Theoretical Foundations of Adversarial Binary Detection

Mauro Barni[1] and Benedetta Tondi[2]

[1]*Department of Information Engineering and Mathematics, University of Siena, Italy; barni@dii.unisi.it*
[2]*Department of Information Engineering and Mathematics, University of Siena, Italy; benedettatondi@gmail.com*

ABSTRACT

The present monograph focuses on the detection problem in adversarial setting. When framed in an adversarial setting, classical detection theory can not be applied any more, since, in order to make a correct decision, the presence of an adversary must be taken into account when designing the detector. In particular, the interplay between the Defender ($\mathscr{D}$), wishing to carry out the detection task, and the Attacker ($\mathscr{A}$), aiming at impeding it, must be investigated. The purpose of this monograph is to lay out the foundations of a general theory of adversarial detection, taking into account the impact that the presence of the adversary has on the design of the optimal detector. We do so by casting the adversarial detection problem into a game theoretical framework, which is then studied by relying on typical methods of information theory. As a final result, the theory allows to state the conditions under which both the false positive and false negative error probabilities tend to zero exponentially fast, and to relate the error exponents of the two kinds of errors to the distortion the attacker can introduce into the test sequence.

Mauro Barni and Benedetta Tondi (2021), "Theoretical Foundations of Adversarial Binary Detection", Foundations and Trends® in Communications and Information Theory: Vol. 18, No. 1, pp 1–172. DOI: 10.1561/0100000102.

1

Introduction

Security-oriented applications of signal processing have received increasing attention in the last decades; digital watermarking, steganography and steganalysis, multimedia forensics, biometrics, network intrusion detection, spam filtering, traffic monitoring, video surveillance are just some examples of such an interest. All these fields are characterized by a unifying feature: the presence of one or more adversaries aiming at making the system fail.

Although each adversarial scenario has its own peculiarities, there are some fundamental questions whose solution under a unified framework would ease the understanding of the underlying security problems and the development of effective and general solutions. Such an observation has prompted the birth of a new discipline, namely *adversarial signal processing* [1], whose final aim is to design signal processing tools which retain their effectiveness even in the presence of an adversary. Within such a framework, classical methods can no longer be applied, since the presence of two contenders with opposite goals and their mutual interaction must be properly taken into account. The goal of this monograph is to present a coherent theory of the most recent

findings regarding the single most common problem in adversarial signal processing, namely binary detection in adversarial setting.

The monograph originates from the research activity carried out by the authors over the last six years, with particular reference to the results proven in [2]–[5]. Other related papers have been published by the same authors and by other researchers, however they are not discussed in this monograph to let the reader focus on the core theory. A brief overview of related works is given in Section 1.3 to introduce the reader to the most interesting extensions of the results presented here.

1.1 Application Areas

Binary detection, sometimes referred to as binary decision or a particular kind of binary hypothesis testing, is a ubiquitous problem in virtually all branches of science and technology. In many cases, binary detection must be carried out in a setting wherein the presence of an adversary aiming at inducing a wrong decision can not be ruled out. Upon restricting the attention to signal processing and data science applications, examples of binary detection problems that, by their nature, are required to work in an adversarial setting include: network monitoring, intrusion detection, spoofing detection in biometric recognition systems, watermarking, steganography and steganalysis, multimedia forensics, spam filtering, video surveillance, anomaly detection, malware detection and many others.

In network monitoring applications, for instance, a common binary detection problem consists in detecting if there is an on-going Denial of Service (DoS) attack. In the simplest case, the presence of the attack can be detected by relying on a few traffic characteristics like the traffic rate, the provenance of data packets and the frequency of traffic bursts [6]. In the likely case that the hacker responsible for the DoS attack is aware of the presence of a network monitoring service, he will try to shape the traffic resulting from the attack in such a way that its characteristics are as close as possible to those of the benign traffic loading the network in the absence of attacks (while of course retaining the effectiveness of the attack). In this way, the hacker is going to alter the statistics of the observed traffic in the presence of the attack, thus impacting heavily

the performance of the monitoring service in case the service had been designed without taking into account the presence of the attacker. Of course, the designer of the monitoring service does not know exactly how the hacker will shape the traffic. In turn, the hacker may not know the exact features the traffic monitoring service is going to rely on to make his decision. This uncertainty, or lack of knowledge, characterizing both the network analyst and the hacker, must be properly taken into account by both parties to optimise the actions they are going to take. It is the goal of adversarial detection theory to model the interplay between the analyst and the hacker to suggest the *best* way for them to reach their (opposite) goals, and derive the performance the monitoring service can achieve despite the presence of the adversary.

A similar situation occurs in spam filtering applications [7], [8]. Even in this case, the spammer and the filter designer engage in a struggle wherein the designer of the spam filtering service looks for a reliable way to distinguish normal e-mails from spam, while the spammer does its best to convey the intended malevolent payload letting spam messages resemble normal e-mails, or, in a similar but not equivalent way, by avoiding that they are recognized as spam. Once again, designing the filter without taking into account the possible efforts made by the spammer to evade detection would result in poor filtering performance. In the same way, creating spam e-mails neglecting the presence of the anti-spam filter would result in most of the spam being filtered out.

Another relevant scenario, even closer to the theory presented in this monograph, is Multimedia Forensics (MF) [9]. Most problems in MF can be formulated as a binary detection or hypothesis testing problem. For instance, the MF analyst may be asked to distinguish between synthetic and natural images, or to decide if a given image has been captured by a specific device or not. In other cases, the analysis aims at deciding if an image or a video has been compressed once or multiple times, since the compression history of the image/video may reveal important aspects of the processing chain the image/video has been subject to. In yet other cases, binary detection requires understanding if a certain media has been manipulated since it has been captured or not. Since the very first days of MF research, it has been recognised that forensic analysis had to cope with the opposite effort, usually

referred to as counter-forensics, made by a counterfeiter [10]. From this perspective, counter-forensics can then be defined as a way to degrade the performance of the hypothesis test envisaged by the analyst. In an attempt to avoid a never-ending loop wherein new defenses and attacks are developed iteratively, and to an extent anticipating the theory developed here, the authors of [10] argued that the Kullback–Leibler distance between the probability density functions of the observed signals after the application of the counter-forensic attack is a proper way to measure the effectiveness of the attack itself. Noticeably, such measure does not depend on the particular technique adopted by the analyst. Even though the formulation in [10] does not explicitly use the game-theoretic approach, this can be seen as the first step towards the definition of the equilibrium point of a general multimedia forensics game.

Prior to multimedia forensics, the arguments used in [10] had already been adopted to model the interplay between steganography and steganalysis. In steganography, the steganographer modifies a cover media, usually an image, to hide within it a hidden message. The resulting image, referred to as a stego image, is sent to the intended receiver of the hidden message in such a way that an external observer does not notice the presence of the hidden message, thus creating a cover channel between the steganographer and the receiver [11]. The goal of the steganalyzer is to observe the communication between the sender and the receiver, trying to distinguish between the cover and stego images. As in the previous examples, the task of the steganalyzer corresponds to a binary detection problem (detecting stego images), taking into account the opposite effort of the steganalyzer who aims at making the cover and stego images indistinguishable. Interestingly, the mathematical model used to describe the interplay between the steganographer and the steganalyzer is very similar to that used in [10], with the steganographer playing the role of the counterfeiter and the steganalyzer the role of the forensic analyst [12].

Biometric authentication is yet another discipline which is often faced with binary decision in settings wherein the presence of an adversary cannot be ignored. In biometric-based user verification, for instance, the authenticating system must decide whether a biometric template (a face

image, a fingerprint, an iris image or any other biometric trait) belongs to a certain individual, despite the opposite efforts of an attacker aiming at building a fake template that passes the authentication test. In other cases, the owner of the biometric template modifies the template to avoid being recognized [13]. In both cases, the distortion introduced within the template as a consequence of the attack should be minimal impede the detection of the attack. Another problem pertaining to biometric security that is naturally modelled as an adversarial binary detection problem, is anti-spoofing. A spoofing attack refers to a situation wherein the attacker attempts to impersonate the target by presenting to the authentication system a synthetic copy of the biometric signal used for authentication. In the case of face-based authentication, for instance, a spoofing attack is easily implemented by showing to the authentication system the face of the victim displayed on the screen of a mobile phone (rebroadcast attack). In this framework, the goal of the anti-spoofing system is to distinguish between natural and rebroadcast images. In his turn, the attacker will try to generate the image or video to be rebroadcast in such a way that it is judged as a natural one by the spoofing detection system. In doing so, the attacker must preserve the quality of the displayed image/video since otherwise it would fail to be recognized as the victim of the attack [14].

In all the examples described so far, the attack is carried out at test time. The situation is rather different in applications entailing the use of machine learning tools, since in such cases the attacker may already act during the training phase [15]. With such detectors, the different distributions of samples observed under the two hypotheses being tested is not known through statistical models, rather, they are learnt during the training phase in which examples of data produced under the two hypotheses are shown to the system. If the attacker can interfere with the training phase, he can try to modify the training data to facilitate a subsequent attack carried out at test time. Many examples of the effectiveness of this kind of attacks have been published recently, due to the ever-increasing popularity of machine learning techniques [16]. In Chapter 6, while addressing the problem of binary detection with corrupted training data, we touch upon attacks carried out at training time.

1.2 Scope of the Theory

The main idea behind adversarial detection theory (and adversarial signal processing in general) consists in casting the detection problem into a *game-theoretic* framework, which permits to rigorously define the goals and the actions available to the two contenders, namely, the designer of the detector, hereafter referred to as Defender ($\mathscr{D}$), and the adversary, referred to as the Attacker ($\mathscr{A}$).

In the following, we introduce the general adversarial binary detection problem addressed in this monograph, which is a binary hypothesis testing problem.[1]

Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete sources belonging to the class of the discrete memoryless sources (DMS) $\mathcal{C}$, with alphabet $\mathcal{X}$. The goal of the Defender, $\mathscr{D}$, is to decide whether a test sequence $z^n \in \mathcal{X}^n$ has been generated by X (hypothesis H_0) or Y (hypothesis H_1). As a result of the test, $\mathcal{X}^n$ is partitioned into two complementary regions Λ^n and $\bar{\Lambda}^n$, such that for $z^n \in \Lambda^n$, $\mathscr{D}$ decides in favor of H_0, while for $z^n \in \bar{\Lambda}^n$, H_1 is preferred. We have a Type-I, or false positive, error when $\mathscr{D}$ decides for H_1 and H_0 is true, and a Type-II, or false negative, error when the decision is in favor for H_0 while H_1 occurs. We indicate the probability of a Type-I, or false positive error as P_{FP} and the probability of a Type-II or false negative error as P_{FN}. Our goal is to design a hypothesis test that encompasses the presence of an attacker aiming at impeding a correct decision. A Neyman–Pearson (NP) setup [17, Chapter 3, p. 63] is considered for the decision test. Accordingly, $\mathscr{D}$ must choose the decision regions Λ^n and $\bar{\Lambda}^n$ in such a way as to ensure that the Type-I error probability is lower than a certain prescribed value. The Attacker, $\mathscr{A}$, takes a sequence y^n generated by Y and transforms it into a sequence z^n so that when presented with the modified sequence, $\mathscr{D}$ still accepts H_0. In doing so, $\mathscr{A}$ must respect a distortion constraint, limiting the amount of modifications that can be introduced into the sequence. In such a scenario, the goal of the Attacker is to cause a false negative decision error. Therefore, $\mathscr{A}$ aims at maximizing the Type-II error probability, while $\mathscr{D}$'s goal is to minimize it by taking into account

[1] For an introduction on the statistical method of hypothesis testing, the reader is referred to [17].

