and $p_1, \ldots, p_n$. Otherwise, when $\epsilon \geq 1 - np_n$, then

$$\epsilon_i = p_i - p_n \quad \textit{for } i = 1, \ldots, n \tag{4.14}$$

and the uniform distribution over $\mathcal{X}$ is induced by $S = 0$, i.e. $H_2(X|U = 0, S = 0) = \log|\mathcal{X}|$.

Proof. The proof consists of two parts: First, the algorithm is shown to establish the statement of the theorem. We then show that no other binary random variable can provide a better bound.

Correctness of algorithm: Typewriter font like $\mathtt{e}_i$ is used for dynamic values `e[i]` of the algorithm. The invariant of the main loop (line 7) is, up to the last repetition,

$$\sum_{i<\mathtt{k}} \mathtt{e}_i + \mathtt{eps} = \epsilon \quad \text{and} \quad \mathtt{e}_i \;=\; p_i - p_{\mathtt{k}} \;\; \text{for } i = 1, \ldots, \mathtt{k} - 1 \tag{4.15}$$

and clearly holds initially. Note that if $\mathtt{d} = \mathtt{eps}/\mathtt{k}$ in line 8, the loop will terminate. Therefore, the invariant is guaranteed by line 9 ($\mathtt{e}_i = \sum_{i \leq j < \mathtt{k}} p_j - p_{j+1} = p_i - p_{\mathtt{k}}$ for $i = 1, \ldots \mathtt{k} - 1$) and by line 11.

The loop terminates either because $\mathtt{eps} = 0$ or $\mathtt{k} = n$. In the first case, the last repetition has changed the second part of the invariant (4.15) to

$$\epsilon_i \;=\; \mathtt{e}_i \;=\; p_i - p_{\mathtt{k-1}} + \mathtt{d} \;\; \text{for } i = 1, \ldots, \mathtt{k} - 1 \tag{4.16}$$

(which holds in line 13). The first part of (4.15) with $\mathtt{eps} = 0$ implies $\sum_{i<\mathtt{k}} \epsilon_i = \epsilon$. Together with (4.16), $p_i - \epsilon_i = p_{\mathtt{k}-1} + \mathtt{d}$ for all $i < \mathtt{k}$ and the first case (4.13) of the theorem follows.

In the second case (termination by $\mathtt{k} = n$), the invariant (4.15) holds also with $\mathtt{k} = n$. Therefore $p_i - \mathtt{e}_i = p_i - \epsilon_i = p_n$ for $i = 1, \ldots, n - 1$ and $\epsilon_n = 0$, which is the second case of the theorem (4.14). The second case results if and only if

$$1 - np_n = \sum_i (p_i - p_n) = \sum_i \epsilon_i \leq \epsilon.$$

Optimality among binary side information: Only the first case (4.13) of the theorem has to be proved. In the second case, S is optimal because $H_2(X|U=0,S=0) = \log|\mathcal{X}|$ is maximal. Suppose that there is a random variable $V \in \{0,1\}$ with $P_{XV|U=0}(i,0) = p_i - \alpha_i$ and $P_{XV|U=0}(i,1) = \alpha_i$ such that $P_{V|U=0}(1) = \sum_i \alpha_i = \alpha \leq \epsilon$ and $H_2(X|U=0,V=0) > H_2(X|U=0,S=0)$.

Let k be the minimal i for which $\epsilon_i = 0$. For all $i \geq k$, $\epsilon_i = 0$ and $p_i \leq \sigma$ (see Figure 4.2). Let $\delta_i = p_i - \alpha_i - \sigma$ for $i = 1, \ldots, k-1$. Then

$$\begin{aligned}
&\sum_i P_{X|V=0}(i)^2 - \sum_i P_{X|U=0}(i)^2 \\
&= \sum_i (p_i - \alpha_i)^2 + \sum_i (p_i + \epsilon_i)^2 \\
&= \sum_{i<k} (\sigma + \delta_i)^2 + \sum_{i \geq k} (p_i - \alpha_i)^2 - \sum_{i<k} \sigma^2 - \sum_{i \geq k} p_i^2 \\
&= \sum_{i<k} \sigma^2 + 2\sigma\delta_i + \delta_i^2 + \sum_{i \geq k} p_i^2 - 2p_i\alpha_i + \alpha_i^2 - (k-1)\sigma^2 - \sum_{i \geq k} p_i^2 \\
&= 2\sigma \sum_{i<k} \delta_i - 2 \sum_{i \geq k} p_i \alpha_i + \sum_{i<k} \delta_i^2 + \sum_{i \geq k} \alpha_i^2 \\
&\geq 2\sigma \sum_{i<k} \delta_i - 2\sigma \sum_{i \geq k} \alpha_i + \sum_{i<k} \delta_i^2 + \sum_{i \geq k} \alpha_i^2 \\
&\geq \sum_{i<k} \delta_i^2 + \sum_{i \geq k} \alpha_i^2 \\
&\geq 0.
\end{aligned}$$

The first inequality follows from $p_i \leq \sigma$ for all $i \geq k$, and the second one from

$$\sum_{i<k} \delta_i = \sum_{i<k} p_i - \sigma - \alpha_i = \epsilon - \sum_{i<k} \alpha_i = \epsilon - \alpha + \sum_{i \geq k} \alpha_i \geq \sum_{i \geq k} \alpha_i$$

because of $\alpha \leq \epsilon$. The inequality above is equivalent to

$$H_2(X|U=0,V=0) \;\leq\; H_2(X|U=0,S=0)$$

in contradiction with the assumption. □

4.5 Bounds Using Spoiling Knowledge

Spoiling knowledge that gives the best lower bounds on smooth entropy has been characterized in the last section. Unfortunately, these characterizations do not translate directly into simple bounds on smooth entropy. Such bounds can, however, be derived using non-optimal side information. This is the subject of the present section. The first bound shows that smooth entropy is lower bounded asymptotically by Rényi entropy of order α for any $\alpha > 1$, and the second bound links smooth entropy with Shannon entropy if some assumptions about the profile (defined below) of the distribution are made.

For the construction of the lower bounds we introduce special side information U with alphabet $\{0, \ldots, m\}$ that partitions the values of X into sets of values with approximately equal probability. Let $U = f(X)$ be the deterministic function of X given by

$$f(x) \;=\; \begin{cases} m & \text{if } P_X(x) \leq 2^{-m} \\ \lfloor -\log P_X(x) \rfloor & \text{otherwise.} \end{cases}$$

We call side information U of this type *log-partition spoiling knowledge* because U partitions the values of X into sets of approximately equal probability and because it is most useful with $m \approx \log |\mathcal{X}|$. For such m, the values of the probability distributions $P_{X|U=u}$ differ at most by a factor of two for all u except for $u = m$.

In the following, let

$$p_{\min} = \min_{x \in \mathcal{X}} P_X(x) \qquad \text{and} \qquad p_{\max} = \max_{x \in \mathcal{X}} P_X(x).$$

The following two lemmas show that Rényi entropy of order 2 and Shannon entropy cannot differ arbitrarily for probability distributions where $p_{\min}$ and $p_{\max}$ are a constant factor apart.

Lemma 4.14. *Let X be a random variable with alphabet $\mathcal{X}$ such that $p_{\max} \leq c \cdot p_{\min}$ for some $c > 1$. Then*

$$\begin{aligned} \frac{1}{|\mathcal{X}| - 1 + c} \;\leq\; & p_{\min} \;\leq\; \frac{1}{|\mathcal{X}|} \\ \frac{1}{|\mathcal{X}|} \;\leq\; & p_{\max} \;\leq\; \frac{c}{|\mathcal{X}| - 1 + c}. \end{aligned}$$

Proof. It is easy to see that maximum of $p_{\max} - p_{\min}$ is reached when $P_X(x) = p_{\min}$ for all x except for the one that has maximal probability $p_{\max} = c \cdot p_{\min}$. The lemma follows directly. □

If the minimum and maximum probability in a distribution P_X do not differ by more than a constant factor, then the Rényi entropy of order 2 of X is at most a constant below the Shannon entropy.

Lemma 4.15. *Let X be a random variable with alphabet $\mathcal{X}$ such that $p_{\max} \leq c \cdot p_{\min}$ for some $c > 1$. Then*

$$H_2(X) \;>\; H(X) - 2\log c.$$

Proof. Lemma 4.14 is used in the second inequality of the following derivation:

$$\begin{aligned}
H(X) - H_2(X) &= H(X) + \log \sum_{x\in\mathcal{X}} P_X(x)^2 \\
&\leq \log|\mathcal{X}| + \log\bigl(|\mathcal{X}|\, p_{\max}^2\bigr) \\
&= 2\log\bigl(|\mathcal{X}|\, p_{\max}\bigr) \\
&\leq 2\log\Bigl(|\mathcal{X}|\, \frac{c}{|\mathcal{X}| - 1 + c}\Bigr) \\
&= 2\Bigl(\log c + \log\Bigl(\frac{|\mathcal{X}|}{|\mathcal{X}| - 1 + c}\Bigr)\Bigr) \\
&< 2\log c
\end{aligned}$$ □

4.5.1 A Bound Using Rényi Entropy of Order $\alpha > 1$

The connection between entropy smoothing and Rényi entropy was established independently by Bennett et al. [BBCM95] and Impagliazzo et al. [ILL89]. Their results (recited as Theorem 2.7 and Theorem 4.2, respectively) show that Rényi entropy of order 2 is a lower bound for smooth entropy. That is, for any random variable X by assuming only a lower bound t on $H_2(X)$, approximately t almost uniform random bits can be extracted from X and the deviation from a uniform distribution decreases exponentially when fewer bits are extracted.

In some applications, only the stronger bound $H_\infty(X) \geq t$ in terms of min-entropy is assumed, equivalent to bounding the maximum probability of any value of X. Indeed, the smoothing results (Theorems 2.7 and 4.2) hold if an assumption about $H_\alpha(X)$ for any $\alpha \geq 2$ is made because $H_2(X) \geq H_\alpha(X)$ for $\alpha \geq 2$ by Proposition 2.4.

On the other hand, it follows from Example 4.4 (on page 61) that a lower bound on $H_1(X) = H(X)$ is not sufficient to guarantee a nontrivial amount of smooth entropy. The smooth entropy could be arbitrarily small if no further assumptions are made. In this section we examine the remaining range for $1 < \alpha < 2$. We show that, with high probability, the smooth entropy of X is lower bounded by $H_\alpha(X)$, up to the logarithm of the alphabet size and some security parameters depending on α and on the error probability.

Our approach uses a spoiling knowledge argument. We will use side information U such that for any distribution of X, with high probability, U takes on a value u for which $H_2(X|U = u)$ is not far below $H_\alpha(X)$. A simple and very weak bound that always holds follows from the next lemma.

Lemma 4.16. *For any random variable X and for any $\alpha > 1$,*

$$\frac{\alpha}{\alpha - 1} H_\infty(X) \;\geq\; H_\alpha(X) \;\geq\; H_\infty(X).$$

Proof. Because $\alpha > 1$,

$$\begin{aligned}
\frac{\alpha}{\alpha - 1} H_\infty(X) &= \frac{1}{1-\alpha} \log \max_{x \in \mathcal{X}} P_X(x)^\alpha \\
&\geq \frac{1}{1-\alpha} \log \sum_{x \in \mathcal{X}} P_X(x)^\alpha \\
&= H_\alpha(X).
\end{aligned}$$

The lower bound follows from (2.14). □

We conclude that

$$H_2(X) \;\geq\; H_\infty(X) \;\geq\; \frac{\alpha - 1}{\alpha} H_\alpha(X)$$

for any $\alpha > 1$. However, this bound is multiplicative in $\alpha - 1$ which limits its usefulness for $\alpha \to 1$. The tighter bound derived below is only additive in $(\alpha - 1)^{-1}$. It is based on the following theorem which provides the connection between the Rényi entropy of order $\alpha > 1$ conditioned on side information and the Rényi entropy of the joint distribution.

Theorem 4.17. *Let $\alpha > 1$ and let $r, t > 0$. For arbitrary random variables X and Y, the probability that Y takes on a value y for which*

$$H_\alpha(X|Y = y) \;\geq\; H_\alpha(XY) - \log |\mathcal{Y}| - \frac{r}{\alpha - 1} - t$$

is at least $1 - 2^{-r} - 2^{-t}$.

Proof. It is straightforward to expand the Rényi entropy of XY as

$$\begin{aligned} H_\alpha(XY) &= \frac{1}{1-\alpha} \log \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} P_{XY}(x,y)^\alpha \\ &= \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} P_Y(y) \cdot P_Y(y)^{\alpha-1} \sum_{x\in\mathcal{X}} P_{X|Y=y}(x)^\alpha \\ &= \frac{1}{1-\alpha} \log \sum_{y\in\mathcal{Y}} P_Y(y)\, 2^{(\alpha-1)\log P_Y(y) + (1-\alpha) H_\alpha(X|Y=y)}. \end{aligned}$$

We introduce the function $\beta(y) = H_\alpha(X|Y=y)$ to interpret $H_\alpha(X|Y=y)$ as a function of y and consider the random variables $P_Y(Y)$ and $\beta(Y)$. The equation above is equivalent to

$$\mathrm{E}_Y\left[\, 2^{(1-\alpha)\beta(Y) + (\alpha-1)\log P_Y(Y)} \,\right] \;=\; 2^{(1-\alpha)H_\alpha(XY)}$$

or

$$\mathrm{E}_Y\left[\, 2^{(1-\alpha)\beta(Y) + (\alpha-1)\log P_Y(Y) - (1-\alpha)H_\alpha(XY) - r} \,\right] \;=\; 2^{-r}.$$

Inserting this into the right-hand side of inequality (2.7) yields

$$\mathrm{P}_Y\left[\, (1-\alpha)\beta(Y) + (\alpha-1)\log P_Y(Y) - (1-\alpha)H_\alpha(XY) \;\geq\; r \,\right] \;\leq\; 2^{-r}$$

from which we see, after dividing by $1-\alpha$, that with probability at least $1-2^{-r}$, Y takes on a value y for which

$$H_\alpha(X|Y=y) \;\geq\; H_\alpha(XY) + \log P_Y(y) - \frac{r}{\alpha-1}. \tag{4.17}$$

The only thing missing is a bound for the term $\log P_Y(y)$. However, large values of $|\log P_Y(Y)|$ occur only with small probability. For any $t > 0$,

$$\mathrm{P}\big[\, P_Y(Y) \;<\; 2^{-t}/|\mathcal{Y}| \,\big] \;=\; \sum_{y\,:\,P_Y(y)\,<\,2^{-t}/|\mathcal{Y}|} P_Y(y) \;<\; 2^{-t}$$

because there are only $|\mathcal{Y}|$ terms in the summation. Therefore, with probability at least $1-2^{-t}$, Y takes on a value y for which

$$\log P_Y(y) \;\geq\; -t - \log|\mathcal{Y}| \tag{4.18}$$

and the theorem follows from (4.17) and (4.18) by the union bound. $\square$

Applying this bound for log-partition side information gives the main result of this section and shows how smooth entropy is lower bounded by Rényi entropy of order α for any $\alpha > 1$.

Theorem 4.18. *Fix $r, t > 0$, let m be an integer such that $m - \log(m+1) > \log|\mathcal{X}| + t$, and let s be the security parameter for smooth entropy. For any $\alpha > 1$, the smooth entropy of a random variable X within $2^{-s}/\ln 2$ in terms of relative entropy with probability $1 - 2^{-r} - 2^{-t}$ is lower bounded by Rényi entropy of order α in the sense that*

$$\Psi(X) \;\geq\; H_\alpha(X) - \log(m+1) - \frac{r}{\alpha - 1} - t - 2.$$

Proof. We use log-partition spoiling-knowledge $U = f(X)$ with alphabet $\{0, \ldots, m\}$ as defined above. Because f is a deterministic function of X, we have $H_\alpha(XU) = H_\alpha(X)$ and Theorem 4.17 shows that U takes on a value u for which

$$H_\alpha(X|U = u) \;\geq\; H_\alpha(X) - \log|\mathcal{U}| - \frac{r}{\alpha - 1} - t$$

with probability at least $1 - 2^{-r} - 2^{-t}$. Because $m > \log|\mathcal{X}|$, Lemma 4.15 can be applied with $c \leq 2$ and, by (2.14), it follows for all $u \neq m$ that

$$H_2(X|U = u) \;>\; H(X|U = u) - 2 \;\geq\; H_\alpha(X|U = u) - 2.$$

Combining these results shows that the probability that U takes on a value $u \neq m$ for which

$$H_2(X|U = u) \;\geq\; H_\alpha(X) - \log(m+1) - \frac{r}{\alpha - 1} - t - 2 \qquad (4.19)$$

is at least $1 - 2^{-r} - 2^{-t}$.

Recall that in (4.18) in the proof of Theorem 4.17, values of U with probability less than $2^{-t-\log|\mathcal{U}|}$ have been excluded. Therefore, if m is chosen such that

$$\mathrm{P}[U = m] \;=\; \sum_{x \,:\, P_X(x) \,<\, 2^{-m}} P_X(x) \;\leq\; |\mathcal{X}| \cdot 2^{-m} \;<\; 2^{-t-\log|\mathcal{U}|}$$

then $U = m$ does not occur in (4.19). Choosing m such that $m - \log(m+1) > \log|\mathcal{X}| + t$ achieves this and applying Theorem 4.12 completes the proof. □

Corollary 4.19. *Let $\mathbb{X}$ be a family of random variables and let r, t, m, and s be defined as in the theorem above. For any $\alpha > 1$, the smooth entropy of $\mathbb{X}$ within $2^{-s}/\ln 2$ in terms of relative entropy with probability $1 - 2^{-r} - 2^{-t}$ satisfies*

$$\Psi(\mathbb{X}) \geq \min_{X \in \mathbb{X}} H_\alpha(X) - \log(m+1) - \frac{r}{\alpha - 1} - t - 2.$$

The corollary follows from the fact that the oracle knows the distribution of the random variable $X \in \mathbb{X}$ to be smoothed and can prepare the side information accordingly. Especially for large alphabets, these results can yield much better bounds on smooth entropy than Rényi entropy of order 2. The logarithmic term vanishes asymptotically with the alphabet size: For any $\alpha > 1$, the ratio between smooth entropy and the logarithm of the alphabet size is asymptotically lower bounded by the ratio between Rényi entropy of order α and the logarithm of the alphabet size.

Example 4.6. Consider the random variables X_β with alphabet $\{0,1\}^n$ and distribution

$$P_{X_\beta}(x) = \begin{cases} 2^{-n/(2\beta)} & \text{for } x = 0^n \\ \frac{1-2^{-n/(2\beta)}}{2^n - 1} & \text{otherwise} \end{cases}$$

for $\beta \ll n$. (With $\beta = 2$ this is the example from [BBCM95].) The lower bound on $\Psi(X)$ by Rényi entropy of order 2 is weak because $H_2(X) < n/\beta$. However, $H(X_\beta)$ is very close to n bits. Figure 4.3 displays the Rényi entropy $H_\alpha(X_\beta)$ for $1 \leq \alpha \leq 2$. For α close to 1, it is almost equal to $H(X_\beta) \approx n$.

Using Rényi entropy of order 2, Corollary 4.1 shows that $\Psi(X_8)$ within $2^{-s}/\ln 2$ with probability 1 is at least $H_2(X_8) \approx n/8$. Allowing failure of the bound with probability 2^{-19}, the lower bound by Theorem 4.18 on $\Psi(X_8)$ with probability $1 - 2^{-19}$ is about $n - \log n - 222$ (using Rényi entropy of order $\alpha = 1.1$, $r = t = 20$, and simplifying the choice of m such that $m = \log|\mathcal{X}| = n$). With $n = 10000$ (as in Figure 4.3), $\Psi(X_8) \geq 9764$ with probability $1 - 2^{-19}$, compared to Rényi entropy of order 2 from which we can conclude only $\Psi(X_8) \geq 1250$. ○

For $\alpha \to 1$, the bound of Theorem 4.18 is reduced to the Shannon entropy. But as shown in Example 4.4 in Section 4.3.3, $H(X)$ yields a weak lower bound for $\Psi(X)$. The next example shows this transition for $\alpha \to 1$.

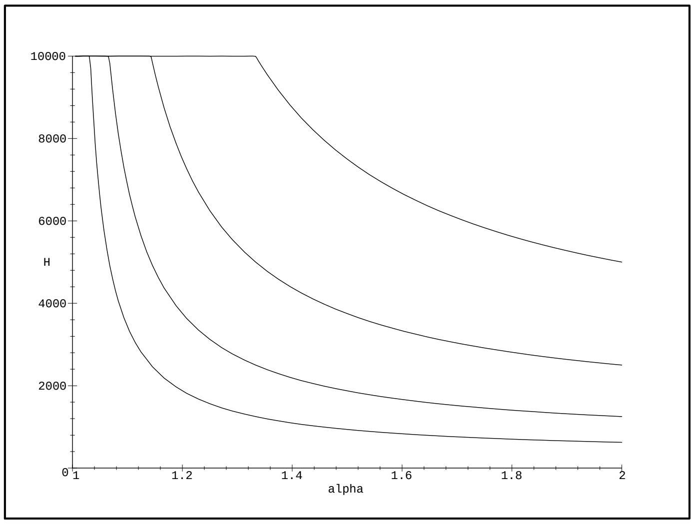

Figure 4.3. Rényi entropy $H_\alpha(X_\beta)$ as function of α between 1 and 2. The random variables X_β for $\beta = 16, 8, 4, 2$ (from below) are defined as in Example 4.6 with $n = 10000$. The graph shows that, together with Theorem 4.18, Rényi entropy of order α close to 1 can yield much better bounds on smooth entropy than Rényi entropy of order 2.

Example 4.7. Let X be a random variable with alphabet $\{0,1\}^{10000}$. We now examine the lower bounds on $\Psi(X)$ when $H_\alpha(X) \geq 9000$ is assumed for various α (see Figure 4.4). For $\alpha \geq 2$, $\Psi(X) \geq H_2(X) \geq 9000$ is guaranteed by Corollary 4.1. Theorem 4.18 shows that $\Psi(X)$ with probability $1 - 2^{-19}$ is close to 9000 for α between 2 and about 1.05. The bound decreases sharply with $\alpha \to 1$. For $\alpha = 1$, if only $H(X) \geq 9000$ is assumed, the random variable constructed in Example 4.4 has $H_2(X) = 6.64$ and has almost no smooth entropy. ○

4.5.2 A Tighter Bound Using the Profile of the Distribution

The last section has shown how smooth entropy can be lower bounded by Rényi entropy of order α for any $\alpha > 1$. This bound, however, is

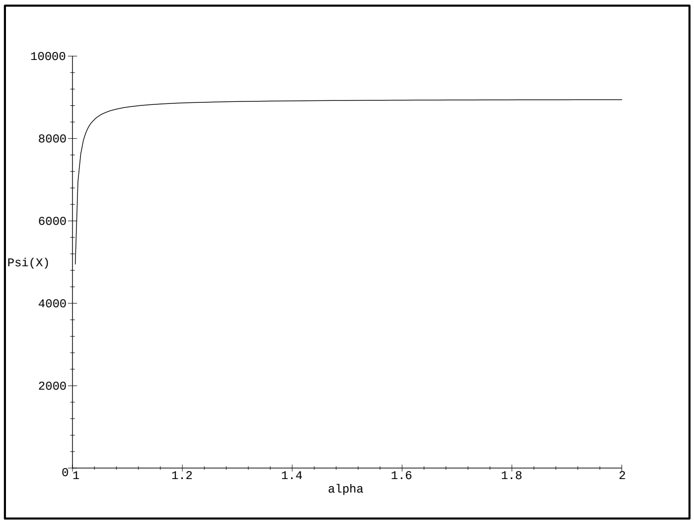

Figure 4.4. The dependence of the lower bound for $\Psi(X)$ on the order α of Rényi entropy. The graph shows the lower bound of Theorem 4.18 on the smooth entropy $\Psi(X)$ within $2^{-s}/\ln 2$ with probability $1 - 2^{19}$ that can be deduced from $H_\alpha(X) \geq 9000$ as a function of α. Note the sharp decrease with $\alpha \to 1$. (See also Example 4.7.)

not tight for small alphabet sizes. We derive a tighter bound in this section that depends on an assumption about the profile of the probability distribution (defined below). The bound is tighter than the one of Theorem 4.18, especially for smaller alphabets.

We use again log-partition spoiling knowledge $U \in \mathcal{U} = \{0, \ldots, m\}$ as defined above. For a fixed value m, define the *profile* π_X of the random variable X as the function $\pi_X : \mathcal{U} \to \mathbb{N}$ such that for $u < m$

$$\pi_X(u) \;=\; \Big|\big\{x \in \mathcal{X} \,|\, 2^{-u-1} < P_X(x) \leq 2^{-u}\big\}\Big|$$

and

$$\pi_X(m) \;=\; \Big|\big\{x \in \mathcal{X} \mid P_X(x) \leq 2^{-m}\big\}\Big|.$$

The expected difference (over U) between the logarithm of the profile $\pi_X(u)$ and the conditional entropy of X given U, $H(X|U = u)$, can be used to obtain a lower bound on smooth entropy. Examining the structure of the probability distributions $P_{X|U=u}$ for all u such that $\pi_X(u) \geq 2$, we see that the logarithm of the profile, $\pi_X(u)$, is close to the conditional entropy, $H(X|U=u)$, in the sense that

$$\log \pi_X(u) \;\geq\; H(X|U=u) \;\geq\; h\Big(\frac{2}{\pi_X(u)+1}\Big) + \log\big(\pi_X(u)-1\big). \tag{4.20}$$

(h denotes the binary entropy function.) We note that $H(X|U=u)=0$ for the remaining u with $\pi_X(u) < 2$. Therefore,

$$\mathrm{E}_U\Big[\log \pi_X(U)\Big] \;\geq\; H(X|U) \;\geq\; \mathrm{E}_U\Big[\log\big(\pi_X(U)-1\big)\Big]. \tag{4.21}$$

We are now ready to state the main result of this section.

Theorem 4.20. *Let X be a random variable, let $\epsilon > 0$, let m be an integer such that $m \geq \log|\mathcal{X}| + \log\frac{1}{\epsilon}$, let $t > 0$, and let k be a positive integer. Let U be the log-partition side information for X introduced above and let*

$$\mu(u) \;=\; \max\Big\{ \log \pi_X(u) - \mathrm{E}_U\Big[\log\big(\pi_X(U)-1\big)\Big], \\ \mathrm{E}_U\Big[\log \pi_X(U)\Big] - \log\big(\pi_X(u)-1\big) \Big\}$$

for all u such that $\pi_X(u) \geq 2$ and $\mu(u) = \mathrm{E}_U[\log \pi_X(U)]$ for u such that $\pi_X(u) < 2$. If

$$\mathrm{E}_U\big[\mu(U)^k\big] \;\leq\; \epsilon \cdot t^k,$$

the following lower bound on the smooth entropy of X within $2^{-s}/\ln 2$ in terms of relative entropy holds with probability at least $1-2\epsilon$:

$$\Psi(X) \;\geq\; H(X|U) - t - 2 \;\geq\; H(X) - \log(m+1) - t - 2.$$

Proof. Let $\gamma(u) = H(X|U=u)$ be a function of $u \in \mathcal{U}$ that denotes the entropy of X given $U=u$ and consider the random variable $C = \gamma(U)$. The expectation $\mathrm{E}[C]$ is equal to $H(X|U) \geq H(X) - \log(m+1)$. Applying the k-th moment inequality (2.4), we see that

$$\mathrm{P}\big[|C - E[C]| \geq t\big] \;\leq\; \frac{\mathrm{E}\big[|C - E[C]|^k\big]}{t^k}. \tag{4.22}$$

If this probability is small, then $H(X|U = u) \geq H(X|U) - t$ with high probability. Using (4.20) and (4.21), we can bound the probability in (4.22):

$$
\begin{aligned}
&\mathrm{E}\Big[\big|C - E[C]\big|^k\Big] \\
&= \sum_{u\in\mathcal{U}} P_U(u)\big|H(X|U=u) - H(X|U)\big|^k \\
&= \sum_{u\,:\,\pi_X(u)<2} P_U(u) H(X|U)^k + \\
&\quad \sum_{u\,:\,H(X|U=u)>H(X|U)} P_U(u)\Big(H(X|U=u) - H(X|U)\Big)^k + \\
&\quad \sum_{u\,:\,H(X|U=u)<H(X|U)} P_U(u)\Big(H(X|U) - H(X|U=u)\Big)^k \\
&\leq \sum_{u\,:\,\pi_X(u)<2} P_U(u) H(X|U)^k + \\
&\quad \sum_{u\,:\,H(X|U=u)>H(X|U)} P_U(u)\Big(\log\pi_X(u) - \mathrm{E}_U\left[\log(\pi_X(U)-1)\right]\Big)^k + \\
&\quad \sum_{u\,:\,H(X|U=u)<H(X|U)} P_U(u)\Big(\mathrm{E}_U\left[\log\pi_X(U)\right] - \log\big(\pi_X(u)-1\big)\Big)^k \\
&= \sum_{u\in\mathcal{U}} P_U(u)\mu(u)^k
\end{aligned}
$$

where the last step follows form the definition of $\mu(u)$. We conclude from (4.22) and from the assumption of the theorem that $H(X|U = u) \geq H(X|U) - t$ occurs with probability at least $1 - \epsilon$. It follows from Lemma 4.15 that for $u \neq m$

$$H_2(X|U = u) \;\geq\; H(X|U) - t - 2. \tag{4.23}$$

But the event $U = m$ has small probability because the choice of m guarantees that

$$\mathrm{P}[U = m] \;=\; \sum_{x\,:\,P_X(x)<2^{-m}} P_X(x) \;\leq\; |\mathcal{X}|\cdot 2^{-m} \;\leq\; \epsilon.$$

By the union bound, the total probability that (4.23) fails is at most 2ϵ and the proof is completed by applying Theorem 4.12. □

Example 4.8. Consider again the random variable X_8 from Example 4.6 (page 84). For $n = 100$ and desired total failure probability 2^{-19}, the bound of Theorem 4.18 cannot be applied and we have to resort to Rényi entropy of order 2 that shows $\Psi(X_8) \geq 12.5$ (within $2^{-s}/\ln 2$ in terms of relative entropy).

Applying Theorem 4.20 with $\epsilon = 2^{-20}$, $t = 12$, and $k = 6$, however, shows that $\Psi(X_8) \geq 84.6$. Therefore, a 60-bit string Y can be extracted from X_8 by a randomly chosen universal hash function such that $H(Y|T) \geq 60 - 2^{-24}/\ln 2$. ○

As the example shows, the bound on smooth entropy by Theorem 4.20 can be much tighter than Rényi entropy of order 2 and also tighter than the bound of Theorem 4.18. However, this comes at the cost of the stronger assumption that must be made in terms of the profile of the distribution to be smoothed.

4.6 Smoothing an Unknown Distribution

The discussion in this chapter has concentrated on the smooth entropy of a random variable X with given distribution. The distribution of X was either given or at least known to be one of many given distributions of a family $\mathbb{X}$. What happens if a universal hash function is applied to X if its distribution is not known? Is it possible to quantify the ignorance about the bits extracted from a random variable with unknown probability distribution? In this section, we answer this question positively. We show that if P_S is the assumed distribution of a random variable X and P_X is its true distribution, the amount of uniformly looking randomness that can be extracted corresponds to $-\log \mathrm{P}[X = S]$, where X and S are interpreted as independent random variables in the same probability space.

The uncertainty about the outcome of a random variable with unknown distribution consists of two parts that can be related to the uncertainty of the random variable itself and to the lack of knowledge about the correct distribution. More specifically, if it is assumed that the distribution of a random variable X is P_S, the ignorance about X is characterized by $H(X) + D(P_X\|P_S)$ which is called the *inaccuracy* [CK81]. The inaccuracy consists of two additive parts: $D(P_X\|P_S)$ is due to the ignorance of the correct distribution and $H(X)$ is due to the uncertainty of X itself.

Inaccuracy measures the lack of knowledge about the correct distribution similar to entropy [CT91]. Consider the construction of an optimal binary prefix-free code for a random variable X. If the distribution of X is known, the optimal code has average length between $H(X)$ and $H(X) + 1$ which is one justification of entropy as a fundamental measure of uncertainty. Thus, the expected number of arbitrary binary questions needed to describe an outcome of X is at least $H(X)$. The "guessing entropy" discussed in Chapter 3 refers to a different guessing scenario where only questions of the form "$X = x$" are allowed.

If P_X is unknown and a code optimal for a distribution P_S is used, the average length of the code used for X lies between $H(X) + D(P_X\|P_S)$ and $H(X) + D(P_X\|P_S) + 1$. Thus, the expected number (over X) of binary questions needed to describe an outcome of X is at least $H(X)+ D(P_X\|P_S)$. It is clear that $\lceil \log|\mathcal{X}|\rceil$ questions are always sufficient by assuming the uniform distribution over $\mathcal{X}$. This corresponds to (2.12).

We now extend Theorem 2.7 by considering universal hashing of a random variable X with unknown distribution. We show that the ignorance about the output of this process depends only on the probability that a value of X can be guessed correctly. More precisely, when S denotes a random variable with the distribution we assume for X and when a shorter value Y is extracted from X by universal hashing, the size of the largest Y with ignorance close to the maximum, $\log|\mathcal{Y}|$, is characterized by $-\log \mathrm{P}[X = S]$.