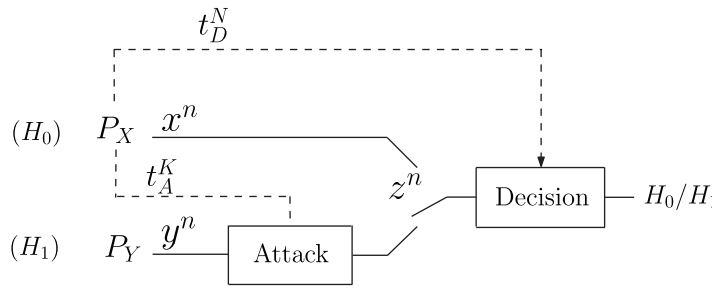


Figure 1.1: General setup of the adversarial binary decision test. P_X and P_Y govern the generation of the test sequence under H_0 and H_1 respectively. P_X also underlies the generation of the training sequences t_D^N and t_A^K for the case of binary decision based on training data.

the presence of $\mathscr{A}$. The above scenario provides a suitable model for the detection problems found in many practical applications, where the rejection of H_0 corresponds to raising some kind of alarms and $\mathscr{A}$ aims at preventing it (e.g., to avoid that an anomalous situation is detected, or to allow the access to a system or service to a unauthorized user).

A schematic representation of the adversarial binary detection test in its general form is depicted in Figure 1.1. The continuous line drawing refers to the most basic scenario. Let $x^n \in \mathcal{X}^n$, resp. $y^n \in \mathcal{X}^n$, be a sequence drawn from X, resp. Y, and let $z^n \in \mathcal{X}^n$ denote the sequence observed by the $\mathscr{D}$. We then have $z^n = x^n$ under H_0, whereas, under H_1, z^n is a modified version of y^n produced by $\mathscr{A}$ in the attempt to deceive $\mathscr{D}$. In the rest of this monograph, we assume that X and Y are discrete memoryless sources (DMS).

In this monograph, we address several variants of the above problem, depending on the knowledge available to the Defender and the Attacker about the statistical characterization of the system under the two hypotheses, which can be full or based on training data, and on the capability of the adversary, who may attack the system at test time only or both during the training and testing phases.

Below, we summarize the setups of the adversarial binary decision test considered in this monograph.

1.2.1 Adversarial Binary Detection Setups

In the simplest setup, referred to as *binary detection with known sources*, the Defender and the Attacker have full knowledge of the statistics characterizing the system, i.e., they know the probability mass function ruling the emission of the test sequence under H_0. The scheme illustrating this setup is the one corresponding to the continuous-line drawing in Figure 1.1. Binary detection with known sources is studied in Chapters 3 and 4. The second setup studied in this monograph considers the more realistic case in which the sources are not fully known to the Defender and the Attacker. In this case, $\mathscr{D}$ and $\mathscr{A}$ obtain their knowledge about X through the observation of a training sequence. This setup is schematized in Figure 1.1 (solid and dashed line drawing). In the most general case, the training sequences observed by $\mathscr{D}$ and $\mathscr{A}$, namely t_D^N and t_A^K, are different and have different length ($N \neq K$). Such a setup is referred to as *binary detection with training data*, and is studied in Chapter 5. We also consider a setup that accounts for the possibility that the Attacker corrupts part of the training data available to the Defender. This corresponds to a more complicated situation, since the action of the Attacker also affects the decision under H_0, thus impacting on both Type-I and Type-II error probabilities (while in the previous cases, the action of the attack had an impact on H_1 only). This setup, referred to as *binary detection with corrupted training*, is studied in Chapter 6. More specifically, two different scenarios are considered in Chapter 6, one corresponding to the case where the attacker can only add some samples to the training sequence, and the other to the case where he replaces a percentage of samples of the training sequence. A schematic representation of the adversarial detection test in the corrupted training setup is reported in Figure 1.2. With reference to the notation in the figure, the original training sequence t_A^K is corrupted by $\mathscr{A}$ producing t_A^m. The corrupted training sequence t_A^m is the one observed by $\mathscr{D}$, upon which he bases the decision. Such a sequence has length $m > K$ in the case of sample addition, while in the scenario of sample replacement, $m = K$. The scheme presented in Figure 1.2 is a very general one. A more detailed representation for each of the two scenarios with corrupted training is provided in Chapter 6. The two

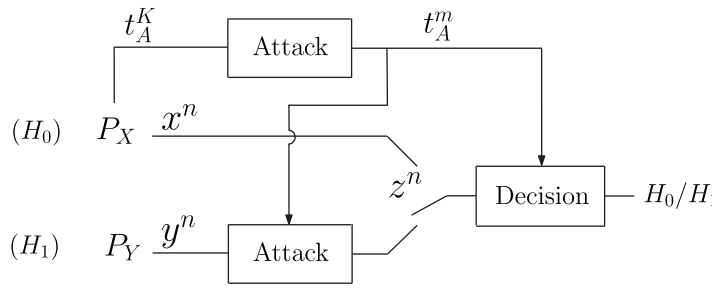


Figure 1.2: Setup of the adversarial binary decision test with corruption of the training set. P_X and P_Y rule the generation of the test sequence under H_0 and H_1 respectively. P_X also rules the generation of the training sequence t_A^K.

Table 1.1: Summary of the adversarial binary detection tests addressed in this monograph

	Defender		Adversary		
Setup	**Source Knowledge**	**Goal**	**Source Knowledge**	**Goal**	**Capability**
Known sources	P_X		P_X		Modify y^n
Detection with training data	t_D^N	min P_{FN}[2]	t_A^K	max P_{FN}	Modify y^n
Corrupted training	t_A^m		t_A^K		Modify y^n and t_A^K (sample addition or replacement)

variants of the game corresponding to sample addition and replacement are discussed in Sections 6.3 and 6.6, respectively.

Table 1.1 summarizes the three adversarial detection setups considered in this monograph.

In all the setups, the game between the Defender and the Attacker is solved by relying on information-theoretic methods, notably on the *method of types*, under some limiting, yet reasonable, assumptions on the statistics used by the Defender to make a decision. The analysis starts with a formal definition of the game, and proceeds by looking for the equilibrium point and with the evaluation of the payoff at the equilibrium. The analysis of the payoff permits one to draw some conclusions about the outcome of the games. From the analysis of

[2]The minimization of P_{FN} is subject to a constraint on P_{FP}.

the achievable performance of the various games, and by drawing a parallelism with *optimal transport theory*, we are also able to define a measure of statistical distinguishability of information sources under adversarial conditions.

In fact, it turns out that as long as the distortion the adversary is allowed to introduce is smaller than a certain quantity, called Security Margin ($\mathcal{SM}$), at the equilibrium both the false positive and false negative error probabilities tend to zero exponentially fast (hence ensuring strictly positive error exponents). On the other hand, if the allowed distortion is larger than $\mathcal{SM}$, the error probabilities can not tend to zero simultaneously. The exact value of $\mathcal{SM}$ depends on the probability density functions governing the emission of the test sequence under H_0 and H_1 and the particular version of the game played by $\mathscr{A}$ and $\mathscr{D}$. Comparing the Security Margin to the distortion introduced by the attacker permits one to anticipate the results of the race of arms between $\mathscr{D}$ and $\mathscr{A}$ for a given strength of the attack when the length of the observed sequence tends to infinity.

1.3 Related Work

In this monograph, we focus on the core of adversarial binary detection theory, paying particular attention to the game-theoretic framework wherein such a theory is cast, and prove theorems stating the most important results of the theory. We do so by analyzing first the basic binary detection game under the assumption that the sources underlying the two hypotheses being tested are known, then we extend the analysis to the more complicate case of sources known through the observation of (possibly corrupted) training data. The theory presented in this monograph, however, does not exhaust the problems addressed and the results proven in the last years pertaining to the general field of adversarial detection. Several extensions of the basic theory have been published both by the authors of this monograph and by other researchers, and several related problems have been addressed as described in the following.

One recent extension of the theory concerns the case of a *fully active* attacker, that is an attacker that acts also when the null hypothesis holds. In many cases, it is reasonable to assume that the attacker

is active under both hypotheses with the goal of causing both false positive and false negative detection errors. As an example, we may consider the case of a radar target detection system, where the defender wishes to distinguish between the presence and the absence of a target, by considering the presence of a hostile jammer. To maximize the damage caused by his actions, the jammer may decide to act under both hypotheses: when H_1 holds, to avoid that the defender detects the presence of the target, and in the H_0 case, to increase the number of false alarms inducing a waste of resources deriving from the adoption of possibly expensive countermeasures even when they are not needed. In a completely different scenario, we may consider an image forensic system aiming at deciding whether a certain image has been shot by a given camera, for instance because the image is involved in a legal procedure. Even in this case, the attacker may be interested in causing a missed detection event, or induce a false alarm to accuse an innocent party. The binary detection game with a fully active adversary is studied extensively in [18], where various versions of the game are considered according to whether the attacker is aware of the real status of the observed system.

A different adversarial hypothesis testing game is introduced in [19]. In this work, the price the attacker has to pay to modify the distribution of samples emitted under H_1 is expressed as a cost added to the payoff of the game, rather than as a hard constraint on the admissible attacking strategies. This results in a non-zero sum game admitting a Nash equilibrium point, for which the authors derive exponential rates of convergence of classification errors.

Another extension of the theory presented in this monograph concerns the case of binary detection based on multiple observations. This scenario is relevant in several applications, including multimedia forensics, data fusion, distributed hypothesis testing and detection, sensor networks, and cognitive radio networks. In all these cases, a fusion center has to take a binary decision about the status of a system by relying on a number of observations made available by different sensors or a number of traces detected by different investigation tools. In many situations, it is possible that an attacker corrupts the observations or deliberately provides misleading data to induce a decision error at the

fusion center. The binary detection game with multiple observations studied in [20] models several situations that can be traced back to the above general formulation, accounting for attackers altering a different number of observations and with different attacking capabilities.

Data fusion with corrupted observations is itself a widely studied topic. Such a problem, often referred to as distributed binary detection in the presence of Byzantines [21], deals with a situation wherein a fusion center must make a decision about the status of a system based on the reports submitted by local agents observing the system at different locations or under different conditions. In particular, binary detection must be carried out despite the possible presence of corrupted agents (referred to as Byzantines) submitting possibly corrupted reports with the goal of inducing a decision error. The Byzantines must satisfy two opposite requirements: (i) maximize the error probability at the fusion center and (ii) avoid being identified. To accomplish this, they can choose among many corruption strategies, however they must do so without knowing the precise detection strategy adopted by the fusion center. In its turn the fusion center must select its detection strategy without knowing the exact attack strategy implemented by the Byzantines. This is a typical dilemma encountered in adversarial binary detection games, thus opening the way to the study of the data fusion problem with corrupted reports via the game-theoretic methods discussed in this monograph (see [22], [23, Chapter 5] for specific examples). Other approaches to distributed binary detection with Byzantines are discussed in [24]–[26]. An example of distributed estimation in the presence of tampered sensors can be found in [27]. For a thorough review of distributed inference in the presence of Byzantines readers are referred to [28].

As a last remark, we mention interesting relationships – deserving further investigation – between adversarial binary detection with training data and the vast body of research devoted to studying the security of Machine Learning (ML) [29], [30]. Despite the difficulty of applying the theory described in this monograph to practical applications, due to the difficulty of building precise statistical models to describe the kind of data ML systems usually involve, such a theory can be conveniently used to get useful insights about the security level that can be reached

by binary detectors in practice. An example of such an analysis applied to image forensics is described in [31]. The theoretical framework behind Generative Adversarial Networks (GANs) also presents interesting connections to adversarial detection with training data. As explained in the seminal work by Goodfellow *et al.* [32], GANs are based on a game played by a generator and a discriminator, the former aiming at generating samples that mimic those of a certain class (e.g., natural images), in such a way that the discriminator can not distinguish between natural samples and samples produced by the generator. The generator, in turn, iteratively updates its decision strategy by learning the characteristics of the samples output by the generator. Interestingly, [32] shows that the equilibrium point of the game is reached when the data produced by the generator minimizes the Jensen–Shannon divergence between the distributions of natural and synthetic samples, which is by any means equivalent to the generalized log-likelihood ratio function appearing in Theorem 5.3 defining the equilibrium point of the binary detection game with training data.

1.4 Outline of the Monograph

This monograph is organized as follows: in Chapter 2 we review the basic tools required to derive and understand the results of our analysis. In Chapter 3, we define and study the simple case of binary detection when the statistical characterization of the observed system is known to both the Defender and the Attacker. The achievable performance of this game are studied in Chapter 4 where we also introduce the source distinguishability concept. The analysis of Chapters 3 and 4 is extended in Chapter 5 to the case in which the statistics of the observed system are known through training data. Then, in Chapter 6, we generalize the adversarial setup studied in Chapter 5 by considering a version of the game in which the adversary can corrupt part of the training data available to the Defender. A summary of the main contributions of the theory and a discussion of its possible extensions are given in Chapter 7.

For a good comprehension of the theory treated in the monograph, the reader is assumed to have a solid background in information theory. Some basic knowledge of classical detection theory is also required.

2

Background Notions and Tools

This chapter provides the reader with the background and the tools necessary for understanding the rest of the monograph. Such tools are of a heterogeneous nature and belong to a number of diverse disciplines including information theory, game theory, optimal transport, and large deviation theory.

2.1 Notation and Definitions

We start by introducing the notation and definitions used throughout the monograph.

Capital letters will be used to indicate scalar discrete random variables (RVs), whose specific realizations will be represented by the corresponding lower case letters. Random sequences, whose length will be denoted by n, are indicated by $X^n = (X_1, X_2, \ldots, X_n)$, where X_i denotes the i-th element of the sequence, $i = 1, \ldots, n$. Instantiations of random sequences are denoted by the corresponding lowercase letters, so x^n indicates a specific realization of the random sequence X^n, and x_i the i-th element of x^n. Information sources will also be defined by capital letters. Throughout the monograph we will focus exclusively on discrete sources. The alphabet of a source will be denoted by the corresponding

calligraphic capital letter (e.g., $\mathcal{X}$). Calligraphic letters will also be used to indicate classes of information sources ($\mathcal{C}$) and classes of probability density functions ($\mathcal{P}$). The *probability mass function* (pmf) of a random variable X will be denoted by P_X. The same notation will be used to indicate the probability measure ruling the emission of sequences from a source X, so we will use the expressions $P_X(a)$ and $P_X(x^n)$ to denote, respectively, the probability of symbol $a \in \mathcal{X}$ and the probability that the source X emits the sequence x^n, the exact meaning of P_X being always clearly recoverable from the context wherein it is used. Similarly, P_{XY} denotes the joint pmf of a pair of random variables (X, Y). The notation $X \sim P_X$ indicates that the source X emits symbols according to P_X. Generic sets will also be denoted with capital letters. Given an event A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$), we will use the notation $P_X(A)$ to denote the probability of the event A under the probability measure P_X. Notation $\bar{A}$ will be used to denote the complementary set of A.

Let x^n be a sequence with elements belonging to a finite alphabet $\mathcal{X}$. The *type* P_{x^n} of the sequence x^n is defined as the empirical probability distribution induced by the sequence x^n, that is, the vector of the relative frequencies of the various alphabet symbols in x^n. In the following, we denote by $\mathcal{P}_n$ the set of types with denominator n, i.e., the set of types induced by sequences of length n. Given $P \in \mathcal{P}_n$, we denote by $\mathcal{T}(P)$ the *type class* of P, i.e., the set of all the sequences in $\mathcal{X}^n$ having type P. Similarly, given a sequence x^n we denote by $\mathcal{T}(P_{x^n})$, or simply $\mathcal{T}(x^n)$, the set of the sequences having the same type as x^n. Given a pair of sequences (x^n, y^n), $P_{y^n \,|\, x^n}$ denotes the empirical conditional probability distribution, i.e., the conditional type. The conditional type class $\mathcal{T}(P_{y^n \,|\, x^n})$, or $\mathcal{T}(y^n \,|\, x^n)$, is the set of sequences y^n having empirical conditional probability distribution (i.e., conditional type) $P_{y^n \,|\, x^n}$. Some basic results concerning types are provided in Section 2.4.1.

Regarding information theoretic measures, the empirical entropy of the sequence x^n, that is the entropy associated with P_{x^n}, is defined as

$$H(P_{x^n}) = -\sum_{a \in \mathcal{X}} P_{x^n}(a) \log P_{x^n}(a), \tag{2.1}$$

sometimes simply referred to as H_{x^n}. Similar definitions hold for other information theoretic quantities (e.g., joint and conditional entropy)

governed by empirical distributions. The Kullback–Leibler (KL) divergence between two distributions P and Q defined on the same finite alphabet $\mathcal{X}$ is:

$$\mathcal{D}(P\|Q) = \sum_{a\in\mathcal{X}} P(a)\log\frac{P(a)}{Q(a)}, \tag{2.2}$$

where, according to usual conventions, $0\log 0 = 0$ and $p\log p/0 = \infty$ if $p > 0$. If P_{x^n} and P_{y^n} are empirical distributions induced, respectively, by x^n and y^n, $\mathcal{D}(P_{x^n}\|P_{y^n})$ is the empirical KL-divergence.

With regard to binary hypothesis testing, relying on the notation already introduced in Section 1.2, the two decision error probabilities P_{FP} and P_{FN} are given by

$$P_{\mathrm{FP}} = P(H_1\,|\,H_0) = P(z^n \in \bar{\Lambda}^n\,|\,H_0) = P_X(z^n \in \bar{\Lambda}^n), \tag{2.3}$$

$$P_{\mathrm{FN}} = P(H_0\,|\,H_1) = P(z^n \in \Lambda^n\,|\,H_1) = P_Y(z^n \in \Lambda^n), \tag{2.4}$$

where X and Y are the two sources emitting symbols under H_0 and H_1 respectively, Λ^n is the acceptance region of hypothesis H_0, and z^n is the test sequence.

Our main focus is on the asymptotic behavior of P_{FP} and P_{FN} as n tends to infinity. We define the false positive (η) and false negative (ε) error exponents as follows:

$$\eta = -\limsup_{n\to\infty}\frac{\log P_{\mathrm{FP}}}{n}; \quad \varepsilon = -\limsup_{n\to\infty}\frac{\log P_{\mathrm{FN}}}{n}, \tag{2.5}$$

where the log's are taken in base 2. Note that when the limit exists the above definitions can be simplified by avoiding the use of lim sup: whenever this is the case, we use lim instead of lim sup.

Given two sequences x^n and y^n, the distance induced by the L_p-norm, $p \geq 1$, is referred to as the L_p distance and is denoted by $d_{L_p}(x^n, y^n)$. Then,

$$d_{L_p}(x^n, y^n) = \|x^n - y^n\|_{L_p} = \left(\sum_{i=1}^{n}|x_i - y_i|^p\right)^{1/p}. \tag{2.6}$$

Throughout the monograph, we will use the same notation d_{L_p} also to denote the L_p distortion between two pmf's in $\mathcal{P}$,[1] the exact meaning being always clear from the context.

[1] Given two pmf's P and Q, $d_{L_p}(P,Q) = (\sum_{a\in\mathcal{X}}|P(a) - P(a)|^p)^{1/p}$.

The *Hamming distance* between x^n and y^n is defined as the number of locations for which $x_i \neq y_i$, i.e., $d_H(x^n, y^n) = n - \sum_{i=1}^{n} \delta(x_i, y_i)$, with $\delta(x_i, y_i) = 1$ if $x_i = y_i$ and 0 otherwise (Kronecker delta).

Definition 2.1. A distance function d: $\mathcal{X}^n \times \mathcal{X}^n \to \mathbb{R}^+$ is said to be *permutation-invariant* if for every two sequences x^n and y^n, $d(y^n, z^n) = d(\sigma(y^n), \sigma(z^n))$ for all permutations σ of the elements of the sequences.

Below we introduce the concept of distances between subsets and the definition of the *Hausdorff distance* as a way to measure the distance between subsets of a metric space [33, Chapter 2]. We denote by (S, d) a metric space with the distance functions d. For any point $x \in S$ and any non-empty subset $A \subseteq S$, the distance of x from A is defined as:

$$d(x, A) = \inf_{a \in A} d(x, a). \tag{2.7}$$

Definition 2.2. For any given pair (A, B) of subsets of S let $\delta_A(B) = \sup_{b \in B} d(b, A)$. Let δ_H be a function that associates to the pair of subsets (A, B) the quantity

$$\delta_H(A, B) = \max\{\delta_A(B), \delta_B(A)\}. \tag{2.8}$$

$\delta_H(A, B)$ is called the Hausdorff distance between A and B.