Theorem 4.21. *Let X and S be independent random variables over the same alphabet $\mathcal{X}$ with probability distributions P_X and P_S, respectively, and let G be the random variable corresponding to the random choice (with uniform distribution) of a member of a universal hash function $\mathcal{G} : \mathcal{X} \to \mathcal{Y}$, and let $Y = G(X)$ and $Z = G(S)$. Then*

$$H(Y|G) + D(P_{Y|G}\|P_{Z|G}) \;\geq\; \log|\mathcal{Y}| - \frac{|\mathcal{Y}| \cdot \mathrm{P}[X = S]}{\ln 2}.$$

Proof. Expanding the definition of conditional relative entropy (2.11), we obtain

$$\begin{aligned}
&D(P_{G(X)|G}\|P_{G(S)|G}) \\
&\quad= \sum_{g\in\mathcal{G}} P_G(g) \sum_{y\in\mathcal{Y}} P_{G(X)|G=g}(y) \log \frac{P_{G(X)|G=g}(y)}{P_{G(S)|G=g}(y)} \\
&\quad= \sum_{g\in\mathcal{G}} P_G(g) \sum_{y\in\mathcal{Y}} P_{G(X)|G=g}(y) \log P_{G(X)|G=g}(y)
\end{aligned}$$

$$
\begin{aligned}
& - \sum_{g \in \mathcal{G}, y \in \mathcal{Y}} P_G(g) P_{G(X)|G=g}(y) \log P_{G(S)|G=g}(y) \\
\stackrel{(a)}{\geq}\ & -H(G(X)|G) - \log \sum_{g \in \mathcal{G}, y \in \mathcal{Y}} P_G(g) P_{G(X)|G=g}(y) P_{G(S)|G=g}(y) \\
\stackrel{(b)}{=}\ & -H(G(X)|G) - \log \mathrm{P}[G(X) = G(S)] \\
=\ & -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + \\
& \qquad \mathrm{P}[X \neq S] \cdot \mathrm{P}[G(X) = G(S)|X \neq S]\Big) \\
\stackrel{(c)}{\geq}\ & -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + \mathrm{P}[X \neq S] \cdot |\mathcal{Y}|^{-1}\Big) \\
>\ & -H(G(X)|G) - \log\Big(\mathrm{P}[X = S] + |\mathcal{Y}|^{-1}\Big) \\
=\ & -H(G(X)|G) - \log\Big(|\mathcal{Y}|^{-1}(1 + |\mathcal{Y}| \cdot \mathrm{P}[X = S])\Big).
\end{aligned}
$$

The inequality (a) follows from the application of the Jensen inequality (2.2) to the negative sum. In step (b), the argument of the logarithm corresponds to the probability that $g(X) = g(S)$ if X and S are independent of each other and of G, which is selected randomly according to P_G. The second inequality (c) is a consequence of the universality of the hash function $\mathcal{G}$ from which G is chosen with uniform probability. Because the logarithms are to base 2, the theorem follows from the last expression and the inequality $\log(1 + x) \leq x / \ln 2$. □

Theorem 4.21 is a strict generalization of Theorem 2.7, because $2^{-H_2(X)} = \mathrm{P}[X_1 = X_2] = \sum_{x \in \mathcal{X}} P_X(x)^2$ (when X_1 and X_2 are independent random variables with distributions equal to P_X) and $D(P_X \| P_S) = 0$ if and only if $P_X(x) = P_S(x)$ for all $x \in \mathcal{X}$. In the other extreme case, if nothing is known about P_X except the size of the alphabet, then the uniform distribution has to be assumed and applying a hash function is unnecessary.

Between these extreme cases are those situations in which P_X is unknown and the assumed distribution of X is P_S. To find a value of $Y = G(X)$ by a series of binary questions, a code optimal for $Z = G(S)$ is used and the theorem shows that the average number of questions needed is close to the maximum $\log |\mathcal{Y}|$ whenever $\mathrm{P}[X = S] < |\mathcal{Y}|^{-1}$. Moreover, the average number of questions can be made arbitrarily close to the maximum by decreasing the size of $\mathcal{Y}$.

Example 4.9. Consider a random variable X with alphabet $\{1, \ldots, 100\}$ and distribution $P_X(i) = 1.01/(i(i+1))$. Let P_S be the assumed distribution of X such that $P_S(i) = P_X(101 - i)$ for $i = 1, \ldots, 100$. Thus, the probabilities are assumed to be exactly in reverse order. We note that $H(X) = 2.81$ and $H(X) + D(P_X \| P_Y) = 10.36$. The Rényi entropy of order 2 is $H_2(X) = 1.76$, but the guessing probability using P_S satisfies $-\log \mathrm{P}[X = S] = 12.20$.

Let $\mathcal{G}$ be a universal hash function from $\mathcal{X}$ to $\{0, 1\}$. If the distribution of X was known, the average number of binary questions to determine $G(X)$ is $H(G(X)|G) \geq 0.147$ according to Theorem 2.7. But since the distribution of X is assumed to be P_S, at least $H(G(X)|G) + D(P_{G(X)|G} \| P_{G(S)|G}) \geq 0.999$ are needed on the average as demonstrated by Theorem 4.21. ○

This result shows how smoothing by universal hashing extends to random variables with unknown distributions. In learning theory, for example, such a situation is conceivable if nothing is known about the distribution of learning examples (the target distribution, see Section 4.3.5). However, extending the notion of smooth entropy to the effect of allowing unknown distributions is most likely in vain. Applications of entropy smoothing usually require X to satisfy a certain condition (e.g. characterizing the information that an adversary has). Thus, any X satisfying the condition can be assumed and distinguishing between unknown and assumed distributions is meaningless.

4.7 Conclusions

In this chapter, the concept of smooth entropy has been introduced to quantify the number of uniform bits that can be extracted from a random source by probabilistic algorithms. Smooth entropy unifies work on privacy amplification in cryptography and on entropy smoothing in complexity theory. It turns out that the notion of smooth entropy falls between intrinsic randomness, a formalization of the uniform entropy extractable by deterministic functions, and the concept of extractors, which differs in the respect that it also takes into account the number of auxiliary random bits.

The formalization of smooth entropy allows a systematic investigation of the spoiling knowledge proof technique for obtaining lower bounds on extractable uniform randomness. Through the application of a special type of spoiling knowledge, we were able to show that smooth entropy is lower bounded by Rényi entropy of order α for any $\alpha > 1$. This closes the gap for values of α between 1 and 2. It was previously only known that Rényi entropy of order 2 (and higher) is a lower bound and that no such statement is possible for Rényi entropy of order 1.

Our results can be applied to all scenarios using entropy smoothing and, in particular, to all applications of privacy amplification in cryptography. Our analysis shows that entropy smoothing by universal hashing is generally much more efficient than what was known from previous results using Rényi entropy of order 2.

Several questions with respect to smooth entropy remain open. One of them is the existence of different smoothing functions. Are there other functions that could extract more randomness than universal hashing? Extractors, in general, consist of composed universal hash functions and other steps that, taken together, do not achieve better parameters than universal hashing alone (except for the size of the auxiliary input).

Smooth entropy has been investigated for relative entropy and L_1 distance as nonuniformity measures. We do not know whether smooth entropy is substantially different for other measures. However, there is some evidence that this is not the case because most nonuniformity measures can be linked with general bounds (see Section 3.3). Furthermore, the investigation of intrinsic randomness shows that the intrinsic randomness rate is the same for three different nonuniformity measures [VV95].

Smooth entropy could also have been formalized asymptotically in terms of a "smooth entropy rate," in the tradition of information theory and similar to the intrinsic randomness rate. Benefits of such a formulation would be simpler statements of the results and the use of asymptotic effects. However, we have explicitly chosen a non-asymptotic treatment because of the focus on applications.

Chapter 5

Unconditional Security in Cryptography

5.1 Introduction

One of the most important properties of a cryptographic system is a proof of its security under reasonable and general assumptions. However, every design involves a trade-off between the level of security and further important qualities of a cryptosystem, such as efficiency and practicality.

The security models currently used in cryptography include *computational security*, *provable computational security*, and *unconditional security* in the terminology of Menezes et al. [MvOV97].

The notion of *computational security* is based on the amount of computational work required to break a system by the best currently known methods. Discovering the relevant attacks requires a thorough examination and, as yet, no design can be guaranteed to remain secure in the future. Computational security is equivalent to the *historical difficulty* of a problem and is likely to decrease with the development of new cryptanalytic techniques.

A cryptosystem is *computationally secure* if the amount of work needed to break it exceeds the computational resources of an adversary by a substantial margin. Most of the currently used public-key cryptosystems, as well as private-key systems, fall into this category. Many public-key schemes and advanced protocols (e.g. voting or electronic payment systems) are based on the difficulty of an intractable computational problem, such as factoring large numbers or computing discrete logarithms, but their security cannot be proved to be equivalent to solving the underlying problem.

In contrast, cryptosystems with *provable computational security* are based on security proofs showing that the ability of the adversary to defeat the cryptosystem with significant probability contradicts the *supposed* intractability of the underlying problem [Riv90]. Thus, provable computational security is always conditional on a security assumption, typically in terms of a number-theoretic problem (e.g. factoring or discrete logarithm). Providing a proof for such an assumption in a sufficiently general model of computation and with a cryptographically relevant notion of solving the problem continues to be among the most difficult tasks in complexity theory. Provable computational security is nevertheless attractive, as the security assumption can be formalized and is concentrated at one specific location. At the same time, this concentration is a drawback because many cryptosystems rely on the same problems and because only a few suitable problems are known. Although the hardness of these problems is unquestioned at the moment, it can be dangerous to base the security of the global information economy on a very small number of mathematical problems. Recent advances in quantum computing show that precisely these two problems, factoring and discrete logarithm, could be solved efficiently if quantum computers could be built [Sho94].

An alternative to proofs in the computational security model is offered by the stronger notion of information-theoretic or *unconditional security* where no limits are imposed on an adversary's computational power. In addition, the security need not be based on intractability assumptions. The first information-theoretic definition of perfect secrecy by Shannon [Sha49] led immediately to his famous impracticality theorem, which states, roughly, that the shared secret key in any perfectly secure cryptosystem must be at least as long as the plaintext to be encrypted (see Section 3.2.1). Vernam's one-time pad is the prime example of a perfectly secure but impractical system. Unconditional security was therefore considered too expensive for a long time.

However, recent developments show how Shannon's model can be modified [Mas91] to make practical provably secure cryptosystems possible. The first modification is to relax the requirement that perfect security means complete independence between the plaintext and the adversary's knowledge and to allow an arbitrarily small correlation. The second, crucial modification removes the assumption that the adversary receives exactly the same information as the legitimate users. The following primitives are the most realistic mechanisms proposed so far for limiting the information available to the adversary.

Quantum Channel: Quantum cryptography was developed mainly by Bennett and Brassard during the 1980's [BBB$^+$92, BC96] and uses photons, i.e. polarized light pulses of very low intensity, that are transmitted over a fiber-optical channel. In the basic quantum key agreement protocol, this allows two parties to generate a secret key by communicating about the received values. The unconditional secrecy of the key is guaranteed by the uncertainty relation of quantum mechanics. Current implementations of quantum key distribution span distances of 20–30 km [MZG95]. For some time, it seemed also possible to realize bit commitment based on the security of the quantum channel [BCJL93], but recent results show that this is not the case [May96, Cré96].

Noisy Channel: The use of a noisy channel for cryptographic purposes was introduced by Wyner with the wiretap channel scenario [Wyn75]. A sender wants to transmit data secretly over a noisy channel to a receiver. The adversary (the wiretapper) views the channel output via a second noisy channel. Wyner's result shows that secret information transmission from the sender to the receiver is possible with unconditional security. In a later and perhaps more realistic model proposed by Maurer, the output of a random source is transmitted to the participants over partially independent noisy channels that insert errors with certain probabilities [Mau93]. Two parties can then generate a secret key from their received values by public discussion. The secrecy of the key is based on the information differences between the channel outputs and on the assumption that no channel is completely error-free. This system is practical because it works also in the realistic case where the adversary receives the random source via a much better channel than the legitimate users. The power of a noisy channel was also demonstrated by Crépeau and Kilian who showed that unconditionally secure bit commitment and oblivious transfer can be based on this primitive [CK89, Cré97].

Memory Bound: We showed [CM97c] how to realize unconditionally secure cryptosystems based on the assumption that the memory size of the adversary is limited (see Section 5.4). This means that

> an enemy can use unlimited computing power to compute any probabilistic function of some huge amount of public data, which is infeasible to store. As long as the function's output size does not exceed the number of available storage bits, we can prove from this sole assumption that the proposed secret-key system and public key agreement protocol are information-theoretically secure.

In this chapter we focus on the use of these mechanisms for realizing public key agreement and secret-key encryption. We start in Section 5.2 with an overview of unconditionally secure key agreement protocols that involve three phases, called advantage distillation, information reconciliation, and privacy amplification. Section 5.3 investigates the effect of side information that an adversary obtains during information reconciliation on privacy amplification. In Section 5.4, we propose unconditionally secure cryptosystems that are based only on the assumption that an adversary's memory capacity is bounded. The rest of the chapter is partially taken from [CM97a, CM97c].

5.2 Unconditionally Secure Key Agreement Protocols

The generation of a shared secret key by two partners Alice and Bob is one of the fundamental problems in cryptography. The partners do not share secret information at the beginning. We assume in this chapter that they are linked by an authenticated communication channel that is insecure in the sense that the passive adversary Eve can read but not modify messages.

In this scenario, key agreement with computational security can be realized with the Diffie-Hellman protocol [DH76] or the RSA public key system [RSA78] that are both widely used today. Key agreement with provable computational security is also possible, as was first demonstrated by the Goldwasser-Micali public-key encryption scheme [GM84]. But the focus of this chapter is on realizing unconditional security.

As discussed in the previous paragraphs, practical unconditional security can be achieved if some basic mechanism is used initially in such a way that Eve does not receive exactly the same information as do Alice and Bob. In general, this initialization phase gives the participants Alice, Bob, and Eve access to correlated random variables X, Y, and Z, respectively, distributed according to a joint distribution P_{XYZ}.

The distribution P_{XYZ} may be under partial control of Eve and details of it may be unknown to either Alice and Bob or to Eve. This is the case in quantum cryptography [BBB+92] for instance, where Eve can undertake partial measurements on the quantum channel on which Alice sends polarized photons to Bob. This disturbs the probability distribution of the values Y measured by Bob in an arbitrary way unknown to him and limited only by the laws of quantum mechanics. But these physical laws also guarantee that any measurement can be detected by Bob if it gives Eve more information than allowed by the protocol. On the other hand, it may be the case that Alice and Bob do not know the distribution of Eve's information Z about their common knowledge as in the privacy amplification scenario (see Section 2.6).

The goal of a key agreement protocol between Alice and Bob is to establish a secret key K by exchanging a series of messages on the public channel. Alice and Bob should be able to determine K from the public communication and from X and Y, respectively, whereas Eve should have no information about K from Z and the protocol messages exchanged. The protocol and the requirements are described below in more detail. To provide secret communication for Alice and Bob, K can be used directly as the key for one-time pad encryption (see page 27). Alternatively, K can be used by Alice and Bob for unconditionally secure message authentication (see Section 3.2.2).

We assume that the public channel between Alice and Bob is authenticated (Eve has only read access) and that the channel can be used in abundance. However, it can be difficult to guarantee unconditional authenticity in practical applications. In this case, we assume that Alice and Bob initially share a short secret key K_0 to be used for authentication. A public key agreement protocol is then a method for expanding K_0 to a key K of arbitrary length. Key agreement without authentic channels has been investigated by Maurer [Mau97] and is not pursued further here.

A key agreement protocol generally consists of three phases. Some phases may be missing for certain scenarios, as explained later.

Advantage Distillation [Mau93]: The purpose of the first phase is to create a random variable T about which both Alice and Bob have more information than Eve. Advantage distillation is only needed when such a T is not immediately available from X and Y. Alice and Bob create T by exchanging messages over the public channel that we denote by the random variable C. In terms of entropies, the goal of advantage distillation is to establish the conditions $H(T|CY) < H(T|CZ)$ and $H(T|CX) < H(T|CZ)$.

Information Reconciliation [BBB$^+$92, BS94]: To agree on a common string W, Alice and Bob exchange redundant error-correction information U over the public channel. U can be a sequence of parity checks from a systematic error-correcting code [Bla83]. Alternatively, Alice and Bob can carry out an interactive protocol, such as *Cascade* proposed by Brassard and Salvail [BS94]. Specialized protocols have a number of advantages compared to error-correcting codes, in particular because the communication on the public channel is always assumed to be error-free. After information reconciliation, Alice and Bob must be able to determine a common string W from U, C, and X or Y, respectively. Eve's (incomplete) information about W consists of Z, C, and U.

When Alice knows a string T at the beginning of reconciliation about which Bob has more information than Eve, Alice and Bob can choose W to be this string and transmit the missing information from Alice to Bob during reconciliation. In information-theoretic terms, $W = T$ is a random variable with $H(W|XC) = 0$ and $H(W|YC) < H(W|ZC)$. In this case, Bob tries to determine W from Y and the reconciliation string U. Reconciliation serves to establish $H(W|YCU) \approx 0$, while Eve still has a substantial amount of uncertainty about W, i.e. $H(W|ZCU) \gg 0$.

Privacy Amplification [BBCM95, BBR88]: In the final phase, Alice and Bob agree publicly on a compression function G to distill from W a shorter string K about which Eve has only a negligible amount of information. In terms of entropy, $H(K)$ should be as large as possible, Eve's information about K should be arbitrarily close to zero, $I(K; ZCUG) = H(K) - H(K|ZCUG) \approx 0$, and Alice and Bob can both compute K, or $H(K|WG) = 0$. K can subsequently be used as a secret key.