According to the above definition, the Hausdorff distance is not a true metric, but only a pseudometric, since $\delta(A, B) = 0$ implies only that the closures of the sets coincide, namely $cl(A) = cl(B)$, but not necessarily that $A = B$. In order for δ_H to be a metric, the definition must be restricted to closed subsets. Let $\mathcal{L}(S)$ denote the space of non-empty closed and bounded subsets of S and let δ_H: $\mathcal{L}(S) \times \mathcal{L}(S) \to [0, \infty)$. The assumption of boundedness of the sets[2] guarantees that the Hausdorff distance takes a finite value. Then, the space $\mathcal{L}(S)$ endowed with the Hausdorff metric δ_H is a metric space [33, Chapter 2].

Definition 2.3. Let $\{K_n\}$ be a sequence of closed and bounded subsets of (S, d), i.e., $K_n \in \mathcal{L}(S)$ $\forall n$. We use the notation $K_n \xrightarrow{H} K$ to indicate that the sequence has a limit in $(\mathcal{L}(S), \delta_H)$ and the limiting set is K.

[2] Recall that boundedness of the sets depends on the distance measure d defined in the metric space.

2.2 Game Theory in a Nutshell

Game Theory is a branch of mathematics devoted to the study of the interplay, of conflict and/or cooperation, between *decision makers* or *players*. Game-theoretic concepts apply whenever the actions of several decision makers are interdependent, that is their choices potentially affect, and are affected by, the choices of the other players. Game Theory is also referred to as *interactive decision theory*, as opposed to *classical decision theory*. While classical decision theory has been used to study signal processing problems, Game Theory can be naturally advocated for the study of adversarial signal processing, where the simple adoption of a worst-case analysis for the design of the system (carried out under the assumption that the attacker always acts in such a way as to cause the greatest damage to the system) leads to suboptimum solutions.

The birth of modern Game Theory as a unique field traces back to 1944, with the book "Theory of Games and Economic Behavior" by von Neumann and Morgenstern [34]. Game Theory provides tools to formulate, model, and study strategic scenarios in a wide variety of application fields, from economics and political science to computer science. A central assumption in most variants of Game Theory is that each decision maker is *rational* and *intelligent*. A rational player is one who has a relation of preferences regarding the outcomes of the game.[3] An intelligent player is able to act in a rational way and then always chooses the action that gives the most preferable outcome to him, i.e., the action that maximizes his gain or payoff, given his expectation on the other players. The goal of game theory analysis is to predict how the game will be played, or, relatedly, to give advice on how to play the game against rational opponents.

Game Theory models are highly abstract representations of classes of real-life situations for which equilibrium solutions are suggested, having some desirable properties. Game Theory encompasses a large variety of situations depending on the number of players, the way the players interact, the knowledge that a player has of the strategies adopted by the others, the deterministic or probabilistic nature of

[3] Axioms of rationality, Von Neumann-Morgenstern utility theorem [34, Chapter 3].

the game, etc. In all the models, the basic entity is the *player*, which should be interpreted as an individual or as a group of individuals making a decision. A distinction can be made between situations in which the players have common goals, and are allowed to form binding agreements (*cooperative* games) and situations in which the players have different and possibly conflicting goals (*non-cooperative* games). Another common distinction is made between *simultaneous* and *sequential* games. Simultaneous games are games where both players move simultaneously, or if they do not move simultaneously, they are unaware of the earlier players' actions (making their action effectively simultaneous). On the contrary, sequential games (or dynamic games) are games where players have some knowledge about earlier actions. The difference between simultaneous and sequential games is captured in the different ways of representing the game. With reference to non-cooperative games, the *strategic* form is used when the players choose their action or plan of actions once and for all at the beginning, that is, when all the players' decisions are made simultaneously (strategic form games are discussed below). By contrast, the so-called *extensive* form is used for sequential games, when each player needs to reconsider his plan of action whenever it is his turn to move [35, Chapters 5 and 10].

The extensive form of a game is an explicit, highly descriptive, representation of a number of important aspects, like the sequence of players' moves, their choices at every stage, the (possibly imperfect) information each player has about the other player's moves when he makes a decision, and his payoffs for all possible game outcomes [35]. A game in extensive form is represented using a game tree, which is composed of nodes and branches. Each non-terminal node represents a move and the departing branches represent actions associated with the move (at every node, it is one player's turn to move). The sequence of moves that precedes a node is the history of the game up to that point. A player then chooses his action for every history after which it is his turn to play. A play corresponds to a path through the tree, from the root to a terminal node. A payoff for each player is then associated to each terminal node (outcome of the game). A simple example of a 2-player game in extensive form is shown in Figure 2.1, representing a two-step (sequential) interaction where $\{a, b\}$ are the actions available

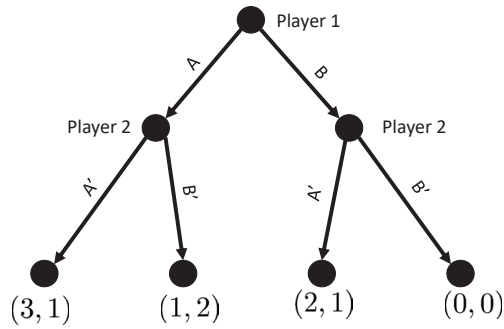


Figure 2.1: Example of game in extensive form.

to the first player in Step 1 (when it is his turn to move), while $\{a', b'\}$ are the actions of the second player in Step 2. The pairs with the payoffs for both players are shown for every leaf. Given a game, determining the best strategy that each player should follow to maximize his payoff is not easy, all the more that a profile which is optimum for both players may not exist.

A common goal in Game Theory is to determine the existence of equilibrium points, i.e., profiles that in *some sense* represent a *satisfactory* choice for all the players. Strategic and extensive form games are characterized by different equilibrium notions and can be studied using different tools.

In this monograph, we focus on non-cooperative, 2-player, strategic games.

2.2.1 Strategic Games

The strategic form, also called normal form, is the basic type of games studied in non-cooperative Game Theory. A game in strategic form lists each players' strategies, and the outcomes that result from each possible combination of choices. In the 2-player case, a strategic game is defined as a quadruple $G(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, where $\mathcal{S}_1 = \{s_{1,1}, \ldots, s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1}, \ldots, s_{2,n_2}\}$ are the sets of strategies (or actions) the first and the second player can choose from, and $u_l(s_{1,i}, s_{2,j}), l = 1, 2$ is the payoff of the game for player l, when the first player chooses the strategy $s_{1,i}$ and the second chooses $s_{2,j}$. A pair of strategies $(s_{1,i}, s_{2,j})$

Table 2.1: Example of game representation in normal form

	$s_{2,1}$	$s_{2,2}$
$s_{1,1}$	$(u_1(s_{1,1}, s_{2,1}), u_2(s_{1,1}, s_{2,1}))$	$(u_1(s_{1,1}, s_{2,2}), u_2(s_{1,1}, s_{2,2}))$
$s_{1,2}$	$(u_1(s_{1,2}, s_{2,1}), u_2(s_{1,2}, s_{2,1}))$	$(u_1(s_{1,2}, s_{2,2}), u_2(s_{1,2}, s_{2,2}))$
$s_{1,3}$	$(u_1(s_{1,3}, s_{2,1}), u_2(s_{1,3}, s_{2,1}))$	$(u_1(s_{1,3}, s_{2,2}), u_2(s_{1,3}, s_{2,2}))$

is called a profile and it corresponds to an outcome of the game. Games in strategic form are compactly represented by matrices, referred to as *payoff matrices*. For the 2-player case, one player is considered as the row player, and the other as the column player. Each row or column represents a strategy (which is the move selected by the player) and each entry in the matrix represents the payoff, that is the outcome of the game for each player for every combination of strategies. A simple example of 2-player game in normal form (with three strategies for the row player and two strategies for the column player) is shown in Table 2.1. The row player is Player 1 and the column player is Player 2. The entries of the table for each pair of strategies are the payoffs of the players.

A particular class of 2-player strategic games is the class of strictly-competitive games. In a *strictly-competitive* game, also referred to as *zero-sum*, the two players have opposite goals; in this case, the two payoff functions are strictly related to each other since for any profile we have $u_1(s_{1,i}, s_{2,j}) + u_2(s_{1,i}, s_{2,j}) = 0$. In other words, the win of a player is equal to the loss of the other player. In the particular case of a zero-sum game, then, only one payoff function needs to be defined, referred to as payoff of the game. The payoff of the game, generally denoted by u, can be defined by adopting the perspective of one of the two players, e.g., without loss of generality, $u(s_{1,i}, s_{2,j}) = u_1(s_{1,i}, s_{2,j})$, with the understanding that, for the second player, $u_2(s_{1,i}, s_{2,j}) = -u(s_{1,i}, s_{2,j})$. In the most common formulation of strategic games, the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players (game with perfect information [35, Chapter 5]). In addition, as discussed before, it is assumed that the players choose their strategies before

starting the game so that they have no hints about the strategy actually chosen by the other player.

Nash Equilibrium

The most popular notion of equilibrium for strategic games is due to Nash [36], [37]. For a 2-player game, a profile (s_{1,i^*}, s_{2,j^*}) is a Nash equilibrium if

$$\begin{aligned} u_1(s_{1,i^*}, s_{2,j^*}) &\geq u_1(s_{1,i}, s_{2,j^*}) \quad \forall s_{1,i} \in \mathcal{S}_1 \\ u_2(s_{1,i^*}, s_{2,j^*}) &\geq u_2(s_{1,i^*}, s_{2,j}) \quad \forall s_{2,j} \in \mathcal{S}_2, \end{aligned} \tag{2.9}$$

where for a zero-sum game $u_2 = -u_1$. Then, a profile is a Nash equilibrium if no player can improve his payoff by unilaterally changing his strategy. The notion of Nash equilibrium captures a steady state of a strategic game, however the process whereby the steady state is reached is not examined.

For zero-sum games, Nash equilibria have interesting properties. Let (s_{1,i^*}, s_{2,j^*}) be the Nash equilibrium of a 2-player zero-sum game G. Then, s_{1,i^*} maximizes the first player's payoff in the worst-case scenario, i.e., assuming that the second player selects his most profitable strategy corresponding to the most harmful move for the first player. Similarly, s_{2,j^*} maximizes the second player worst-case payoff. We also have

$$\max_{s_{1,i} \in \mathcal{S}_1} \min_{s_{2,j} \in \mathcal{S}_2} u_1(s_{1,i}, s_{2,j}) = \min_{s_{2,j} \in \mathcal{S}_2} \max_{s_{1,i} \in \mathcal{S}_1} u_1(s_{1,i}, s_{2,j}) = u_1(s_{1,i^*}, s_{2,j^*}). \tag{2.10}$$

As a consequence of relation (2.10), if many equilibria exist, they all yield the same payoff. In a 2-player game, a player's min-max value is always equal to his max-min value, and both are equal to the Nash equilibrium value as shown by Von Neumann's Minimax Theorem [38].

It is possible to show that the solution of Equation (2.10) can be found by solving two separate Linear Programming (LP) problems [39, Chapter 8], one for each player that moves first (corresponding to the outer maximization of the max-min and to the outer

minimization of the min-max).[4] Since the problems are *duals*,[5] it turns out that only one LP has to be solved to find the optimal strategies for the players [40], [41].

Equation (2.9) defines a pure-strategy Nash equilibrium profile, where the equilibrium strategies for the players are the *pure* strategies s_{1,i^*} and s_{2,j^*}. More generally, a Nash equilibrium can be defined in mixed strategies. A *mixed* strategy for a player is defined as a probability distribution over his set of (pure) strategies. This allows for a player to randomize the choice over his set of strategies. Since probabilities are continuous, there are infinitely many mixed strategies available to the player. More formally, given a 2-player game $G(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, let $\Pi(\mathcal{Z})$ be the set of all the probability distributions over the set $\mathcal{Z} = \{z_1, \ldots, z_k\}$. Then, the *set of mixed strategies* for a player i is formed of all probability distributions over his strategy set $\mathcal{S}_i$, namely, $\Pi(\mathcal{S}_i)$, and the set of mixed strategy profiles is the cartesian product of single mixed strategy sets $\Pi(\mathcal{S}_1) \times \Pi(\mathcal{S}_2)$. When mixed strategies are adopted by the players, the *expected payoff* can be computed for them.

An important result in Game Theory states that every strategic game with finite sets of strategies for the players has *at least* one Nash equilibrium in mixed strategies [36].

Dominance Solvable Games

Despite its popularity, the practical meaning of Nash equilibrium is often unclear, since there is no guarantee that the players will end up playing at the Nash equilibrium. A particular kind of strategic games for which stronger forms of equilibrium exist are the so-called *dominance solvable games* [37]. The concept of dominance-solvability is directly related to the notion of dominant and dominated strategies. A strategy is said to be *strictly dominant* for one player if it is the best strategy for the player, i.e., the strategy which maximizes the payoff, no matter what

[4] LP deals with the maximization or minimization of a linear objective function, subject to linear equality and inequality constraints. The admissible region of an LP problem is then a convex polytope, namely, a set defined as the intersection of finitely many half-spaces.

[5] We refer to [39, Chapter 15], for the concept of duality.

the strategy of the opponent is. Reasonably, when one such strategy exists for one of the players, he will surely adopt it. In a similar way, we say that a strategy $s_{l,i}$ is strictly dominated by strategy $s_{l,j}$, if the payoff achieved by player l choosing $s_{l,i}$ is always lower than that obtained by playing $s_{l,j}$ regardless of the strategy of the other player. Formally, in the 2-player case, we say that strategy $s_{1,i}$ is *strictly dominated* by strategy $s_{1,k}$ for player 1 if

$$u_1(s_{1,k}, s_{2,j}) > u_1(s_{1,i}, s_{2,j}) \quad \forall s_{2,j} \in \mathcal{S}_2. \tag{2.11}$$

Accordingly, a strictly dominant strategy is a strategy which strictly dominates all the other strategies.

Recursive elimination of dominated strategies is a possible technique for solving dominance-solvable games working as follows: in the first step, all the dominated strategies are removed from the set of available strategies, since no rational player would ever choose them. In this way, a new, smaller game is obtained. At this point, some strategies, that were not dominated before, may be dominated in the new game, and hence are eliminated as well. The process goes on until no dominated strategy exists for any player. A *rationalizable equilibrium* is any profile that survives the iterative elimination of dominated strategies [42], [43]. If at the end of the process only one profile is left, the remaining profile is said to be the *only rationalizable equilibrium* of the game, which is also the only Nash equilibrium point. A dominance solvable game is a game that can be solved according to the procedure described above. It goes without saying that the concept of rationalizable equilibrium is a much stronger notion than that of the Nash equilibrium, and its practical meaning is easier to grasp [44, Chapter 3]: in fact, under the assumption of rational players, we can anticipate that the players will choose the strategies corresponding to the unique rationalizable equilibrium. While *every* game with finitely many players, each of whom has finitely many pure strategies, has a Nash equilibrium in mixed strategies, a rationalizable equilibrium only exists for dominance solvable games. Another, related, interesting notion of equilibrium is that of dominant equilibrium. A *dominant equilibrium* is a profile which corresponds to dominant strategies for both players and is the strongest kind of equilibrium that a strategic game may have.

Continuous Games

The concept of continuous game extends the notion of discrete game, where the players choose from a finite sets of pure strategies. Continuous strategic games include sets of pure strategies which may be uncountably infinite [45, Chapter 3, p. 140]. More specifically, continuous games are a special case of the broad category of infinite games (i.e., games with infinite strategy sets) with the following main features: the number of players is finite, the sets of strategies are compact sets and the payoff functions are continuous. An important property of such games is that they can be approximated with a sequence of finite games corresponding to a successively finer discretization of the original game. As a consequence, all the main concepts stated for discrete games (Nash equilibrium, dominance solvability, ...) can be extended to this category of games. In particular, it is possible to prove that every continuous game has a mixed strategy Nash equilibrium (Glicksberg theorem) [46].

2.3 Introduction to Optimal Transport (OT)

The theory of optimal transportation (OT) has its origins in the eighteenth century when the problem of transporting resources at a minimal cost was first formalized by the mathematician Monge [47]. The problem of "déblais and remblais" addressed by Monge is the following: given a pile of soil and a hole (of the same volume), filling the hole with the soil from the pile with the minimum effort. Equivalently, Monge's problem is the one of moving a certain amount of soil from a source location to a sink location by minimizing some cost function of the transportation per unit of mass (see Figure 2.2 for a pictorial illustration of Monge's problem of mass transportation). The pile and hole can be modeled as probability measures μ and ν, defined on some spaces X and Y. For any A and B, measurable subsets of X and Y respectively, $\mu(A)$ gives the measure of the amount of soil located within A, while $\nu(B)$ tells how much soil can be piled in B. The cost of moving one unit of mass from location $x \in X$ to location $y \in X$ is denoted by $c(x, y)$, which is assumed nonnegative. Then c: $X \times Y \to \mathbb{R} \cup \{+\infty\}$.

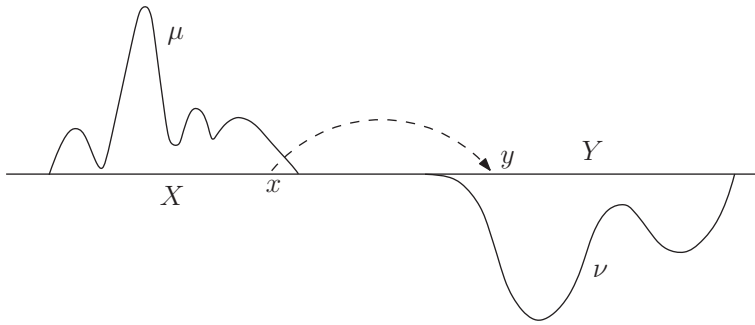


Figure 2.2: The mass transportation problem addressed by Monge.

Similarly, source and sink can be regarded two as two different ways of piling up a certain amount of soil and the goal is to move one pile into the other introducing the minimum average transportation cost.

Below, we give a rigorous formulation of the mass transportation problem. In doing so, we consider the (modern) relaxed version of the original transportation problem due to Kantorovich [48, p. 2], known as Monge–Kantorovich optimal transportation problem.

The *transportation map* is modeled as a probability measure π on the product space $X \times Y$. The quantity $d\pi(x, y)$ measures the quantity of mass moved from x to y (the mass at a given location x in X may be split into several parts moved to different destinations y in Y). An admissible transportation map π has to satisfy, for every x and y,

$$\int_Y d\pi(x, y) = d\mu(x), \quad \int_X d\pi(x, y) = d\nu(y), \tag{2.12}$$

that is, the mass taken from x equals $d\mu(x)$, and the mass moved to y equals $d\nu(x)$. Then, for all measurable subsets A of X and B of Y, π has to satisfy

$$\pi(A \times Y), \quad \pi(X \times B) = \nu(B). \tag{2.13}$$

Then, an admissible map π has marginals μ and ν. The set of admissible maps is denoted by $\Pi(\mu, \nu)$.

Solving the OT problem corresponds to searching for a map π with minimum transportation cost associated to it. Formally, the Monge–Kantorovich transportation problem (TP) corresponds to the following

minimization problem:

$$\min_{\pi \in \Pi(\mu,\nu)} \int_{X \times Y} c(x,y) d\pi(x,y). \tag{2.14}$$

The original Monge formulation of the TP differs from Kantorovich formulation for the fact that it requires that the mass be not split. Then, a unique destination y is associated to each location x. In this case, the transfer can be (more simply) described by a function $T\colon X \to Y$. Then, π has the special property

$$d\pi(x,y) \equiv d\mu(x)\delta(y = T(x)), \tag{2.15}$$

where δ is the Kronecker delta defined before. The transportation function T has to be a bijection and, for any measurable set $B \subseteq Y$, we must have $\nu(B) = \mu(T^{-1}(B))$. This property is compactly indicated as $T\#\mu = \nu$. Then, Monge addresses the following problem:

$$\min_{T\colon T\#\mu=\nu} \int_X c(x, T(x)) d\mu(x). \tag{2.16}$$

2.3.1 The Hitchcock Transportation Problem (HTP)

The one-dimensional discrete version of the Monge–Kantorovich mass transportation problem [49, Chapter 1], is also known as the Hitchcock Transportation Problem (HTP).

The general formulation of the HTP is the following:

$$\begin{aligned} \min_{\{x(i,j)\in\mathbb{R},\ \forall i,j\}} \quad & \sum_{i=1}^{m}\sum_{j=1}^{n} c(i,j)x(i,j) \\ \text{subject to} \quad & \sum_{j=1}^{n} x(i,j) = a_i, \quad i = 1,\dots,m \\ & \sum_{i=1}^{m} x(i,j) = b_j, \quad j = 1,\dots,n \\ & x(i,j) \geq 0, \end{aligned} \tag{2.17}$$

where a_i (called supplies), $i = 1,\dots,m$, represent the source pile, and b_j (called demands), $j = 1,\dots,n$, the sink or destination pile. Without loss of generality, we can assume that the problem is balanced, that

is $\sum_{i=1}^{m} a_i = \sum_{j=1}^{n} b_j$, with $a_i, b_j \geq 0$ (note that a^m and b^n do not necessarily sum to 1). Moreover, $c(i,j) \geq 0$.