We now examine four different scenarios of unconditionally secure key agreement (privacy amplification, quantum cryptography, key agreement based on noisy channels, key agreement against memory-bounded adversaries) and illustrate the concrete realizations of the three phases in these scenarios.

Although *privacy amplification* is described above as a the last phase of a protocol, it can itself be described as a key agreement protocol in which the first two phases are absent. Alice and Bob both know W (i.e. $X = Y = W$) and Eve has partial information Z about W. This corresponds to the scenario as described in Section 2.6 with Y in Theorem 2.7 replaced by K. Alice and Bob do not know P_{WV} except for a lower bound on $H_\alpha(W|V = v)$ for any $\alpha > 1$, that is, Eve's Rényi entropy of order α about W given her particular value v of V. By the Privacy Amplification Theorem and by the results of Section 4.5.1 (Theorem 4.18), Alice and Bob can generate about $H_\alpha(W|V = v)$ bits of secret key K.

In *quantum cryptography* [BC96], advantage distillation is not necessary because the raw bits transmitted over the quantum channel cannot be measured more accurately by Eve than by Bob. Such eavesdropping of Eve is guaranteed to be detected by Alice and Bob by the uncertainty relation of quantum mechanics. Therefore, quantum cryptographic protocols proceed directly with information reconciliation applied to the raw bits to correct errors caused by dark counts and by other defects of the non-ideal devices.

Unconditionally secure secret-key agreement based on *noisy channels* [Mau93, Mau94] takes place in a scenario where X, Y, and Z result from a binary random string broadcast by a satellite and received by Alice, Bob, and Eve over independent noisy channels (i.e. binary symmetric channels [CT91]) with bit error probabilities ϵ_A,ϵ_B, and ϵ_E. Secret-key agreement is possible even when Eve's channel is much more reliable than Alice's and Bob's channels ($\epsilon_A > \epsilon_E$ and $\epsilon_B > \epsilon_E$). The protocol described by Maurer and Gander [GM94] that operates on pairs of bits is used by Alice and Bob in the advantage distillation phase. It generates a much shorter bit string such that Bob's bit error rate with respect to Alice is smaller than Eve's error rate with respect to Alice. An information reconciliation protocol and a privacy amplification step follow to extract the secret key.

Unconditionally secure key agreement against *memory-bounded adversaries* (Section 5.4) is based on a random bit string of length slightly larger than the adversary's memory capacity that can be received by Alice, Bob, and Eve. The random bit string can for instance be broadcast by a satellite or over an optical network, or communicated over an insecure channel between Alice and Bob. The legitimate users Alice and Bob select and store independently a randomly selected subset of the broadcast. After some predetermined interval, they exchange the indices of their selected positions in public messages and determine the positions contained in both subsets. Unless Eve stores almost the complete broadcast string, she will have only partial information about the randomly selected subsets. Privacy amplification is applied to the part of the broadcast that Alice and Bob have in common. Because the participants can receive the random bit string without errors, no information reconciliation is necessary. However, the random selection of the subset can be interpreted as an advantage distillation phase.

5.3 Linking Information Reconciliation and Privacy Amplification

5.3.1 Introduction

As described in the previous section, information reconciliation allows two parties knowing correlated random variables, such as a noisy version of the partner's random bit string, to agree on a shared string. Privacy amplification allows two parties sharing a partially secret string about which an opponent has some partial information, to distill a shorter but almost completely secret key by communicating only over an insecure channel, as long as an upper bound on the opponent's knowledge about the string is known. The relation between these two techniques has not been developed in the literature. But it is important to understand the effect of side information, obtained by the opponent through an initial reconciliation step, on the size of the secret key that can be distilled safely by subsequent privacy amplification.

The purpose of the work presented in this section is to provide the missing link between these techniques by presenting bounds on the reduction of the Rényi entropy of a random variable induced by side information. We show that, with high probability, each bit of side information reduces the size of the key that can be safely distilled by at most one bit. Moreover, in the important special case of side information and raw key data generated by many independent repetitions of a random experiment, each bit of side information *always* reduces the size of the secret key by only about one bit.

It is known that the Rényi entropy of order 2 after reconciliation with $U = u$ (i.e. of the distribution $P_{W|V=v,U=u}$) directly provides a lower bound on the size of the secret key that can be distilled safely by privacy amplification. By applying Theorem 4.18, a similar statement can be made for Rényi entropy of order α for any $\alpha > 1$. However, the results of this section are stated in terms of Rényi entropy of order 2 because it provides the direct connection to privacy amplification.

In the following, let $V = ZC$ summarize Eve's total knowledge about W before reconciliation (adhering to the notation of the previous section). For deriving lower bounds on Eve's final information about the secret key K, one can either consider a particular value $V = v$ that Eve knows or one can average over all possible values of V. Results for a particular $V = v$, which will be considered here, are stronger than averaging results because they hold for the very instance of the protocol execution. Thus, Eve's information about W is modeled by the probability distribution $P_{W|V=v}$ about which Alice and Bob have some incomplete knowledge.

The rest of Section 5.3 is organized as follows. Section 5.3.2 presents upper bounds on the reduction of Rényi entropy due to side information for arbitrary probability distributions. Non-interactive reconciliation protocols with uniform and close-to-uniform probability distributions are investigated in Section 5.3.3. These results are applied in Section 5.3.4 to analyze the important class of scenarios in which a given random experiment is repeated many times independently.

5.3.2 The Effect of Side Information on Rényi Entropy

The reconciliation step consists of Alice and Bob exchanging suitable error-correction information U over the public channel. This information decreases Eve's Shannon entropy and usually also her Rényi entropy about W. For non-interactive reconciliation, Alice chooses an appropriate error-correction function f and sends $U = f(W)$ to Bob who can then reconstruct W with high probability from U and his prior knowledge YC.

The results of this section will be derived for an arbitrary random variable X with probability distribution P_X and a side information random variable U jointly distributed with X according to P_{XU}. However, they can just as well be applied to conditional distributions; our intended application is the key agreement scenario mentioned before, i.e. when P_X and $P_{X|U}$ are replaced by $P_{W|V=v}$ and $P_{W|V=v,U}$, respectively.

In general, giving side information implies a reduction of entropy. Our goal is to derive upper bounds on the size of this reduction. Giving as side information the fact that U takes on a particular value u, it is possible for both Shannon and Rényi entropies that the entropy increases or decreases. Moreover, the size of a reduction can be arbitrarily large.

However, the expected reduction over all values of U of the Shannon entropy of X by giving U is bounded by $H(U)$ (and is called the mutual information between X and U).

$$H(X) - H(X|U) \;=\; I(X;U) \;\leq\; H(U) \tag{5.1}$$

which follows from the symmetry of $I(X;U)$ and the fact that Shannon entropy (conditional or not) is always positive.

Example 5.1 below illustrates two facts. First, the reduction of Rényi entropy of order 2 implied by giving side information $U = u$ can exceed the reduction of Shannon entropy, i.e.

$$H_2(X) - H_2(X|U = u) \;>\; H(X) - H(X|U = u)$$

is possible. Second, it shows that the natural generalization of (5.1) to Rényi entropy of order 2, namely $H_2(X) - H_2(X|U) \leq H_2(U)$, is not true in general. However, Theorem 5.1 demonstrates that the weaker inequality $H_2(X) - H_2(X|U) \leq H(U)$ is always satisfied.

Example 5.1. Let X be a random variable with alphabet

$$\mathcal{X} \;=\; \{a_1, \ldots, a_{10}, b_1, \ldots, b_{10}\},$$

distributed according to $P_X(a_i) = 0.01$ and $P_X(b_i) = 0.09$ for $i = 1, \ldots, 10$. We have $H(X) \approx 3.79$ and $H_2(X) \approx 3.61$. Let $f : \mathcal{X} \rightarrow \{0, 1\}$ be defined as

$$f(x) \;=\; \begin{cases} 0 & \text{if } x \in \{a_1, \ldots, a_9, b_{10}\} \\ 1 & \text{if } x \in \{a_{10}, b_1, \ldots, b_9\} \end{cases}$$

and let $U = f(X)$. Then $H(X|U = 0) \approx 2.58$ and $H_2(X|U = 0) \approx 1.85$. The reduction of Rényi entropy of order 2 when given $U = 0$ exceeds the reduction of Shannon entropy, i.e. $H_2(X) - H_2(X|U = 0) \approx 1.76$ whereas $H(X) - H(X|U = 0) \approx 1.21$.

Because because f is deterministic, $H(U) = H(X) - H(X|U) \approx 0.69$. The expected entropy reductions are $H(X) - H(X|U) \approx 0.69$ and $H_2(X) - H_2(X|U) \approx 0.65$. In addition, $H_2(U) \approx 0.50$ and $H_2(X) - H_2(X|U)$ is indeed greater than $H_2(U)$ but less than $H(U)$. ○

$H(U)$ is not only the maximal expected decrease of Shannon entropy, but $H(U)$ is also an upper bound on the expected decrease of Rényi entropy of order 2, as the following theorem states.

Theorem 5.1. *Let X and U be two random variables with alphabets $\mathcal{X}$ and $\mathcal{U}$, respectively. The expected reduction of the Rényi entropy of order 2 of X, when given U, does not exceed the Shannon entropy of U, i.e.*

$$H_2(X) - H_2(X|U) \;\leq\; H(U),$$

with equality if and only if U is defined uniquely for each $x \in \mathcal{X}$ and P_U is the uniform distribution over $\mathcal{U}$ or a subset of $\mathcal{U}$.

Proof. The collision probability of X can be written as

$$\begin{aligned}
P_2(X) &= \sum_{x\in\mathcal{X}} P_X(x)^2 \\
&= \sum_{x\in\mathcal{X}} \Big(\sum_{u\in\mathcal{U}} P_{XU}(x,u) \Big)^2 \\
&\geq \sum_{x\in\mathcal{X}} \sum_{u\in\mathcal{U}} P_{XU}(x,u)^2 \\
&= \sum_{u\in\mathcal{U}} P_U(u)^2 \sum_{x\in\mathcal{X}} P_{X|U=u}(x)^2 \\
&= \sum_{u\in\mathcal{U}} P_U(u)^2 P_2(X|U=u), \qquad (5.2)
\end{aligned}$$

where the inequality follows from $(\sum_{i=1}^n p_i)^2 \geq \sum_{i=1}^n p_i^2$ for nonnegative p_i $(1 \leq i \leq n)$ and equality holds if and only if $p_i = 0$ for all but one i. Inserting (5.2) into the definition of Rényi of order 2 entropy gives

$$\begin{aligned}
H_2(X) &= -\log P_2(X) \\
&\leq -\log \sum_{u\in\mathcal{U}} P_U(u)^2 P_2(X|U=u) \\
&= -\log \sum_{u\in\mathcal{U}} P_U(u)\big[P_U(u) P_2(X|U=u)\big] \\
&\leq -\sum_{u\in\mathcal{U}} P_U(u) \log\big[P_U(u) P_2(X|U=u)\big] \\
&= -\sum_{u\in\mathcal{U}} P_U(u)\big[\log P_U(u) + \log P_2(X|U=u)\big] \\
&= -\sum_{u\in\mathcal{U}} P_U(u) \log P_U(u) - \sum_{u\in\mathcal{U}} P_U(u) \log P_2(X|U=u)
\end{aligned}$$

$$\begin{aligned} &= H(U) + \sum_{u \in \mathcal{U}} P_U(u)\ H_2(X|U=u) \\ &= H(U) + H_2(X|U), \end{aligned}$$

where the second inequality follows from the Jensen inequality (2.2), which holds with equality if and only if P_U is the uniform distribution over $\mathcal{U}$ or a subset of $\mathcal{U}$. □

In contrast to Shannon entropy, the expected conditional Rényi entropy can *increase* when side information is revealed, i.e. $H_\alpha(X) < H_\alpha(X|U)$ is possible for any $\alpha > 1$. This property is related to the spoiling knowledge proof technique described in Section 4.4 with respect to smooth entropy.

According to Theorem 5.1 and to the Markov inequality (2.5), the probability that the leaking information $U = u$ decreases Rényi entropy of order 2 by more than $kH(U)$ is at most $1/k$, i.e. the probability that U takes on a value u for which $H_2(X) - H_2(X|U=u) \geq kH(U)$ is at most $1/k$. However, such a high probability of partially exposing the string W is unacceptable in a key agreement scenario. The following theorem provides a much stronger result by showing that the Rényi entropy of order 2 decreases in fact at most by $\log|\mathcal{U}|$ except with negligible probability. It is based on Theorem 4.17 on page 81.

Theorem 5.2. *Let X and U be random variables with alphabets $\mathcal{X}$ and $\mathcal{U}$, respectively, and let $s > 0$ be an arbitrary security parameter. With probability at least $1 - 2^{-s}$, U takes on a value u for which*

$$H_2(X) - H_2(X|U=u) \ \leq\ \log|\mathcal{U}| + 2s + 2.$$

Remark. In an earlier version of this work [CM95], a weaker version of this theorem was derived showing that the Rényi entropy of order 2 decreases at most by $2\log|\mathcal{U}|$ except with negligible probability. The present version improves on this by a factor of 2 and is almost optimal, as can be seen from Lemma 5.4.

Proof. Apply Theorem 4.17 for $\alpha = 2$ and $r = t = s + 1$. □

Because of its importance here, we restate Theorem 5.2 for the key generation scenario, replacing P_X by $P_{W|V=v}$, with the side information consisting of k bits, for instance k parity checks of W when W is an n-bit string.

Corollary 5.3. *Let W be a random variable with alphabet $\mathcal{W}$, let v and u be particular values of the correlated random variables V and U with alphabets $\mathcal{V}$ and $\mathcal{U}$, respectively, with $k = \log |\mathcal{U}|$, and let $s > 0$ be a given security parameter. Then, with probability at least $1 - 2^{-s}$, U takes on a value u such that the decrease in Rényi entropy of order 2 by giving u,*

$$H_2(W|V = v) - H_2(W|V = v, U = u),$$

is at most $k + 2s + 2$.

5.3.3 Almost Uniform Distributions

As shown above, giving side information of the form $U = u$ can reduce the Rényi entropy of order 2 by an arbitrary amount, although the probability that this happens is bounded by Theorem 5.2. In this section and the next, we derive better bounds that are not probabilistic for the reduction in the case of non-interactive reconciliation and special probability distributions. For uniform distributions and deterministic side information $U = f(W)$, the reduction of Rényi entropy depends only on the size of the preimage of $u = f(x)$.

Lemma 5.4. *Let X be a random variable with alphabet $\mathcal{X}$, let $f : \mathcal{X} \to \mathcal{U}$ be an arbitrary function taking on values in a given set $\mathcal{U}$, let U be defined as $U = f(X)$, and set $\mathcal{X}_u = \{x \in \mathcal{X} : f(x) = u\}$. If X is distributed uniformly over $\mathcal{X}$, then*

$$H_2(X) - H_2(X|U = u) = \log \frac{|\mathcal{X}|}{|\mathcal{X}_u|}.$$

In particular, if f is such that $|\mathcal{X}_u|$ is the same for all $u \in \mathcal{U}$, knowledge of $U = u$ reduces the Rényi entropy of order 2 by $\log |\mathcal{U}|$.

Proof. Because Rényi entropy equals Shannon entropy for the uniform distribution, we have $H_2(X) = \log |\mathcal{X}|$ and $H_2(X|U = u) = \log |\mathcal{X}_u|$, from which the first claim immediately follows. To prove the second claim, we note that in this case $\frac{|\mathcal{X}|}{|\mathcal{X}_u|} = |\mathcal{U}|$ for all $u \in \mathcal{U}$. □

Theorems 5.5 and 5.6 state bounds on the reduction of Rényi entropy for almost uniform distributions. These results are applied in the next section to the analysis of the important class of scenarios where a given random experiment is repeated many times independently.

Theorem 5.5. *For given $\alpha > 1$ and $\beta > 1$, let X be a random variable with alphabet $\mathcal{X}$ and probability distribution P_X such that $1/(\alpha\,|\mathcal{X}|) \leq P_X(x) \leq \beta/|\mathcal{X}|$ for all $x \in \mathcal{X}$. Define f, U and $\mathcal{X}_u$ as in Lemma 5.4. Then*

$$H_2(X) - H_2(X|U = u) \;\leq\; \log \frac{|\mathcal{X}|}{|\mathcal{X}_u|} + 4\log\alpha + 2\log\beta.$$

In particular, if f is such that $|\mathcal{X}_u|$ is the same for all $u \in \mathcal{U}$, then $H_2(X) - H_2(X|U = u) \;\leq\; \log|\mathcal{U}| + 4\log\alpha + 2\log\beta$.

Proof. We can bound $P_2(X)$ as

$$P_2(X) \;=\; \sum_{x\in\mathcal{X}} P_X(x)^2 \;\geq\; |\mathcal{X}|\,\frac{1}{(\alpha\,|\mathcal{X}|)^2} \;=\; \frac{1}{\alpha^2\,|\mathcal{X}|}. \tag{5.3}$$

Using $P_U(u) \geq |\mathcal{X}_u|\frac{1}{\alpha|\mathcal{X}|}$, we get a similar upper bound for $P_2(X|U = u)$:

$$\begin{aligned} P_2(X|U = u) \;&=\; \sum_{x\in\mathcal{X}_u} P_{X|U=u}(x)^2 \\ &=\; \frac{1}{P_U(u)^2}\sum_{x\in\mathcal{X}_u} P_X(x)^2 \\ &\leq\; \left(\frac{\alpha\,|\mathcal{X}|}{|\mathcal{X}_u|}\right)^2 |\mathcal{X}_u| \left(\frac{\beta}{|\mathcal{X}|}\right)^2 \;=\; \frac{\alpha^2\beta^2}{|\mathcal{X}_u|}. \end{aligned} \tag{5.4}$$

Combining (5.3) and (5.4) gives

$$\frac{P_2(X|U = u)}{P_2(X)} \;\leq\; \frac{|\mathcal{X}|}{|\mathcal{X}_u|}\alpha^4\beta^2$$

and the theorem follows by taking logarithms on both sides. □

The following theorem provides a tighter bound for distributions that are very close to uniform. In particular, Theorem 5.6 is strictly tighter than Theorem 5.5 for $\gamma \leq 0.4563$. For $0 \leq \gamma \leq 0.3$ it is about 30% tighter.

Theorem 5.6. *For given $\gamma < \frac{1}{2}$, let X be a random variable with alphabet $\mathcal{X}$ and probability distribution P_X such that $(1-\gamma)/|\mathcal{X}| \leq P_X(x) \leq (1+\gamma)/|\mathcal{X}|$ for all $x \in \mathcal{X}$. Define f, U and $\mathcal{X}_u$ as in Lemma 5.4. Then*

$$H_2(X) - H_2(X|U = u) \;\leq\; \log \frac{|\mathcal{X}|}{|\mathcal{X}_u|} + \log \frac{(1+\gamma)^2}{1-2\gamma}.$$

Proof. For each x, define δ_x as the deviation of $P_X(x)$ from the uniform distribution: $P_X(x) = \frac{1}{|\mathcal{X}|} + \delta_x$. Hence we have $|\delta_x| \leq \gamma/|\mathcal{X}|$, and $P_2(X|U = u)$ can be expressed as

$$
\begin{aligned}
P_2(X|U=u) &= \sum_{x\in\mathcal{X}_u} P_{X|U}(x,u)^2 \\
&= \sum_{x\in\mathcal{X}_u} \left(\frac{P_{XU}(x,u)}{P_U(u)}\right)^2 \\
&= \frac{\sum_{x\in\mathcal{X}_u} P_X(x)^2}{P_U(u)^2} \\
&= \frac{\sum_{x\in\mathcal{X}_u} P_X(x)^2}{\left(\sum_{x\in\mathcal{X}_u} P_X(x)\right)^2} \\
&= \frac{\sum_{x\in\mathcal{X}_u} (\frac{1}{|\mathcal{X}|} + \delta_x)^2}{\left(\sum_{x\in\mathcal{X}_u} \frac{1}{|\mathcal{X}|} + \sum_{x\in\mathcal{X}_u} \delta_x\right)^2} \\
&= \frac{\frac{|\mathcal{X}_u|}{|\mathcal{X}|^2} + \frac{2}{|\mathcal{X}|}\sum_{x\in\mathcal{X}_u}\delta_x + \sum_{x\in\mathcal{X}_u}\delta_x^2}{\frac{|\mathcal{X}_u|^2}{|\mathcal{X}|^2} + 2\frac{|\mathcal{X}_u|}{|\mathcal{X}|}\sum_{x\in\mathcal{X}_u}\delta_x + \left(\sum_{x\in\mathcal{X}_u}\delta_x\right)^2} \\
&\leq \frac{\frac{|\mathcal{X}_u|}{|\mathcal{X}|^2} + \frac{2}{|\mathcal{X}|}|\mathcal{X}_u|\frac{\gamma}{|\mathcal{X}|} + |\mathcal{X}_u|\frac{\gamma^2}{|\mathcal{X}|^2}}{\frac{|\mathcal{X}_u|^2}{|\mathcal{X}|^2} - 2\frac{|\mathcal{X}_u|}{|\mathcal{X}|}|\mathcal{X}_u|\frac{\gamma}{|\mathcal{X}|}} \\
&= \frac{1+2\gamma+\gamma^2}{|\mathcal{X}_u| - 2\gamma|\mathcal{X}_u|}.
\end{aligned}
$$

In the third step we have made use of the fact that U is a deterministic function of X and thus $P_{XU}(x,u) = P_X(x)$. Using $P_2(X) \geq 1/|\mathcal{X}|$, we get

$$
\frac{P_2(X|U=u)}{P_2(X)} \leq \frac{|\mathcal{X}|}{|\mathcal{X}_u|} \cdot \frac{(1+\gamma)^2}{1-2\gamma},
$$

from which the theorem follows. □

5.3.4 Independent Repetition of a Random Experiment

In many practical scenarios, a certain random experiment is repeated independently a large number of times. For example, W could be the result of receiving independently generated bits over a memoryless channel, as in the satellite scenario mentioned in Section 5.2. The Asymptotic Equipartition Property (AEP), which is fundamental for information theory, states that in such a scenario all occurring sequences can be divided into a typical set and a non-typical set, where the probability that a randomly selected sequence of length n lies in the typical set approaches 1 for all sufficiently large n (see Section 2.5). Furthermore, all sequences in the typical set are almost equally probable. This allows us to bound the decrease of Rényi entropy by the results of the last section.

In the following, we will make use of strongly typical sequences as introduced in Section 2.5. Consider a random variable X with alphabet $\mathcal{X}$ and let $x^n = [x_1, \ldots, x_n]$ be a sequence of n symbols of $\mathcal{X}$. Define $N_a(x^n)$ to be the number of occurrences of the symbol $a \in \mathcal{X}$ in the sequence x^n. A sequence $x^n \in \mathcal{X}^n$ is called ϵ-strongly typical if and only if it satisfies

$$\left|\frac{1}{n}N_a(x^n) - P_X(a)\right| \;\leq\; \frac{\epsilon}{|\mathcal{X}|},$$

for all $a \in \mathcal{X}$. Let $\mathcal{S}_\epsilon^n$ be the set of all ϵ-strongly typical sequences of length n and define $X^n = X_1, \ldots, X_n$ to be a sequence of n independent and identically distributed (i.i.d.) random variables X_i with $P_{X_i} = P_X$ for $i = 1, \ldots, n$. In other words, $P_{X^n}(x^n) = \prod_{i=1}^n P_X(x_i)$. Let $o(n)$ be any function of n such that $\lim_{n\to\infty} o(n)/n = 0$. The next lemma is an asymptotic formulation of the AEP (Proposition 2.6) and asserts that for sufficiently large n, the probability that $X^n \in \mathcal{S}_\epsilon^n$ approaches 1 and that the cardinality of $\mathcal{S}_\epsilon^n$ is close to $2^{nH(X)}$.

Lemma 5.7 ([Bla87]). *Let X^n be a sequence of i.i.d. random variables distributed according to P_X. Then*

1. *For every $\delta > 0$, $P[X^n \in \mathcal{S}_\epsilon^n] \geq 1 - \delta/n$, for sufficiently large n.*
2. *For all $x^n \in \mathcal{S}_\epsilon^n$: $P_{X^n}(x^n) = 2^{-nH(X)+o(n)}$.*
3. $|\mathcal{S}_\epsilon^n| = 2^{nH(X)+o(n)}$.

Because all sequences in $\mathcal{S}_\epsilon^n$ are almost equally probable for sufficiently large n, the reduction of Rényi entropy is similar to the case of the uniform probability distribution where Rényi entropy of order 2 behaves like Shannon entropy. This observation is stated as the next theorem.

Theorem 5.8. *Let X^n be a sequence of i.i.d. random variables distributed according to P_X, let $f : \mathcal{X}^n \to \mathcal{U}$ be an arbitrary function taking on values in a given set $\mathcal{U}$, and define $\mathcal{S}_\epsilon^n(u) = \{x^n \in \mathcal{S}_\epsilon^n : f(x^n) = u\}$ and $U = f(X^n)$. For any $\delta > 0$ and sufficiently large n, the following holds with probability at least $1 - \delta/n$: X^n lies in $\mathcal{S}_\epsilon^n$ and the reduction of Rényi entropy by giving $U = u$ is upper bounded by*

$$H_2(X^n) - H_2(X^n|U = u) \ \le\ nH(X) - \log|\mathcal{S}_\epsilon^n(u)| + o(n).$$

In particular, if f is such that $|\{x \in \mathcal{X} : f(x) = u\}|$ is the same for all $u \in \mathcal{U}$ and $|\mathcal{U}| = 2^k$, then knowledge of $U = u$ reduces the Rényi entropy by at most $k + o(n)$.

Proof. By Lemma 5.7, $P_{X^n}(x^n) = 2^{-nH(X)+o(n)}$ for all $x^n \in \mathcal{S}_\epsilon^n$, and $|\mathcal{S}_\epsilon^n| = 2^{nH(X)+o(n)}$. Application of Theorem 5.5 with $\alpha = 2^{o(n)}$ and $\beta = 2^{o(n)}$ gives

$$\begin{aligned} H_2(X^n) - H_2(X^n|U = u) &\le \log\frac{|\mathcal{S}_\epsilon^n|}{|\mathcal{S}_\epsilon^n(u)|} + o(n) \\ &= nH(X) - \log|\mathcal{S}_u^n(\epsilon)| + o(n). \end{aligned}$$

□

The second part of the theorem applies in particular to all linear functions, such as parity checks from linear error-correcting codes. Due to their widespread use, linear error-correcting codes are most likely to be employed during the reconciliation phase. Theorem 5.8 can replace the spoiling knowledge argument in Maurer's proof [Mau94] that the known results on secret key rate [Mau93] hold also for a much stronger notion of secrecy.

5.3.5 Conclusions

The described link between information reconciliation and privacy amplification for unconditionally secure key agreement can be summarized as follows. Assume that Alice knows a random variable W and that Bob and Eve have partial knowledge about W, characterized by the random variables W' and V, respectively. These random variables could for instance result from the described satellite scenario with W and W' being functions of XC and YC, respectively, and with $V = ZC$. In order to state the results in the strongest possible form, we consider a particular value $V = v$ held by Eve rather than the average over all values of V.

When V gives less information than W' about W, i.e. $H(W|V) > H(W|W')$, and a lower bound $t > 0$ on the Rényi entropy of order 2 of Eve's probability distribution of W is known, i.e. $H_2(W|V = v) \geq t$, then Alice and Bob can generate a shared secret key K as follows. Alice and Bob exchange error-correcting information U consisting of $k > H(W|W')$ bits over the public channel such that Bob can reconstruct W, i.e. $H(W|W'U) \approx 0$. Eve gains additional knowledge about W by seeing $U = u$. However, Corollary 5.3 shows that with probability at least $1 - 2^{-s}$ (over all values of U) where the security parameter s can be chosen arbitrarily, her Rényi entropy of order 2 is bounded from below by $H_2(W|V = v, U = u) \geq t - k - 2s - 2$. Using privacy amplification, Alice and Bob can now generate an r-bit secret key K, where r has to be chosen smaller than $t - k - 2s - 2$ and Eve's total information about K is exponentially small in $t - k - 2s - r - 2$, namely less than $2^{r-(t-k-2s-2)}/\ln 2$ bits.

The main advantage of Theorem 5.2 is that it applies to any distribution and any reconciliation protocol whereas previously obtained results held only for particular distributions and protocols. Sharper bounds than the probabilistic statement of Theorem 5.2 that hold with probability 1 can be obtained for special distributions, as Theorems 5.5, 5.6, and 5.8 show.

5.4 Memory-Bounded Adversaries

5.4.1 Introduction

In this section, we propose a private-key cryptosystem and a protocol for key agreement by public discussion that are unconditionally secure based on the sole assumption that an adversary's memory capacity is limited. No assumption about her computing power is made. The scenario assumes that a random bit string of length slightly larger than the adversary's memory capacity can be received by all parties. The random bit string can for instance be broadcast by a satellite or over an optical network, or transmitted over an insecure channel between the communicating parties. The proposed schemes require very high bandwidth but can nevertheless be practical.

The public data is the output of a random source that is broadcast at very high rate. The legitimate users, Alice and Bob, each randomly select a small subset of the broadcast and store these values. (How this selection is performed will be described below.) Because of the random selection process, the average fraction of the information of an adversary Eve about the selected subset is the same as her fraction of information about the complete broadcast. By applying privacy amplification [BBCM95], Alice and Bob can then eliminate Eve's partial knowledge about the selected subset.

We describe how two different cryptographic tasks can be implemented using this mechanism, depending on how Alice and Bob select the random subset. If they share a secret key initially, the system realizes *private-key encryption* because the key can be used to select identical subsets. If Alice and Bob do not share secret information at the beginning, they can perform a *key agreement protocol* by public discussion: They select and store independently a random subset of the broadcast. After some predetermined interval they exchange the indices of their selected positions in public and determine the positions contained in both subsets. Privacy amplification is applied to the part of the broadcast they have in common.

Our model seems realistic because current communication and high-speed networking technologies allow broadcasting at rates of multiple gigabits per second. Storage systems that are hundreds of terabytes large, on the other hand, require a major investment by a potential adversary. Although this is within reach of government budgets, for example, the method is attractive for the following three reasons: First, the security can be based only on the assumption about the adversary's memory capacity. Second, storage costs scale linearly and can therefore be estimated accurately. Third, the system offers 'proactive' security in the sense that a future increase in storage capacity cannot break the secrecy of messages encrypted earlier.

A precursor of this system is the Rip van Winkle cipher proposed by Massey and Ingemarsson [MI85, Mas91]. This private-key system is provably computationally secure but totally impractical because a legitimate receiver must wait even longer for receiving a message than it takes an adversary to decrypt it.

Related to our work is a paper by Maurer [Mau92] that describes a system based on a large public randomizer which cannot be read entirely within feasible time. Maurer's paper contains also the idea of realizing provably secure encryption based only on assumptions about an enemy's available memory. Such a system for key agreement was described by Mitchell [Mit95], but without security proof. Our analysis provides the first proof that unconditional security can be achieved against memory-bounded adversaries. (Recently, Aumann and Rabin [Rab97] proved a conjecture of Maurer's paper [Mau92] with the same effect.)

We borrow some methods from the work of Zuckerman and others on so-called extractors of uniform randomness from weak random sources [Zuc96b] that are described in Section 4.3.6.

This section is organized as follows. In Section 5.4.2 we introduce the building blocks of our system. The main result concerning Eve's information about the randomly selected subset is given in Section 5.4.3. Based on this primitive, Sections 5.4.4 and 5.4.5 describe how to realize private-key encryption and public key agreement, respectively. The section concludes with a discussion of the underlying assumptions and future perspectives.

5.4.2 Pairwise Independence and Entropy Smoothing

In this section we present the construction of a sequence of k-wise independent random variables using universal hash functions. Universal hash functions were introduced in 1979 by Carter and Wegman [CW79, WC81] and have found many applications in cryptography and theoretical computer science [LW95] (see Definitions 2.1 and 2.2 in Section 2.6).

A strongly universal hash function can be used to generate a sequence of k-wise independent random variables in the following way: Select $G \in \mathcal{G}$ uniformly at random and apply it to any fixed sequence $x_1, \ldots, x_l$ of distinct values in $\mathcal{X}$. Let $Y_j = G(x_j)$ for $j = 1, \ldots, l$. It can easily be verified that $Y_1, \ldots, Y_l$ are k-wise independent and uniformly distributed random variables over $\mathcal{Y}$. The advantage of this technique, compared to selecting n independent samples of Y, is that it requires only $k \log |\mathcal{Y}|$ instead of $n \log |\mathcal{Y}|$ random bits.

An often-used example for a strongly k-universal hash function from $GF(2^n)$ to $GF(2^m)$ is the set

$$\mathcal{G} \;=\; \Big\{ g(x) = \mathrm{msb}_m \big(\sum_{h=0}^{k-1} a_h x^h\big) \;\Big|\; a_h \in GF(2^n),\, h = 0, \ldots, k-1 \Big\}$$

where $\mathrm{msb}_m(x)$ denotes the m most significant bits of x and the operations are in $GF(2^n)$. This construction has the nice property that when $\mathcal{G}$ is used to generate a sequence of pairwise independent random variables ($k = 2$), all values in the sequence are distinct if and only if

$a_1 \neq 0$. We will assume that $a_1 \neq 0$ whenever the pairwise independence construction is used and refer to the resulting distribution as "uniform and pairwise independent" although repeating values are excluded.

The strongly 2-universal family $\mathcal{G}$ is 2-universal even if a_0 is always set to 0. Thus, a member of the 2-universal family can be specified with only n bits.