The Hitchcock Transportation Problem in (2.17) is an LP problem (see Footnote 4), hence, as such, it can be solved by direct application of LP solving algorithms.[6] One of the most famous is the simplex algorithm [39, Chapter 19], that has polynomial-time complexity.

A greedy algorithm that can be used to solve the HTP in some special cases is presented in Section 2.3.2.

Probabilistic Version of the HTP

Given two sources X and Y, with pmf's P_X and P_Y defined over the same alphabet $\mathcal{X}$, we can interpret P_X and P_Y as two different ways of piling up the same amount of soil. The *joint pmf* S_{XY}, usually referred to as P_{XY}, can be regarded as the *transportation map* moving P_X into P_Y. By adopting an OT point of view, $S_{XY}(i,j)$ denotes the quantity of soil shipped from location i in P_X to j in P_Y. We let $d(i,j)$ be the cost, sometimes referred to as distortion, associated to the modification of the i-th symbol of the alphabet into the j-th one. The transportation map that minimizes the average distortion necessary to move P_X into P_Y can be obtained by solving the following constrained minimization problem:

$$\min_{S_{XY}:\, \sum_y S_{XY}=P_X, \sum_x S_{XY}=P_Y} \sum_{i,j} d(i,j) S_{XY}(i,j), \tag{2.18}$$

where $\sum_x S_{XY}$ is a short form for $\sum_i S_{XY}(i,j)$. The minimization in (2.18) is a particular version of the Hitchcock Transportation Problem (HTP), where the source and destination piles are probability distributions.

In the rest of the monograph, the acronym TP always refers to the discrete probabilistic formulation in (2.18).

Due to the earth transportation analogy, in computer vision applications, the minimum in Equation (2.18) is often known as *Earth Mover Distance* (*EMD*) between P_X and P_Y [51], and is denoted by

[6] To be more specific, for the expert reader, HTP is a particular minimum cost flow problem [50, Section 1.2].

$EMD_d(P_X, P_Y)$ (the subscript d is sometimes omitted for brevity). The term *EMD* is used in general, also when the source and sink piles are general mass functions (as in the TP formulation in (2.17)). When the soil piles are probability mass functions, and $d(i,j) = l(i,j)^p$ for some distance measure l (with $p \geq 1$), the *EMD* has a more general statistical meaning. Let X and Y be two random variables with probability distributions P_X and P_Y; the *EMD* between P_X and P_Y corresponds to the minimum expected p-th power distance between X and Y taken over all joint probability distributions P_{XY} with marginal distributions respectively equal to P_X and P_Y:

$$EMD_{l^p}(P_X, P_Y) = \min_{P_{XY}:\, \sum_y P_{XY}=P_X, \sum_x P_{XY}=P_Y} E_{XY}[l(X,Y)^p]. \quad (2.19)$$

In transport theory terminology, expression (2.19) is the p-th power of the Wasserstein distance [49, Chapter 1, p. 40], [52, Chapter 6, p. 105], or the Monge–Kantorovich metric of order p [48, Chapter 7, p. 207], [53]. In particular, when the L_2^2 distance is considered and then $d(i,j) = |i-j|^2$ (i.e., $l(i,j) = |i-j|$ and $p = 2$), the Earth Mover Distance, namely $EMD_{L_2^2}(P_X, P_Y)$, is equivalent to the squared Mallows distance between P_X and P_Y [54].[7] In the following, we will continue to refer to (2.18) as $EMD(P_X, P_Y)$.

2.3.2 Hoffman's (Greedy) Algorithm

Let $X \sim P_X$ and $Y \sim P_Y$ be discrete sources defined on the sets $\mathcal{X}$ and $\mathcal{Y}$ respectively.[8] The TP in (2.18), that has to be solved for computing $EMD_d(P_Y, P_X)$, is a Linear Programming problem. In general, the solution of TP depends on the cost function $d(\cdot,\cdot)$, however there are some classes of cost functions for which the solution can be found through a simple greedy algorithm. Specifically, the algorithm proposed by A. J. Hoffman in 1963 [55], allows to solve the transportation problem

[7] L_2^2 indicates the squared Euclidean distance (similarly, throughout the monograph, L_p^p denotes the p-power of the L_p distance).

[8] In this chapter, we refer to $\mathcal{X}$ and $\mathcal{Y}$ as possibly different subsets of bins from the same alphabet, where P_X and P_Y, have non-zero mass (non-empty bins).

whenever $d(\cdot,\cdot)$ satisfies the so-called Monge property [56], that is when:

$$d(i,j) + d(r,s) \leq d(i,s) + d(r,j), \tag{2.20}$$

for all (i,j,r,s) such that $1 \leq i < r \leq |\mathcal{X}|$ and $1 \leq j < s \leq |\mathcal{Y}|$.

It is easy to verify that the Monge property is satisfied by any cost function of the form $d(i,j) = |i-j|^p$, and, more generally, by any convex function of the quantity $|i-j|$. The iterative procedure proposed by Hoffman to solve the optimal transport problem is known as *North-West Corner (NWC) rule* [55] and works as follows: take the bin of $\mathcal{X}$ with the smallest value and start moving its elements into the bin with the smallest value in $\mathcal{Y}$. When the smallest bin of $\mathcal{Y}$ is filled, go on with the second smallest bin in $\mathcal{Y}$. Similarly, when the smallest bin in $\mathcal{X}$ is emptied, go on with the second smallest bin in $\mathcal{X}$. The procedure is iterated until all the bins in $\mathcal{X}$ have been moved into those of $\mathcal{Y}$. Let i^{low} (i^{up}) and j^{low} (j^{up}) denote the lower (upper) bins of $\mathcal{X}$ and $\mathcal{Y}$ respectively. A pseudocode description of the *NWC* rule is given below.

1. Initialize: $i := i^{low}$, $j := j^{low}$.
2. Set $S_{XY}(i,j) := \min\{P_X(i), P_Y(j)\}$.
3. Adjust the "supply" distribution $P_X(i) := P_X(i) - S_{XY}(i,j)$ and the "demand" distribution $P_Y(j) := P_Y(j) - S_{XY}(i,j)$. If $P_X(i) = 0$ then $i := i+1$ and if $P_Y(j) = 0$ then $j := j+1$.
4. If $j < j^{up}$ or $P_Y(j^{up}) > 0$, go back to Step 2.

The above procedure is described graphically in Figure 2.3. For the sake of clarity the figure shows two distributions with disjoint supports, however this assumption is not necessary for the validity of the procedure. Interestingly, the *NWC* rule does not depend explicitly on the cost matrix, so the transportation map obtained through it is the same regardless of the Monge cost. According to Hoffman's greedy algorithm, when the cost function satisfies Monge's property, the *EMD* can be computed in linear running time: the number of elementary operations, in fact, is at most equal to $|\mathcal{X}| + |\mathcal{Y}|$. This represents a dramatic simplification with respect to the complexity required to solve a general

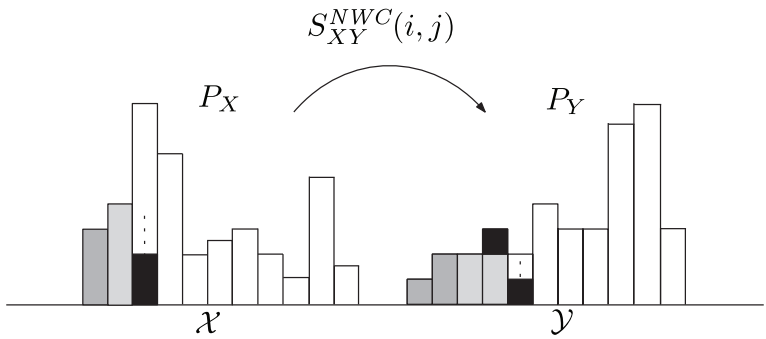


Figure 2.3: Graphical representation of the NWC rule for the Monge transportation problem. P_X and P_Y are two generic soil piles (source and sink) defined on $\mathcal{X}$ and $\mathcal{Y}$, while $S_{XY}^{NWC}(i,j)$ denotes the amount of soil moved from location i to j.

Hitchcock transportation problem (the interested reader may refer to [57] for more details).

2.4 Elements of Large Deviation Theory

Large deviation theory deals with rare events, whose probability is exponentially small, and has applications in many different scientific fields. As a matter of fact, most of the results presented in this monograph can be seen as the solution of large deviation theory problems. The mathematical machinery used to prove the main results of large deviation theory relies on the methods of types, whose main properties are stated in the section below. Moreover, many results in the monograph are derived exploiting a generalization of Sanov's theorem [58, Chapter 12, p. 292], [59], provided in Section 2.4.2.

2.4.1 Basics of the Method of Types

The method of types is a powerful technique in large deviation theory.

Following the notation introduced in the beginning of this chapter, for a given sequence x^n over a finite alphabet $\mathcal{X}$, the type (or empirical probability distribution) of the sequence is defined as $P_{x^n}(a) = N(a\,|\,x^n)/n$, $a \in \mathcal{X}$, where $N(a\,|\,x^n)$ is the number of times symbol a

occurs in the sequence x^n. For a type $P \in \mathcal{P}_n$, the type class is defined as $\mathcal{T}(P) = \{x^n \in \mathcal{X}^n \colon P_{x^n} = P\}$.

In the following, we summarize the main results concerning the method of types.

- The number of types is at most polynomial in n, that is $|\mathcal{P}_n| \leq (n+1)^{|\mathcal{X}|}$.
- If the n random variables are drawn i.i.d. according to P_X, the probability of x^n depends only on its type and is given by

$$P_X(x^n) = 2^{-n(H(P_{x^n}) + \mathcal{D}(P_{x^n} \| P_X))}. \tag{2.21}$$

- For any type $P \in \mathcal{P}_n$, the size of the type class $\mathcal{T}(P)$ can be bounded as $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{nH(P)} \leq |\mathcal{T}(P)| \leq 2^{nH(P)}$.
- For any $P \in \mathcal{P}_n$ and any distribution P_X, the probability of the type class $\mathcal{T}(P)$ under P_X can be bounded as $\frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P\|P_X)} \leq P_X(\mathcal{T}(P)) \leq 2^{-n\mathcal{D}(P\|P_X)}$.

The second to last statement asserts that the cardinality of the type class $\mathcal{T}(P)$ grows exponentially with n, with exponent $H(P)$, that is[9] $|\mathcal{T}(P)| \doteq 2^{nH(P)}$, while its probability tends to zero exponentially fast with decay rate given by $\mathcal{D}(P\|P_X)$, that is, $P_X(\mathcal{T}(P)) \doteq 2^{-n\mathcal{D}(P\|P_X)}$ (last bullet point).

The above relations allow us to estimate the behavior of (asymptotically) long sequences based on the properties of the type. For instance, given a long sequence of samples drawn i.i.d. according to a given distribution, the type of the sequence is close to the generating distribution.[10] In fact, $P_X(\mathcal{T}(P)) \to 0$ for any $P \neq P_X$.

For the derivation of the above results and for more insights into the use of types and type classes in information theory and statistics, interested readers are referred to [58], [60].

[9] Notation $a_n \doteq b_n$ indicates equality to the first order in the exponent, that is, $\lim_{n\to\infty} 1/n \log(a_n/b_n) = 0$.

[10] Ultimately, this is nothing but another way of formulating the Weak Law of Large Numbers.

2.4.2 Generalized Sanov's Theorem

Sanov's theorem constitutes one of the main results in large deviation theory. The statement of this theorem and its generalization, which underlie the results derived in this monograph, are provided in the following.

We start by giving the following definition.

Definition 2.4. The probability simplex in $\mathcal{R}^m$ is the set of points $\{x^m \in \mathbb{R}^m\colon \sum_{i=1}^m x_i = 1, x_i \geq 0\ \forall i\}$.

The space $\mathcal{P}$ of probability distributions defined over a finite alphabet $\mathcal{X}$ is then geometrically represented by the probability simplex in $\mathcal{R}^{|\mathcal{X}|}$, that is,

$$\mathcal{P} = \left\{P \in \mathbb{R}^{|\mathcal{X}|}\colon \sum_{i=1}^{|\mathcal{X}|} P(a) = 1, P(a) \geq 0,\ \forall a \in \mathcal{X}\right\}. \tag{2.22}$$

For $|\mathcal{X}| = 3$, the probability simplex $\mathcal{P}$ is the 2-dimensional manifold represented in Figure 2.4.

Given a sequence of n i.i.d. random variables drawn according to a source distribution P, we denote by P_n the empirical pmf of the sequence.[11] Let $E \subseteq \mathcal{P}$ be a set of probability distributions. From the properties of types, it is easy to argue that if E does not contain P, or a neighborhood of P, then the probability that P_n belongs to E tends to zero as the length n of the sequence of r.v.'s tends to infinity (weak law of large numbers [58, Chapter 3, p. 57]). Sanov's theorem [59], [61, Chapter 2, p. 16], [58, Chapter 12, p. 292] calculates the exponent of the (vanishing) probability that P_n belongs to E, stating that

$$\begin{aligned}\inf_{Q\in E} \mathcal{D}(Q\|P) &\leq -\limsup_{n\to\infty} \frac{1}{n} \log P(P_n \in E)\\ &\leq -\liminf_{n\to\infty}\ \frac{1}{n} \log P(P_n \in E)\\ &\leq \inf_{Q\in int\ E} \mathcal{D}(Q\|P),\end{aligned} \tag{2.23}$$

where *int* E denotes the interior part of the set E.

[11] For brevity, in this section we use notation P_n in place of P_{x^n} for the empirical pmf.

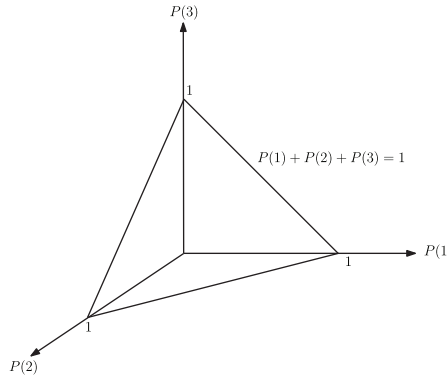


Figure 2.4: Probability simplex $\mathcal{P}$ for $|\mathcal{X}| = 3$.

When $cl(E) = cl(int(E))$,[12] or, $E \subseteq cl(int(E))$, the left- and right-hand side of (2.23) coincide, and we get the exact rate:

$$-\lim_{n\to\infty} \frac{1}{n} \log P(P_n \in E) = \inf_{Q\in E} \mathcal{D}(Q\|P). \tag{2.24}$$

Let Q^* denote the distribution yielding the infimum in (2.24). A geometric illustration of Sanov's theorem in the probability simplex is given in Figure 2.5.

If we define the set $E_n = E \cap \mathcal{P}_n$, we have: $P(P_n \in E) = P(P_n \in E_n)$ and we can rewrite Sanov's theorem as:

$$\begin{aligned} \inf_{Q\in E} \mathcal{D}(Q\|P) &\leq -\limsup_{n\to\infty} \frac{1}{n} \log P(P_n \in E_n) \\ &\leq -\liminf_{n\to\infty} \frac{1}{n} \log P(P_n \in E_n) \\ &\leq \inf_{Q\in int\ E} \mathcal{D}(Q\|P). \end{aligned} \tag{2.25}$$

We now extend the formulation of Sanov's theorem to more general sequences of sets E_n for which we do not necessary have $E_n = E \cap \mathcal{P}_n$ for some set E.

We start by introducing the notion of convergence for sequences of subsets due to Kuratowski, which is a more general notion of convergence compared to the one based on the Hausdorff distance. Let (S, d) be a

[12] $cl(E)$ denotes the closure of E.

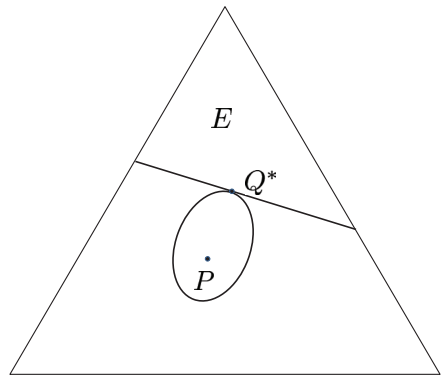


Figure 2.5: Geometric interpretation of Sanov's theorem.

metric space. We first provide the definition of *lower closed limit* or Kuratowski limit inferior [62, Chapter 29, p. 335].

Definition 2.5. A point p belongs to the lower limit $\underset{n\to\infty}{Li}\, K_n$ (or simply LiK_n) of a sequence of sets K_n, if every neighborhood of p intersects all the K_n's from a sufficiently great index n onward.

The expression $p \in \underset{n\to\infty}{Li}\, K_n$ is equivalent to the existence of a sequence of points $\{p_n\}$ such that:

$$p = \lim_{n\to\infty} p_n, \quad p_n \in K_n. \tag{2.26}$$

Stated in another way, LiK_n is the set of the accumulation points of sequences in K_n. An alternative, equivalent, definition is

$$\underset{n\to\infty}{Li}\, K_n = \Big\{p \in X \text{ s.t. } \limsup_{n\to\infty} d(x, K_n) = 0\Big\}. \tag{2.27}$$

Similarly, the following definition of *upper closed limit* or Kuratowski limit superior [62, Chapter 29, p. 335] can be given.

Definition 2.6. A point p belongs to the upper limit $\underset{n\to\infty}{Ls}\, K_n$ (or simply LsK_n) of a sequence of sets K_n, if every neighborhood of p intersects an infinite number of terms in K_n.

The expression $p \in \underset{n\to\infty}{Ls}\, K_n$ is equivalent to the existence of a subsequence of points $\{p_{k_n}\}$ such that

$$k_1 < k_2 < \cdots, \quad p = \lim_{n\to\infty} p_{k_n}, \quad p_{k_n} \in K_{k_n}.$$

An alternative, equivalent, definition is

$$\underset{n\to\infty}{Ls}\, K_n = \Big\{p \in X \text{ s.t. } \liminf_{n\to\infty} d(x, K_n) = 0\Big\}. \tag{2.28}$$

It can be proven that the Kuratowski limit inferior and superior are always closed sets [62, Chapter 29, pp. 335 and 337].

Given the above, we can state the following.

Definition 2.7. The sequence of sets $\{K_n\}$ is said to be convergent to K in the sense of Kuratowski, that is $K_n \overset{K}{\to} K$, if $LiK_n = K = LsK_n$, in which case we write $K = LimK_n$.

Kuratowski convergence is weaker than convergence in Hausdorff metric; in fact, given a sequence of closed sets $\{K_n\}$, $K_n \overset{H}{\to} K$ implies $K_n \overset{K}{\to} K$ [63]. For compact metric spaces, the reverse implication also holds and the two kinds of convergence coincide.

In this monograph, we are interested in the space $\mathcal{P}$ of probability mass functions defined over a finite alphabet $\mathcal{X}$, i.e., the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$, equipped with the L_1 metric. Being $\mathcal{P}$ a closed subset of $\mathbb{R}^{|\mathcal{X}|}$, $\mathcal{P}$ is a complete set. In addition, with the L_1 metric, $\mathcal{P} \in \mathcal{L}(\mathbb{R}^{|\mathcal{X}|})$, that is, $\mathcal{P}$ is bounded. The space $(\mathcal{P}, d_{L_1})$, then, is a compact metric space and therefore, for our purposes, Kuratowski and Hausdorff convergence are equivalent.

With the above ideas in mind, the following generalization of Sanov's theorem can be proven. We use notation $E_{(n)}$ to denote the dependence on n of a generic set in $\mathcal{P}$, and we let $E_n = E_{(n)} \cap \mathcal{P}_n$.

Theorem 2.1 (Generalized Sanov's Theorem)**.** Let $\{E_{(n)}\}$ be a sequence of sets in $\mathcal{P}$, such that $Li(E_{(n)} \cap \mathcal{P}_n) \neq \emptyset$. Then:

$$\begin{aligned}\min_{Q \in\, LsE_{(n)}} \mathcal{D}(Q\|P) &\le -\limsup_{n\to\infty} \frac{1}{n} \log P(P_n \in E_{(n)}) \\ &\le -\liminf_{n\to\infty} \frac{1}{n} \log P(P_n \in E_{(n)}) \\ &\le \min_{Q \in\, Li\,(E_{(n)} \cap \mathcal{P}_n)} \mathcal{D}(Q\|P). \end{aligned} \tag{2.29}$$

If, in addition, $LsE_{(n)} = Li(E_{(n)} \cap \mathcal{P}_n)$, the generalized Sanov's limit exists as follows:

$$-\lim_{n\to\infty} \frac{1}{n} \log P(P_n \in E_{(n)}) = \min_{Q \in LimE_{(n)}} \mathcal{D}(Q\|P). \tag{2.30}$$

The theorem follows from the properties of the Kuratowski limit superior and the boundedness of the probability simplex $\mathcal{P}$. We refer to Appendix A in [5] for the proof.