2-universal hash functions are also the main technique to concentrate the randomness inherent in a probability distribution by a result known in different contexts as the Entropy Smoothing Theorem, the Leftover Hash Lemma [Lub96], or the Privacy Amplification Theorem [BBCM95] (Theorem 2.7 in Section 2.6).

5.4.3 Extracting a Secret Key from a Randomly Selected Subset

We are now going to show how and why Alice and Bob can exploit the fact that an adversary Eve cannot store the complete output of a public random source that is broadcast to the participants. The security proof consists of three steps. In the first step, we use the fact that Eve's storage capacity is limited to establish a lower bound on the min-entropy of Eve about the broadcast bits. The second step shows that Eve's min-entropy about a randomly selected subset of the broadcast bits is large with high probability. In the third step, we apply privacy amplification to the selected subset to obtain the secret key.

Suppose the output of a uniformly distributed binary source R is broadcast over an error-free channel and can be received by all participants. The source can be independent from the participants or it can be operated by one of the legitimate users, e.g. Alice can generate R and transmit it over an authenticated public channel to Bob. More generally, any source that is trusted to output random bits and has a sufficient capacity can be used. The channel must have high capacity, which could be realized, for example, using satellite technology for digital TV broadcasting or all-optical networks. The channel is used n times in succession and the broadcast bits are denoted by $R^n = R_1, \ldots, R_n$. We assume that Eve has a total of $m < n$ storage bits available and therefore cannot record the complete broadcast, leaving her only with partial knowledge about R^n.

During the broadcast, Eve may compute an arbitrary function of R^n with her unlimited computing power and can also use additional private random bits. We model the output of the function to be stored in her

m bits of memory by the random variable Z with alphabet $\mathcal{Z}$, subject to $\log |\mathcal{Z}| \leq m$.

Because R^n is uniformly distributed, its Rényi entropy of any order $\alpha \geq 0$ and its Shannon entropy satisfy $H_\alpha(R^n) = H(R^n) = H_\infty(R^n) = n$. The following lemma shows that the min-entropy of R^n given Z, which corresponds to Eve's knowledge about R^n, is at least $n - m$ for all but a negligible fraction of the values of Z. More precisely, the lemma implies that for any $r > 0$, the particular value z that Z takes on satisfies $H_\infty(R^n|Z = z) \geq n - m - r$, except with probability at most 2^{-r}.

Lemma 5.9. *Let X be a random variable with alphabet $\mathcal{X}$, let Z be an arbitrary random variable with alphabet $\mathcal{Z}$, and let $r > 0$. Then with probability at least $1 - 2^{-r}$, Z takes on a value z for which*

$$H_\infty(X|Z = z) \ \geq \ H_\infty(X) - \log |\mathcal{Z}| - r.$$

Proof. Let $p_0 = 2^{-r}/|\mathcal{Z}|$. Thus, $\sum_{z:P_Z(z)<p_0} P_Z(z) < 2^{-r}$. It follows for all z with $P_Z(z) \geq p_0$

$$\begin{aligned}
H_\infty(X|Z = z) &= -\log \max_{x \in \mathcal{X}} P_{X|Z=z}(x) \\
&= -\log \max_{x \in \mathcal{X}} \frac{P_X(x) P_{Z|X=x}(z)}{P_Z(z)} \\
&\geq -\log \max_{x \in \mathcal{X}} \frac{P_X(x)}{p_0} \\
&= H_\infty(X) - r - \log |\mathcal{Z}|
\end{aligned}$$

which proves the lemma. □

For the rest of this section, we denote Eve's knowledge of R^n, given the particular value $Z = z$ she observed, by the random variable $X^n = X_1, \ldots, X_n$ with alphabet $\mathcal{X}^n = \{0,1\}^n$. The distribution of X^n is arbitrary and only subject to $H_\infty(X^n) \geq n - m - r$ by Lemma 5.9.

The strategy of the legitimate users Alice and Bob is to select the values at l positions

$$\mathbf{S} \ = \ [S_1, \ldots, S_l] \quad \text{with} \quad S_1, \ldots, S_l \in \{1, \ldots, n\}$$

randomly from the broadcast symbols X^n. $\mathbf{S}$ is a vector-valued random variable taking on values $\mathbf{s} \in \{1, \ldots, n\}^l$ and the list of selected positions $X_{S_1}, \ldots, X_{S_l}$ is denoted by $X^{\mathbf{S}}$. Because this selection is performed with the uniform distribution according to the pairwise independence construction of a sequence of l values from $\{1, \ldots, n\}$ as described in Section 5.4.2, the resulting $S_1, \ldots, S_l$ are all distinct and $\mathbf{S}$ can also be viewed as a set of l values. In addition, $\mathbf{S}$ can be determined efficiently from $2 \log n$ bits.

We assume that the value of $\mathbf{S}$ is known whenever the random variable $X^{\mathbf{S}}$ is used. In the private-key system described later, Eve is assumed to obtain $\mathbf{S}$ from an oracle *after* the public random string is broadcast.

How much does Eve know about the bits selected by Alice and Bob? Intuitively, one would expect that the fraction of bits in $X^{\mathbf{S}}$ known to Eve corresponds to the fraction of bits in X^n that Eve knows (here a bit is not to be understood as a binary digit, but in the information-theoretic sense). This is indeed the case, as was observed before by Zuckerman and others in the context of weak random sources [Zuc96b, Nis96]. It is easy to show that the fraction of Eve's Shannon information corresponds to the expected value (see Theorem 5.10). However, privacy amplification can be applied only if a lower bound on the Rényi entropy of order 2 of $X^{\mathbf{S}}$ is known, which follows from the stronger bound on the min-entropy by Lemma 5.11. (Why a lower bound on Shannon entropy is not sufficient for privacy amplification was demonstrated in Example 4.4.)

The following theorem, which is outlined in [Zuc96b], shows that the Shannon entropy of $X^{\mathbf{S}}$ corresponds with high probability to the expected value $\frac{l}{n} H(X^n)$.

Theorem 5.10. *Let X^n be a random variable with alphabet $\{0,1\}^n$, let $X^{\mathbf{S}}$ correspond to l positions $\mathbf{S} = [S_1, \ldots, S_l]$ of X^n that are selected k-wise independently, and let*

$$t \;=\; \sqrt{e^{1/3}\, k\, l\, n^{-1}\, H(X^n)}.$$

Then

$$\mathrm{P}\Big[H(X^{\mathbf{S}}) \;\geq\; \frac{l}{n} H(X^n) - t\Big] \;\geq\; 1 - e^{\lfloor k/2 \rfloor},$$

where the probability is over the choice of $\mathbf{S}$.

Proof. Let $Q(i) = H(X_i | X_1 \cdots X_{i-1})$ be a function of i for $i = 1, \ldots, n$, and let I be a random variable with uniform distribution over $\{1, \ldots, n\}$. Then $\mathrm{E}[Q(I)] = \frac{1}{n} \sum_{i=1}^{n} Q(i) = \frac{1}{n} H(X^n)$ by the chain rule of entropy. We use the fact that conditioning reduces entropy to obtain a lower bound on $H(X^{\mathbf{S}})$,

$$\begin{aligned} H(X^{\mathbf{S}}) &= \sum_{j=1}^{l} H(X_{S_j}|X_{S_1}\cdots X_{S_{j-1}}) \\ &\geq \sum_{j=1}^{l} H(X_{S_j}|X_1\cdots X_{S_j-1}) \;=\; \sum_{j=1}^{l} Q(S_j). \end{aligned} \tag{5.5}$$

Because S_j for $j = 1,\ldots,l$ are chosen k-wise independently from the set $\{1,\ldots,n\}$, we can apply the Chernoff-Hoeffding bound for limited independence (2.9) to the random variable $Q(I)$ and get

$$\mathrm{P}\Bigg[\Big|\sum_{j=1}^{l} Q(S_j) - l\,\mathrm{E}[Q(I)]\Big| \geq \sqrt{e^{1/3}\,k\,l\,\mathrm{E}[Q(I)]}\Bigg] \;\leq\; e^{\lfloor k/2 \rfloor}.$$

Substituting t for $\sqrt{e^{1/3}\,k\,l\,\mathrm{E}[Q(I)]}$ implies

$$\mathrm{P}\Big[H(X^{\mathbf{S}}) \leq \frac{l}{n}H(X^n) - t\Big] \;\leq\; e^{\lfloor k/2 \rfloor}.$$

□

The proof of Theorem 5.10 works only because Shannon entropy has the property that side information can only reduce the average uncertainty. This is not the case for expected conditional Rényi entropy of order $\alpha > 1$ and is the main obstacle for extending the proof to Rényi entropy. However, the following stronger result by Zuckerman [Zuc96b] shows that also the fraction of Eve's min-entropy in the selected positions is, with high probability, close to the corresponding fraction of the total min-entropy. Because the min-entropy of a random variable is a lower bound for its Rényi entropy for any $\alpha > 0$, the lemma is sufficient for applying privacy amplification to the selected subset.

Lemma 5.11. *Let X^n be a random variable with alphabet $\{0,1\}^n$ and min-entropy $H_\infty(X^n) \geq \delta n$ (where $\frac{1}{n} \leq \delta \leq 0.9453$), let $\mathbf{S} = [S_1,\ldots,S_l]$ be chosen pairwise independently as described in Section 5.4.2, let $\rho \in [0,\frac{1}{3}]$ be such that $h(\rho) + \rho\log\frac{1}{\delta} + \frac{1}{n} = \delta$, and let $\epsilon = \sqrt{4/(\rho l) + 2^{\rho n \log\delta}}$. Then, for every value $\mathbf{s}$ of $\mathbf{S}$ there exists a random variable $A^l(\mathbf{s})$ with alphabet $\{0,1\}^l$ and min-entropy $H_\infty(A^l(\mathbf{s})) \geq \rho l/2$ such that with probability at least $1-\epsilon$ (over the choice of $\mathbf{S}$), $P_{X^{\mathbf{S}}}$ is ϵ-close to $P_{A^l(\mathbf{S})}$ in variational distance, i.e.*

$$\forall \mathbf{s} \,:\, \exists A^l(\mathbf{s}) \,:\, \mathrm{P}\big[\,\|P_{X^{\mathbf{S}}} - P_{A^l(\mathbf{S})}\|_v \leq \epsilon\,\big] \;\geq\; 1-\epsilon.$$

Remark. For fixed, large n, the value of ρ resulting from the choice in the lemma increases monotonically with δ and for δ smaller than about 0.9453 there always exists a unique $\rho \in [0, \frac{1}{3}]$ satisfying $h(\rho)+\rho \log \frac{1}{\delta} = \delta$, as can be verified easily.

Proof. Our statement of the lemma is slightly different from Zuckerman's asymptotic result [Zuc96b, Lemma 9] with respect to ρ (that we use in place of α) and ϵ, but follows also from the original proof. We describe here only the differences that lead to our formulation of the lemma.

It is straightforward to verify that $\binom{n}{i-1} = \frac{i}{n-i+1}\binom{n}{i}$ and therefore $\binom{n}{i-1} < \frac{1}{2}\binom{n}{i}$ for $i \leq n/3$. This implies $\binom{n}{i-j} < 2^{-j}\binom{n}{i}$ for $i \leq n/3$ and $0 \leq j \leq i$, from which the bound

$$\sum_{i=0}^{k} \binom{n}{i} \;<\; \binom{n}{k} \sum_{i=0}^{k} 2^{-i} \;<\; 2\binom{n}{k} \tag{5.6}$$

for any $k \leq n/3$ follows immediately. The approximation of $\binom{n}{i}$ by the binary entropy function [CT91], $\frac{1}{n+1}2^{nh(\frac{i}{n})} \leq \binom{n}{i} \leq 2^{nh(\frac{i}{n})}$, implies

$$2 \cdot 2^{nh(\rho)} \;\geq\; 2\binom{n}{\lfloor \rho n \rfloor} \;>\; \sum_{i=0}^{\lfloor \rho n \rfloor} \binom{n}{i},$$

where the second inequality follows from (5.6) for $\rho \leq \frac{1}{3}$. Thus choosing ρ as described in the statement of the lemma guarantees that

$$2^{-\delta n} \cdot \sum_{i=0}^{\lfloor \rho n \rfloor} \binom{n}{i} \;<\; 2^{-\delta n} \cdot 2^{nh(\rho)+1} \;=\; 2^{-\rho n \log \frac{1}{\delta}}$$

as required in the proof of Lemma 12 in [Zuc96b]. The choice of ϵ is the value resulting at the end of the proof of Lemma 10 in [Zuc96b]. □

We are now ready to state the main result of this section. First, we summarize the scenario and the choice of the parameters.

Let R^n be a random n-bit string with uniform distribution that is broadcast to Alice and Bob, who want to generate a secret key, and to the adversary Eve, who has a total of $m < n$ bits of memory available. Let the random variable Z denote Eve's knowledge about R^n, let $\varepsilon_1, \varepsilon_2 > 0$ be arbitrary error probabilities, and let $\Delta > 0$ be a parameter that denotes the amount of information that may leak to Eve. Let the parameters

1. $\delta = \min\{0.9453, \frac{1}{n}(n - m - \log\frac{1}{\varepsilon_1})\}$;
2. ρ such that $h(\rho) + \rho\log\frac{1}{\delta} + \frac{1}{n} = \delta$;
3. $l = \lfloor(\rho\varepsilon_2^2 - \rho 2^{-\rho n \log\frac{1}{\delta} - 2})^{-1}\rfloor$;
4. $r = \lfloor \log\Delta + \rho l/2 - 1\rfloor$.

Alice and Bob select $\mathbf{S} = [S_1, \ldots, S_l]$ randomly from $\{1, \ldots, n\}$ with the pairwise independence construction as described in Section 5.4.2 and store the bits $R^{\mathbf{S}} = R_{S_1}, \ldots, R_{S_l}$. Then they select a function $G \in \mathcal{G}$ uniformly at random from a 2-universal hash function $\mathcal{G}$ from l-bit strings to r-bit strings and compute $K = G(R^{\mathbf{S}})$ as their secret key. The random experiment consists of the choices of R^n, Z, $\mathbf{S}$, and G. As mentioned before, the theorem is proved under the (weaker) assumption that $\mathbf{S}$ is known to Eve, although this may not even be the case.

Theorem 5.12. *In the described scenario, there exists a security event $\mathcal{E}$ with probability at least $1-\varepsilon_1-\varepsilon_2$ such that Eve's information about K, given G, given her particular knowledge $Z = z$ about R^n, given $\mathbf{S} = \mathbf{s}$, and given $\mathcal{E}$, is at most Δ. Formally,*

$$\exists\mathcal{E} \;:\; \mathrm{P}[\mathcal{E}] \geq 1 - \varepsilon_1 - \varepsilon_2 \quad \textit{and} \quad I(K; G | Z = z, \mathbf{S} = \mathbf{s}, \mathcal{E}) \;\leq\; \Delta.$$

Proof. Applying Lemma 5.9 with error probability ε_1 shows that

$$H_\infty(R^n | Z = z) \;\geq\; n - m - \log\frac{1}{\varepsilon_1},$$

leading to the value of δ. Lemma 5.11 shows that $\mathbf{S}$ takes on a value $\mathbf{s}$ such that there is a distribution $P_{A^l(\mathbf{s})}$ within $\varepsilon_2/2$ of $P_{R^{\mathbf{s}}|Z=z}$ with probability $1 - \varepsilon_2/2$. Privacy amplification can be applied because

$$H_2(A^l(\mathbf{s})) \;\geq\; H_\infty(A^l(\mathbf{s})) \;\geq\; \rho l/2.$$

The choice of r guarantees $H(K|G, Z = z, \mathbf{S} = \mathbf{s}) \geq r - \Delta$ by Theorem 2.7 because $2^{r - H_2(A^l(\mathbf{s}))}/\ln 2 \leq \Delta$.

Failure of the uniformity bound, which is equivalent to the event $\overline{\mathcal{E}}$, consists of the union of the following three events. First, the bound of Lemma 5.9 can fail with probability at most ε_1. Second, $A^l(\mathbf{s})$ may deviate from the random variable with distribution $P_{R^{\mathbf{s}}|Z=z}$ with probability at most $\varepsilon_2/2$ and third, an $\mathbf{s}$ such that the distance $\|P_{X^{\mathbf{s}}} - P_{A^l(\mathbf{s})}\|_v$ is outside of the allowed range occurs with probability at most $\varepsilon_2/2$ in

Lemma 5.11. Applying the union bound, we see that $\mathrm{P}[\mathcal{E}] \geq 1 - \varepsilon_1 - \varepsilon_2$ and

$$H(K|G, Z = z, \mathbf{S} = \mathbf{s}, \mathcal{E}) \;\geq\; r - \Delta.$$

The theorem now follows from the definition of mutual information upon noting that $H(K|Z = z, \mathbf{S} = \mathbf{s}, \mathcal{E}) \leq r$. □

In a realistic cryptographic application of Theorem 5.12, the choice of the parameters is somewhat simplified because m is typically very large and because choosing a reasonable safety margin implies $n \gg m$. In this case, the parameters are $\delta = 0.9453$ and $\rho = \frac{1}{3}$, and l depends almost only on ε_2 and is close to $3/\varepsilon_2^2$. Thus, the storage required by Alice and Bob and the size of the resulting secret key are inversely proportional to the square of the desired error probability.

5.4.4 A Private-Key System

We now describe an example of a practical private-key encryption system that offers virtually the same security as the one-time pad. Assume Alice and Bob share a secret key K_0 and have both access to the broadcast public random source R^n. In addition, they are connected by an authenticated public channel on which Eve can read but not modify messages. For the pairwise independent selection of $\mathbf{S}$, the size of K_0 must be $2 \log n$ bit. However, no initial communication between the partners is needed because the interval to observe R and other parameters like l, r, ε_1, and ε_2 are fixed. The authenticated public channel is needed to exchange the description of the hash function G, which is used to extract the secret value K from $R^{\mathbf{S}}$.

In a straightforward implementation, Alice and Bob need $l(\log n + 1)$ bit of storage to hold $\mathbf{S} = [S_1, \ldots, S_l]$ and the values of $R^{\mathbf{S}}$. Because R^n is broadcast at high speed, the positions to observe must be precomputed and be recalled in increasing order. The legitimate users must only be able to synchronize on the broadcast channel and to read one bit from time to time. An adversary, however, needs equipment with high bandwidth from the channel interface through to mass storage in order to store a substantial part of R^n.

The following considerations demonstrate that this system is on the verge of being practical. The broadcast channel could be realized by a satellite. Typically, current communications satellites have a capacity of 1–10 Gbit/s [Sei96]. Commercial satellite communications services offer broadcast data rates up to 0.8 Gbit/s at consumer electronics prices. Far more capacity is offered by fiber optical networks [CHK+96]. The test bed of the All-Optical Networking Consortium, for example, has a capacity of 1 Tbit/s and has been demonstrated at 130 Gbit/s [KDD+96] (which was only limited by the number of sources available). On the other hand, tape libraries with capacities in the PByte range (1 PetaByte $= 2^{50}$ or about 10^{15} bytes) are a major investment [IEE95].

As an example for the private-key system, consider a 16 Gbit/s satellite channel that is used for one day, making $n = 1.5 \times 10^{15}$. The size of the secret key K_0 is only 102 bit. Assume the adversary can store 100 TByte ($m = 8.8 \times 10^{14}$). Using $\Delta = 10^{-20}$ and error probabilities $\varepsilon_1 = 10^{-20}$ and $\varepsilon_2 = 10^{-4}$, we see that $\delta = 0.41$, $\rho = 0.060$, $l = 1.7 \times 10^9$, and $r = 5.0 \times 10^7$, that is, about 6.0 MByte of virtually secret information K can be extracted. The legitimate users need only 10 GByte of storage each to hold the indices and the selected bits. For privacy amplification, one of them has to announce the randomly chosen universal hash function, which takes about 197 MByte. An adversary knows not more than 10^{-20} bit about K except with probability about 10^{-4}. K can be used directly for encryption with a one-time pad, for example.

The memory requirements of Alice and Bob can be reduced if fast computation enables an implicit representation of the indices $\mathbf{S}$. This seems feasible because only simple operations are needed for the pairwise independence selection method. Assuming for example that the l values can be computed in one minute, only the positions to be observed within the next minute must be stored. With the figures of the preceding example, this reduces the storage requirements to only 7 MByte for the current block of indices plus a total of 197 MByte for $R^{\mathbf{S}}$. If the computation of the indices takes longer, observation of the random broadcast could also be halted until the indices are available.

The system can be used repeatedly with only one initial key K_0, because a small part of the secret key K obtained in the first round can be used safely as the secret key for the subsequent round and so forth. In addition, some part of K can be employed to relax the authenticity requirement for the public channel using unconditionally secure message authentication techniques [WC81].

5.4.5 Key Agreement by Public Discussion

Our methods can also be used to establish a secret key between two users not sharing secret information who have access to the random broadcast and are linked by a public channel. Communication on the public channel is assumed to be authenticated, i.e. the adversary can read but not modify messages. This system offers public key agreement with virtually the same security as the one-time pad under the sole assumption that the adversary's memory capacity is limited. (The public communication channel is different from the public broadcast channel whose only purpose is to distribute a large number of random bits.)

To agree on a secret key, Alice and Bob independently select and store a subset of the broadcast random bits R^n. After a predetermined amount of time, they announce the chosen set of positions on the public channel. The secret key can then be extracted from the values of R^n at the common positions using privacy amplification. To keep the communication and storage requirements for Alice and Bob at a reasonable level, it is crucial that they use a memory-efficient description of the index set. Fortunately, the pairwise independent selection method achieves this.

Both Alice and Bob select a sequence of q uniform and pairwise independent indices $T_1, \ldots, T_q$ and $U_1, \ldots, U_q$, respectively, from $\{1, \ldots, n\}$ as described in Section 5.4.2. (The values of q and the other parameters $n, l, r, \varepsilon_1, \varepsilon_2$ are fixed and also known to Eve.) Alice stores the values of R^n at the indices in $\mathbf{T} = [T_1, \ldots, T_q]$, denoted by $R^{\mathbf{T}}$, and Bob stores $R^{\mathbf{U}}$ for his indices $\mathbf{U} = [U_1, \ldots, U_q]$. We assume that they use a memory-efficient, implicit representation of the index set as described earlier for the private-key system, with on-line recomputation of the indices when necessary. In this way, Alice and Bob need approximately $\log q$ bits of memory each.

Because of the pairwise independent selection, both index sets can be determined from $2 \log n$ bits each. The descriptions of $\mathbf{T}$ and $\mathbf{U}$ exchanged on the public channel are therefore short. In order to apply Theorem 5.12 to the set $\{S_1, \ldots, S_l\} = \{T_1, \ldots, T_q\} \cap \{U_1, \ldots, U_q\}$ of common positions, we need the following lemma to make sure that also $S_1, \ldots, S_l$ have a uniform and pairwise independent distribution. It is easy to see that the expected number of common indices is $l = q^2/n$.

Lemma 5.13. *Let $T_1, \ldots, T_q$ and $U_1, \ldots, U_q$ be independent sequences of uniform and pairwise independent random variables, respectively, with alphabet $\{1, \ldots, n\}$ and distribution as described in Section 5.4.2, and let $S_1, \ldots, S_q$ be the sequence $T_1, \ldots, T_q$ restricted to those values occurring in $U_1, \ldots, U_q$, i.e. $S_j = T_j$ if there is an index h such that $U_h = T_j$ and $S_j = \omega$ otherwise. Then, the sequence $S_1, \ldots, S_q$ restricted to those* *positions different from ω is pairwise independent.*

Proof. Because the pairwise independence construction of Section 5.4.2 is used, no values in the **U** and **T** sequences are repeated. This implies

$$\mathrm{P}[T_i = x_1 \wedge T_j = x_2] \;=\; \frac{1}{n(n-1)}$$

for all $i, j \in \{1, \ldots, q\}$ and all $x_1, x_2 \in \{1, \ldots, n\}$ with $x_1 \neq x_2$. The sequence $S_1, \ldots, S_l$ satisfies

$$\mathrm{P}[S_i = x_1 \wedge S_j = x_2 | S_i \neq \omega \wedge S_j \neq \omega] \;=\; \frac{\mathrm{P}[S_i = x_1 \wedge S_j = x_2]}{\mathrm{P}[S_i \neq \omega \wedge S_j \neq \omega]} \tag{5.7}$$

for any $i, j \in \{1, \ldots, q\}$ and $x_1, x_2 \in \{1, \ldots, n\}$. Considering only those positions of the sequence $S_1, \ldots, S_q$ with values different from ω, we see that for any $i, j \in \{1, \ldots, q\}$ and all $x_1, x_2 \in \{1, \ldots, n\}$ such that $x_1 \neq x_2$ and $S_i \neq \omega$ and $S_j \neq \omega$,

$$\begin{aligned}
&\mathrm{P}[S_i = x_1 \wedge S_j = x_2] \\
&\quad = \;\mathrm{P}[T_i = x_1 \wedge \exists h_1 : U_{h_1} = x_1 \wedge T_j = x_2 \wedge \exists h_2 : U_{h_2} = x_2] \\
&\quad = \;\mathrm{P}[T_i = x_1 \wedge T_j = x_2] \cdot \mathrm{P}[\exists h_1, h_2 : U_{h_1} = x_1 \wedge U_{h_2} = x_2] \\
&\quad = \;\frac{1}{n(n-1)} \cdot \frac{q(q-1)}{n(n-1)}.
\end{aligned}$$

Furthermore, for all $i, j \in \{1, \ldots, q\}$, we have

$$\mathrm{P}[S_i \neq \omega \wedge S_j \neq \omega] \;=\; \mathrm{P}[\exists h_1, h_2 : U_{h_1} = T_i \wedge U_{h_2} = T_j] \;=\; \frac{q(q-1)}{n(n-1)}$$

because every pair of distinct x_1, x_2 occurs with the same probability in the sequence $U_1, \ldots, U_q$. Thus, the probability in (5.7) is equal to $\frac{1}{n(n-1)}$ for any $i, j \in \{1, \ldots, q\}$ and all $x_1 \neq x_2$, and the lemma follows. □

To illustrate a concrete example of the system, assume Alice and Bob both have access to a 40 Gbit/s broadcast channel. We need more network capacity for public key agreement than for private-key encryption to achieve a similar error probability. The channel is used for 2×10^5 seconds (about two days), thus $n = 8.6 \times 10^{15}$. Eve is allowed to store 1/2 PByte or $m = 4.5 \times 10^{15}$ bit. With $\Delta = 10^{-20}$ and error probabilities $\varepsilon_1 = 10^{-20}$ and $\varepsilon_2 = 10^{-3}$, the parameters are $\delta = 0.476$, $\rho = 0.077$, $l = 1.3 \times 10^7$, and $r = 5.0 \times 10^5$. In order to have l common indices on the average, Alice and Bob must store $q = \sqrt{ln} = 3.3 \times 10^{11}$ bit or about 39 GByte each (assuming the index sequences $\mathbf{T}$ and $\mathbf{U}$ are represented implicitly). The public communication between Alice and Bob consists of $2 \log n = 106$ bit in each direction for the selected indices plus 1.5 MByte in one direction for privacy amplification. Except with probability about 10^{-3}, Eve knows less than 10^{-20} bit about the 61 KByte secret key that Alice and Bob obtain.

Because l is on the order of the inverse squared error probability ε_2, the probabilities in the example are relatively large to keep the storage requirements of Alice and Bob at a reasonable level. Generating a shorter key does not help to reduce the storage space, which depends primarily on ε_2. It is an interesting open question whether Lemma 5.11 can be improved in order to reduce the influence on the error probability.

The large size of the hash function that has to be communicated for privacy amplification can be reduced by using "almost universal" hash functions based on almost k-wise independent random variables that can be constructed efficiently [AGHP92]. Such functions $g : \mathcal{X} \to \mathcal{Y}$ can be described with about $5 \log |\mathcal{Y}|$ instead of $\log |\mathcal{X}|$ bits and can replace universal hash functions in privacy amplification [GW96, SZ94].

5.4.6 Discussion

Our results show that unconditional security can be based on assumptions about the adversary's available memory. In essence, such a system exploits the capacity gap between fast communication and mass storage technology. We discuss a few implications of this fact.

First of all, generating random bits at a sufficiently high rate may be more expensive than merely transmitting them. However, a large investment in a random source can be amortized by the potentially high number of participants that can use the source simultaneously.

A drawback of our system is that the security margin is linear in the sense that memory costs are directly proportional to the offered storage capacity, at least up to technological advances. In most computationally secure encryption systems, the complexity of a brute-force attack grows exponentially in the length of the keys.

Our system is provably secure taking into account the current storage capacity of an adversary because the only possible attack is to store the broadcast data when it is sent. In contrast, most computationally secure systems can be broken retroactively, once better algorithms are discovered or faster processing becomes possible.

We have used the broadcast channel as an error-free black-box communication primitive in our system, although the legitimate users do not need its full functionality: They need not receive the complete broadcast, but only a small part of it. It is conceivable that a receiving device could be much simpler and less expensive if it can only synchronize and read a small, but arbitrary part of the traffic. Such receivers could also allow for a greater capacity of the channel.

The described protocols offer no resilience to errors on the broadcast channel. To take into account such errors, Alice and Bob can perform information reconciliation [BS94] on the selected subset. Methods for bounding the effect of this additional information provided to Eve are known [CM97a].

The system rests on the gap between two technologies—fast communication and mass storage. Impressive future developments can be expected in both fields. We mention only the enormous potential of all-optical networks on one hand and the recent developments in holographic and molecular storage on the other hand.

Chapter 6

Concluding Remarks

The challenge in the development of cryptographic algorithms is to find practical systems with strong and guaranteed security properties. Today, the unconditional security model offers an attractive way for realizing practical provably secure cryptosystems because proofs for the computational security of a cipher still seem to be far away.

One goal that has been formulated at the outset of the work presented in this thesis could not be reached. It is the formalization of a general theory of unconditionally secure key agreement from correlated information.

A step in this direction is the formalization of smooth entropy presented in Chapter 4. This measure allows a comprehensive view of an important part of the general key agreement scenario and shows that entropy smoothing is an information-theoretic concept with applications in many areas of theoretical computer science. The formalization paved also the way for closing the gap in the connection between smooth entropy and Rényi entropy of order α for $1 < \alpha < 2$. But a general theory and a corresponding information measure to characterize the number of secret key bits that can generated by key agreement from common information remain a distant goal.

With respect to smooth entropy, it is an open question whether there are more efficient smoothing algorithms than universal hash functions for extracting uniform random bits. It is also unclear whether Rényi entropy of order $\alpha > 1$ is relevant for extractors used in complexity theory. The current literature on extractors is formulated only in terms of the min-entropy of a weak random source. It seems natural to generalize these results to Rényi entropy of any order $\alpha > 1$. This would considerably relax the assumptions on the weak random sources needed for polynomial-time probabilistic computation.

The main part of such an extension would be a generalization of Lemma 5.11 from min-entropy to Rényi entropy of order 2 or any order $\alpha > 1$. This would show that not only the min-entropy of a randomly selected subset but also the Rényi entropy is with high probability close to the expected fraction of the entropy of the longer string.

Chapter 5 shows how to make unconditionally secure key agreement possible based only on an assumption about the adversary's memory capacity. Our methods provide a third mechanism, in addition to quantum channels and to noisy channels, upon which proofs of unconditional security can be built. It is conceivable that this mechanism can be used for realizing other cryptographic primitives than key agreement, such as oblivious transfer or bit commitment. An efficiency improvement for the key agreement protocol would also result from the described generalization of Lemma 5.11 to Rényi entropy.

Bibliography

[AD75] J. Aczél and Z. Daróczy, *On measures of information and theiry characterizations*, Mathematics in science and engineering, vol. 115, Academic Press, New York, 1975.

[AGHP92] Noga Alon, Oded Goldreich, Johan Håstad, and René Peralta, *Simple constructions of almost k-wise independent random variables*, Random Structures and Algorithms **3** (1992), no. 3, 289–304, Preliminary version presented at 31st FOCS (1990).

[Ari77] S. Arimoto, *Information measures and capacity of order α for discrete memoryless channels*, Topics in Information Theory (I. Csiszar and P. Elias, eds.), North-Holland, 1977, pp. 41–52.

[Ari96] Erdal Arikan, *An inequality on guessing and its application to sequential decoding*, IEEE Transactions on Information Theory **42** (1996), no. 1, 99–105.

[BBB+92] Charles H. Bennett, François Bessette, Gilles Brassard, Louis Salvail, and John Smolin, *Experimental quantum cryptography*, Journal of Cryptology **5** (1992), no. 1, 3–28.

[BBCM95] Charles H. Bennett, Gilles Brassard, Claude Crépeau, and Ueli M. Maurer, *Generalized privacy amplification*, IEEE Transactions on Information Theory **41** (1995), no. 6, 1915–1923.

[BBR86] Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert, *How to reduce your enemy's information*, Advances in Cryptology — CRYPTO '85 (Hugh C. Williams, ed.), Lecture

Notes in Computer Science, vol. 218, Springer-Verlag, 1986, pp. 468–476.

[BBR88] Charles H. Bennett, Gilles Brassard, and Jean-Marc Robert, *Privacy amplification by public discussion*, SIAM Journal on Computing **17** (1988), no. 2, 210–229.

[BC94] Amos Beimel and Benny Chor, *Interaction in key distribution schemes*, Advances in Cryptology — CRYPTO '93 (Douglas R. Stinson, ed.), Lecture Notes in Computer Science, vol. 773, Springer-Verlag, 1994.

[BC96] Gilles Brassard and Claude Crépeau, *25 years of quantum cryptography*, SIGACT News **27** (1996), no. 3, 13–24.

[BC97] Gilles Brassard and Claude Crépeau, *Oblivious transfers and privacy amplification*, Advances in Cryptology — EUROCRYPT '97 (Walter Fumy, ed.), Lecture Notes in Computer Science, vol. 1233, Springer-Verlag, 1997, pp. 334–347.

[BCJL93] Gilles Brassard, Claude Crépeau, Richard Jozsa, and Denis Langlois, *A quantum bit commitment scheme provably unbreakable by both parties*, Proc. 34th IEEE Symposium on Foundations of Computer Science (FOCS), 1993.

[Bil95] Patrick Billingsley, *Probability and measure*, third ed., Wiley, New York, 1995.

[BL90] Josh Benaloh and Jerry Leichter, *Generalized secret sharing and monotone functions*, Advances in Cryptology — CRYPTO '88 (Shafi Goldwasser, ed.), Lecture Notes in Computer Science, vol. 403, Springer-Verlag, 1990, pp. 27–35.

[Bla83] Richard E. Blahut, *Theory and practice of error control codes*, Addison-Wesley, Reading, 1983.

[Bla87] Richard E. Blahut, *Principles and practice of information theory*, Addison-Wesley, Reading, 1987.

[Blo85] Rolf Blom, *An optimal class of symmetric key generation systems*, Advances in Cryptology: Proceedings of EUROCRYPT 84 (Thomas Beth, Norbert Cot, and Ingemar Ingemarsson, eds.), Lecture Notes in Computer Science, vol. 209, Springer-Verlag, 1985, pp. 335–338.

[BS94] Gilles Brassard and Louis Salvail, *Secret-key reconciliation by public discussion*, Advances in Cryptology — EUROCRYPT '93 (Tor Helleseth, ed.), Lecture Notes in Computer Science, vol. 765, Springer-Verlag, 1994, pp. 410–423.

[BSH+93] Carlo Blundo, Alfredo De Santis, Amir Herzberg, Shay Kutten, Uga Vaccaro, and Moti Yung, *Perfectly-secure key distribution for dynamic conferences*, Advances in Cryptology — CRYPTO '92 (Ernest F. Brickell, ed.), Lecture Notes in Computer Science, vol. 740, Springer-Verlag, 1993, pp. 471–486.

[Cac97] Christian Cachin, *Smooth entropy and Rényi entropy*, Advances in Cryptology — EUROCRYPT '97 (Walter Fumy, ed.), Lecture Notes in Computer Science, vol. 1233, Springer-Verlag, 1997, pp. 193–208.

[CG85] Benny Chor and Oded Goldreich, *Unbiased bits from sources of weak randomness and probabilistic communication complexity*, Proc. 26th IEEE Symposium on Foundations of Computer Science (FOCS), 1985, pp. 429–442.

[CHK+96] R. Cruz, G. Hill, A. Kellner, R. Ramaswami, G. Sasaki, and Y. Yamabashi, Eds., *Special issue on optical networks*, IEEE Journal on Selected Areas in Communications **14** (1996), no. 5, 761–1052.

[CK81] Imre Csiszár and János Körner, *Information theory: Coding theorems for discrete memoryless systems*, Academic Press, New York, 1981.

[CK89] Claude Crépeau and Joe Kilian, *Achieving oblivious transfer using weakened security assumptions*, Proc. 29th IEEE Symposium on Foundations of Computer Science (FOCS), 1989.

[CM95] Christian Cachin and Ueli Maurer, *Linking information reconciliation and privacy amplification*, Advances in Cryptology — EUROCRYPT '94 (Alfredo De Santis, ed.), Lecture Notes in Computer Science, vol. 950, Springer, 1995, pp. 266–274.

[CM97a] Christian Cachin and Ueli Maurer, *Linking information reconciliation and privacy amplification*, Journal of Cryptology **10** (1997), no. 2, 97–110.

[CM97b] Christian Cachin and Ueli Maurer, *Smoothing probability distributions and smooth entropy*, Preprint (Abstract in Proc. 1997 IEEE International Symposium on Information Theory, Ulm), 1997.

[CM97c] Christian Cachin and Ueli Maurer, *Unconditional security against memory-bounded adversaries*, Advances in Cryptology — CRYPTO '97 (Burt Kaliski, ed.), Lecture Notes in Computer Science, Springer-Verlag, 1997.

[Cré96] Claude Crépeau, *What is going on with quantum bit commitment?*, Proceedings of Pragocrypt '96, Czech Technical University Publishing House, 1996.

[Cré97] Claude Crépeau, *Efficient cryptographic protocols based on noisy channels*, Advances in Cryptology — EUROCRYPT '97 (Walter Fumy, ed.), Lecture Notes in Computer Science, vol. 1233, Springer-Verlag, 1997, pp. 306–317.

[CSGV92] R. M. Capocelli, A. De Santis, L. Gargano, and U. Vaccaro, *On the size of shares for secret sharing schemes*, Advances in Cryptology — CRYPTO '91 (Joan Feigenbaum, ed.), Lecture Notes in Computer Science, vol. 576, Springer-Verlag, 1992, pp. 101–113.

[Csi95a] László Csirmaz, *The size of a share must be large*, Advances in Cryptology — EUROCRYPT '94 (Alfredo De Santis, ed.), Lecture Notes in Computer Science, vol. 950, Springer-Verlag, 1995, pp. 13–22.

[Csi95b] Imre Csiszár, *Generalized cutoff rates and Rényi's information measures*, IEEE Transactions on Information Theory **41** (1995), no. 1, 26–34.

[CT91] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley, 1991.

[CW79] J. Lawrence Carter and Mark N. Wegman, *Universal classes of hash functions*, Journal of Computer and System Sciences **18** (1979), 143–154.

[DH76] Whitfield Diffie and Martin E. Hellman, *New directions in cryptography*, IEEE Transactions on Information Theory **22** (1976), no. 6, 644–654.

[Fel68] William Feller, *An introduction to probability theory*, 3rd ed., Wiley, 1968.

[GM84] Shafi Goldwasser and Silvio Micali, *Probabilistic encryption*, Journal of Computer and System Sciences **28** (1984), 270–299.

[GM94] Martin J. Gander and Ueli M. Maurer, *On the secret-key rate of binary random variables*, Proc. 1994 IEEE International Symposium on Information Theory, 1994, p. 351.

[GNS75] Robert M. Gray, David L. Neuhoff, and Paul C. Shields, *A generalization of Ornstein's $\bar{d}$ distance with applications to information theory*, The Annals of Probability **3** (1975), no. 2, 315–328.

[Gol95] Oded Goldreich, *Foundations of cryptography (fragments of a book)*, Electronic Colloquium on Computational Complexity (ECCC), `http://www.eccc.uni-trier.de/eccc/`, 1995.

[GW96] Oded Goldreich and Avi Wigderson, *Tiny families of functions with random properties: A quality-size trade-off for hashing*, Preprint available from the authors, preliminary version presented at 26th STOC (1994), January 1996.

[HILL91] Johan Håstad, Russell Impagliazzo, Leonid A. Levin, and Michael Luby, *Construction of a pseudo-random generator from any one-way function*, Tech. Report 91-068, International Computer Science Institute (ICSI), Berkeley, 1991.

[HV93] Te Sun Han and Sergio Verdú, *Approximation theory of output statistics*, IEEE Transactions on Information Theory **39** (1993), no. 3, 752–772.

[IEE95] *Proc. 14th IEEE symposium on mass storage systems*, IEEE Computer Society Press, 1995.

[ILL89] Russell Impagliazzo, Leonid A. Levin, and Michael Luby, *Pseudo-random generation from one-way functions*, Proc.

21st Annual ACM Symposium on Theory of Computing (STOC), 1989, pp. 12–24.

[ISN93] Mitsuru Ito, Akira Saito, and Takao Nishizeki, *Multiple assignment scheme for secret sharing*, Journal of Cryptology **6** (1993), 115–20.

[Kah67] David Kahn, *The codebreakers: The story of secret writing*, Macmillan, New York, 1967.

[KDD+96] I. P. Kaminow, C. R. Doerr, C. Dragone, T. Koch, U. Koren, A. A. M. Saleh, et al., *A wideband all-optical WDM network*, IEEE Journal on Selected Areas in Communications **14** (1996), no. 5, 780–799.

[Kha93] M. Kharitonov, *Cryptographic hardness of distribution-specific learning*, Proc. 25th Annual ACM Symposium on Theory of Computing (STOC), 1993, pp. 372–381.

[KV94] Michael J. Kearns and Umesh V. Vazirani, *An introduction to computational learning theory*, MIT Press, 1994.

[Lub96] Michael Luby, *Pseudorandomness and cryptographic applications*, Princeton University Press, 1996.

[LW95] Michael Luby and Avi Wigderson, *Pairwise independence and derandomization*, Tech. Report 95-035, International Computer Science Institute (ICSI), Berkeley, 1995.

[Mas91] James L. Massey, *Contemporary cryptography: An introduction*, Contemporary Cryptology: The Science of Information Integrity (Gustavus J. Simmons, ed.), IEEE Press, 1991, pp. 1–39.

[Mas93] James L. Massey, *Lecture notes for "Applied Digital Information Theory I"*, Abteilung für Elektrotechnik, ETH Zürich, 1993.

[Mas94] James L. Massey, *Guessing and entropy*, Proc. 1994 IEEE International Symposium on Information Theory, 1994, p. 204.

[Mau92] Ueli M. Maurer, *Conditionally-perfect secrecy and a provably-secure randomized cipher*, Journal of Cryptology **5** (1992), 53–66.

[Mau93] Ueli M. Maurer, *Secret key agreement by public discussion from common information*, IEEE Transactions on Information Theory **39** (1993), no. 3, 733–742.

[Mau94] Ueli M. Maurer, *The strong secret key rate of discrete random triples*, Communications and Cryptography: Two Sides of One Tapestry (Richard E. Blahut et al., eds.), Kluwer, 1994.

[Mau95] Ueli Maurer, *Lecture notes for "Theoretische Informatik II"*, Abteilung für Informatik, ETH Zürich, 1995.

[Mau96] Ueli M. Maurer, *A unified and generalized treatment of authentication theory*, Proc. 13th Annual Symposium on Theoretical Aspects of Computer Science (STACS) (Ruediger Reischuk Claude Puech, ed.), Lecture Notes in Computer Science, vol. 1046, Springer-Verlag, 1996, pp. 190–198.

[Mau97] Ueli Maurer, *Information-theoretically secure secret-key agreement by NOT authenticated public discussion*, Advances in Cryptology — EUROCRYPT '97 (Walter Fumy, ed.), Lecture Notes in Computer Science, Springer-Verlag, 1997.

[May96] Dominic Mayers, *The trouble with quantum bit commitment*, Manuscript, available from Los Alamos reprint archive quant-ph, March 1996.

[MI85] James L. Massey and Ingemar Ingemarsson, *The Rip van Winkle cipher: A simple and provably computationally secure cipher with a finite key*, Proc. 1985 IEEE International Symposium on Information Theory, 1985, p. 146.

[Mit95] C. J. Mitchell, *A storage complexity based analogue of Maurer key establishment using public channels*, Cryptography and Coding: 5th IMA Conference, Cirencester, UK (Colin Boyd, ed.), Lecture Notes in Computer Science, vol. 1025, Springer, 1995, pp. 84–93.

[MR95] Rajeev Motwani and Prabhakar Raghavan, *Randomized algorithms*, Cambridge University Press, 1995.

[MvOV97] Alfred J. Menezes, Paul C. van Oorschot, and Scott A. Vanstone, *Handbook of applied cryptography*, CRC Press, Boca Raton, FL, 1997.

[MW97] Ueli Maurer and Stefan Wolf, *The intrinsic conditional mutual information and perfect secrecy*, Tech. Report 268, Department of Computer Science, ETH Zürich, 1997.

[MY95] Robert J. McEliece and Zhong Yu, *An inequality on entropy*, Proc. 1995 IEEE International Symposium on Information Theory, 1995, p. 329.

[MZG95] A. Muller, H. Zbinden, and N. Gisin, *Underwater quantum coding*, Nature **378** (1995), 449.

[Nis96] Noam Nisan, *Extracting randomness: How and why — a survey*, Proc. 11th Annual IEEE Conference on Computational Complexity, 1996.

[Nun92] Thomas S. Nunnikhoven, *A birthday solution for nonuniform birth frequencies*, American Statistician **46** (1992), no. 4, 270–274.

[NZ95] Noam Nisan and David Zuckerman, *Randomness is linear in space*, Preprint available from the authors, preliminary version presented at 25th STOC (1993), 1995.

[Pap94] Christos H. Papadimitriou, *Computational complexity*, Addison-Wesley, Reading, 1994.

[Pom90] Carl Pomerance, *Factoring*, Cryptology and Computational Number Theory (Carl Pomerance, ed.), Proceedings of Symposia in applied Mathematics, vol. 42, American Mathematical Society, 1990, pp. 27–47.

[Rab97] Michael Rabin, Personal Communication, 1997.

[Rén61] Alfréd Rényi, *On measures of entropy and information*, Proc. 4th Berkeley Symposium on Mathematical Statistics and Probability (Berkeley), vol. 1, Univ. of Calif. Press, 1961, pp. 547–561.

[Rén65] Alfréd Rényi, *On the foundations of information theory*, Rev. Inst. Internat. Stat. **33** (1965), 1–14, reprinted in [Tur76].

[Rén70] Alfred Rényi, *Probability theory*, North-Holland, Amsterdam, 1970.

[Riv90] Ronald L. Rivest, *Cryptography*, Handbook of Theoretical Computer Science (J. van Leeuwen, ed.), Elsevier, 1990, pp. 717–755.

[RSA78] Ronald L. Rivest, Adi Shamir, and Leonard Adleman, *A method for obtaining digital signatures and public-key cryptosystems*, Communications of the ACM **21** (1978), no. 2, 120–126.

[Sar80] Dilip V. Sarwate, *A note on universal hash functions*, Information Processing Letters **10** (1980), 41–45.

[Sei96] Lawrence P. Seidman, *Satellites for wideband access*, IEEE Communications Magazine (1996), 108–111.

[Sha48] Claude E. Shannon, *A mathematical theory of communication*, Bell System Technical Journal **27** (1948), 379–423, 623–656.

[Sha49] Claude E. Shannon, *Communication theory of secrecy systems*, Bell System Technical Journal **28** (1949), 656–715.

[Sha79] Adi Shamir, *How to share a secret*, Communications of the ACM **22** (1979), no. 11, 612–613.

[Sho94] Peter W. Shor, *Algorithms for quantum computation: Discrete log and factoring*, Proc. 35th IEEE Symposium on Foundations of Computer Science (FOCS), 1994, pp. 124–134.

[Sim91] Gustavus J. Simmons (ed.), *Contemporary cryptology: The science of information integrity*, IEEE Press, 1991.

[Spi96] Timothy P. Spiller, *Quantum information processing: Cryptography, computation, and teleportation*, Proceedings of the IEEE **84** (1996), no. 12, 1719–1746.

[SSS95] Jeanette P. Schmidt, Alan Siegel, and Aravind Srinivasan, *Chernoff-Hoeffding bounds for applications with limited independence*, SIAM Journal on Discrete Mathematics **8** (1995), no. 2, 223–250.

[Sti92] D. R. Stinson, *An explication of secret sharing schemes*, Designs, Codes and Cryptography **2** (1992), 357–390.

[Sti95] Douglas R. Stinson, *Cryptography: Theory and practice*, CRC Press, 1995.

[Sti96] Douglas R. Stinson, *On some methods for unconditionally secure key distribution and broadcast encryption*, Preprint, 1996.

[SZ94] Aravind Srinivasan and David Zuckerman, *Computing with very weak random sources*, Preprint available from the authors, preliminary version presented at 35th FOCS (1994), 1994.

[TS96] Amnon Ta-Shma, *On extracting randomness from weak random sources*, Proc. 28th Annual ACM Symposium on Theory of Computing (STOC), 1996, pp. 276–285.

[Tur76] Paul Turan (ed.), *Alfréd Rényi: Selected papers*, Akademiai Kiado, Budapest, 1976, 3 volumes.

[Val84] L. G. Valiant, *A theory of the learnable*, Communications of the ACM **27** (1984), no. 11, 1134–1142.

[VV95] Sridhar Vembu and Sergio Verdú, *Generating random bits from an arbitrary source: Fundamental limits*, IEEE Transactions on Information Theory **41** (1995), no. 5, 1322–1332.

[WC81] Mark N. Wegman and J. Lawrence Carter, *New hash functions and their use in authentication and set equality*, Journal of Computer and System Sciences **22** (1981), 265–279.

[Wyn75] A. D. Wyner, *The wire-tap channel*, Bell System Technical Journal **54** (1975), no. 8, 1355–1387.

[WZ95] Avi Wigderson and David Zuckerman, *Expanders that beat the eigenvalue bound: Explicit construction and applications*, Preprint available from the authors, preliminary version presented at 25th STOC (1993), 1995.

[Zuc91] David Zuckerman, *Simulating BPP using a general weak random source*, Proc. 32nd IEEE Symposium on Foundations of Computer Science (FOCS), 1991, pp. 79–89.

[Zuc96a] David Zuckerman, *Randomness-optimal sampling, extractors, and constructive leader election*, Proc. 28th Annual ACM Symposium on Theory of Computing (STOC), 1996, pp. 286–295.

[Zuc96b] David Zuckerman, *Simulating BPP using a general weak random source*, Algorithmica **16** (1996), 367–391, Preliminary version presented at 32nd FOCS (1991).

Index