In general, the Kuratowski convergence of $E_{(n)}$ is a *necessary* condition for the existence of the generalized Sanov limit in (2.30), but it is not sufficient. In fact, we may have $LiE_{(n)} \supseteq Li(E_{(n)} \cap \mathcal{P}_n)$, in which case the lower and upper bound in (2.29) do not coincide. We notice that when $E_{(n)} \in \mathcal{P}_n$ is a sequence of sets in $\mathcal{P}_n$, then Sanov's limit holds whenever $E_{(n)} \xrightarrow{K} E$ for some set E, or, by exploiting the compactness of $\mathcal{P}$, $E_{(n)} \xrightarrow{H} E$. Based on the above observation, we can state the following corollary.

Corollary 2.2. Let $E_{(n)}$ be a sequence of sets in $\mathcal{P}_n$, such that $E_{(n)} \xrightarrow{H} E$. Then:

$$-\lim_{n\to\infty} \frac{1}{n} \log P(P_n \in E_{(n)}) = \min_{Q\in E} \mathcal{D}(Q\|P). \tag{2.31}$$

As a final observation, it is straightforward to argue that, when $\{E_{(n)}\} = E \; \forall n$ (or from a certain n on), the generalized Sanov's theorem corresponds to the classical Sanov's theorem. In this case, in fact, we have that $LsE_{(n)} = E$, while $Li(E_n) = Li(E \cap \mathcal{P}_n) \supseteq intE$ (since every $p \in int(E)$ is the accumulation point of a sequence in $E \cap \mathcal{P}_n$), and the original Sanov's bounds are obtained.

3

Binary Detection Game with Known Sources

In this chapter, we study the binary detection problem when both the Defender and the Attacker have full knowledge of the sources underlying the two hypotheses. The problem is first cast into a rigorous game-theoretic framework by modeling the interplay between the Defender and the Attacker as a zero-sum game. Under some simplifying assumptions, we derive the optimal strategies of the two players and the equilibrium point of the game. We then analyze the asymptotic payoff at equilibrium, i.e., the limit payoff when the length of the observed sequence tends to infinity, determining under which conditions the Defender succeeds in making a correct detection notwithstanding the presence of the Attacker.

3.1 Detection Game with Known Sources (DG-KS)

In this chapter, we consider the simplest version of the detection game according to which the pmf's P_X and P_Y are known to $\mathscr{D}$ and $\mathscr{A}$. The assumption that the source Y is known to $\mathscr{D}$ may seem a questionable choice, since, in many practical applications, it could be difficult for $\mathscr{D}$ to have full access to the source Y. We will see, however, that, at least asymptotically, the assumption that $\mathscr{D}$ knows Y can be removed, thus leading to a more realistic model. In order to limit the complexity

of the problem and make the analysis tractable, we limit the kind of acceptance regions $\mathscr{D}$ can choose from. Specifically, we confine the decision to depend on a limited set of statistics computed on the test sequence. Such an assumption, according to which the detector has access to a limited set of empirical statistics of the sequence, is referred to as *limited resources assumption* (see [64] for an introduction on this terminology). In particular, we limit the analysis carried out by the detector to first order statistics, which are sufficient statistics for the case of memoryless sources ([58, Section 2.9]). Hence, we require that $\mathscr{D}$ bases his decision by relying only on P_{z^n}, i.e., on the empirical probability distribution induced by the test sequence z^n. Note that, strictly speaking, P_{z^n} is not a sufficient statistics for the test under H_1: in fact, even if Y is a memoryless source, $\mathscr{A}$ could introduce some memory within the sequence as a result of the attack. This is the reason why we need to introduce explicitly the requirement that $\mathscr{D}$ bases his decision only on the empirical distribution, that is, on first order statistics. While the limited resources assumption is mainly introduced to simplify the analysis, we observe that the use of first order statistics is pretty common in a number of application scenarios even if the sources under analysis are not memoryless. In multimedia forensics, for instance, several techniques have been proposed which rely on the analysis of the image histogram or a subset of statistics derived from it (see, for example, [65], [66]). As another example, the analysis of statistics derived from the histograms of block-DCT coefficients is often adopted for detecting multiple JPEG compression [67]. More generally, the assumption of limited resources is reasonable in application scenarios where the detector has a small computational power. Eventually, we emphasize that the theory presented in this and the subsequent chapters can be extended to richer sets of empirical statistics, as long as a suitable extension of the method of types is available (e.g., for Markov source).

A fundamental consequence of the limited resources assumption is that it forces Λ^n to be a union of type classes, i.e., if z^n belongs to Λ^n, then the whole type class of z^n, namely $\mathcal{T}(P_{z^n})$, will be contained in Λ^n. Since a type class is univocally defined by the empirical probability distribution of the sequences contained in it, the acceptance region Λ^n

can be defined as a union of types $P \in \mathcal{P}_n$, where $\mathcal{P}_n$ is the set of all possible types with denominator n.

Moreover, we focus on the *asymptotic* behavior of the game, that is, the behavior when the length n of the observed sequence tends to infinity.

With the above ideas in mind, the binary detection game with known sources (DG-KS) is defined as follows.

Definition 3.1. The DG-KS $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game is a zero-sum, strategic, game played by $\mathscr{D}$ and $\mathscr{A}$, defined by the following strategies and payoff.

- *Defender's strategies.* The set of strategies $\mathscr{D}$ can choose from is the set of acceptance regions for H_0 for which the false positive probability is below a certain threshold:

$$\mathcal{S}_{\mathscr{D}} = \{\Lambda^n \in 2^{\mathcal{P}_n} \colon P_{\text{FP}} \leq 2^{-\lambda n}\}, \tag{3.1}$$

  where $P_{\text{FP}} = P_X(z^n \notin \Lambda^n)$ and $2^{\mathcal{P}_n}$ denotes the power set of $\mathcal{P}_n$, i.e., all the possible unions of types[1] and we require that the false positive error probability decays exponentially fast with n, with an exponential rate *at least* as large as λ.

- *Attacker's strategies.* The set of strategies of $\mathscr{A}$ is formed by all the functions that map a sequence $y^n \in \mathcal{X}^n$ into a new sequence $z^n \in \mathcal{X}^n$ subject to a distortion constraint:

$$\mathcal{S}_{\mathscr{A}} = \{g(\cdot) \colon d(y^n, g(y^n)) \leq nL\}, \tag{3.2}$$

  where $d(\cdot, \cdot)$ is a proper distortion function and L is the maximum allowed average per-letter distortion.

- *Payoff function.* The payoff of the game is defined in terms of the false negative error probability (P_{FN}), namely:

$$u(\Lambda^n, g) = -P_{\text{FN}} = -P_Y(z^n \in \Lambda^n) = -\sum_{y^n \colon g(y^n) \in \Lambda^n} P_Y(y^n), \tag{3.3}$$

  where $\mathscr{D}$ aims at maximizing u, while $\mathscr{A}$ wishes to minimize it.

[1] We will refer to Λ^n as a union of sequences or a union of types interchangeably, the two perspectives being equivalent and clearly understandable from the context.

Regarding the distortion measure $d(\cdot,\cdot)$, throughout the monograph we always consider the most common case of additive distortion. The only exception are Sections 3.5 and 4.5, where we consider a distortion measure based on the infinity norm.[2] We also point that $d(\cdot,\cdot)$ does not need to be a distance, that is why we adopt the general term distortion. For instance, in the following, we will use both the L_1 distortion, corresponding to the L_1 distance, and the L_2^2 distortion, which corresponds to a squared distance. Regarding the constraint imposed by the Attacker, since L is the maximum average per-letter distortion, $\mathscr{A}$ is not forced to introduce a distortion that is lower than L for each sample of the sequence.

Before going on, we pause to clarify some of the choices behind the formulation of the DG-KS game. First of all, the strategies available to $\mathscr{A}$ are limited to deterministic functions of y^n. This may seem a limiting choice, however we will see that, at least asymptotically, i.e., when the length of n tends to infinity, the optimal strategy of $\mathscr{D}$ does not depend on the strategy chosen by $\mathscr{A}$, then, it does not make sense for $\mathscr{A}$ to adopt a randomized strategy to confuse $\mathscr{D}$. The second comment regards the assumption that $\mathscr{D}$ knows P_Y. As it is evident from Equation (3.3), this is a necessary assumption, since for a proper definition of the game it is required that both players have a full knowledge of the payoff for all possible profiles. As we will see later, the asymptotically optimal strategy of $\mathscr{D}$ does not depend on P_Y, thus making the assumption that $\mathscr{D}$ knows P_Y irrelevant.

3.2 Solution of the DG-KS Game

The solution of the DG-KS game passes through the following lemma.

[2] Part of results derived in the monograph, e.g., the derivation of the equilibrium strategies (for all the versions of the binary detection game), can be stated for a wide class of distortion measures. Other results, e.g., the characterization of the attack by means of optimum transportation and the computation of the payoff at the equilibrium, require that some assumptions are made on the distortion measure. In our treatment, we privileged the simplicity of the analysis and hence we avoid to state each time the minimum set of requirements needed to prove the various results. The reader may refer to [68] for a more detailed analysis.

Lemma 3.1. Let $\Lambda^{n,*}$ be defined as follows:

$$\Lambda^{n,*} = \left\{ P \in \mathcal{P}_n \colon \mathcal{D}(P\|P_X) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n} \right\}, \tag{3.4}$$

sometimes referred to as $\Lambda^{n,*}(P_X)$ or $\Lambda^{n,*}(P_X, \lambda)$. Then we have:

1. $P_{\text{FP}} \leq 2^{-n(\lambda-\delta_n)}$, with $\delta_n \to 0$ for $n \to \infty$,
2. for every $\Lambda^n \in \mathcal{S}_{\mathscr{D}}$ (with $\mathcal{S}_{\mathscr{D}}$ defined as in (3.1)) we have $\bar{\Lambda}^n \subseteq \bar{\Lambda}^{n,*}$.

Hence, $\Lambda^{n,*}$ is a *dominant strategy* for $\mathscr{D}$.

Proof. Since $\bar{\Lambda}^{n,*}$ and $\Lambda^{n,*}$ are unions of type classes, $P_{\text{FP}}(\Lambda^{n,*})$ can be rewritten as

$$P_{\text{FP}}(\Lambda^{n,*}) = \sum_{P \in \bar{\Lambda}^{n,*}} P_X(\mathcal{T}(P)), \tag{3.5}$$

where $P_X(\mathcal{T}(P))$ denotes the collective probability (under P_X) of all the sequences in $\mathcal{T}(P)$. For the class of DMS sources, the number of types is upper bounded by $(n+1)^{|\mathcal{X}|}$ and the probability of a type class $\mathcal{T}(P)$ by $2^{-n\mathcal{D}(P\|P_X)}$ (see Section 2.4.1), hence we have:

$$\begin{aligned} P_{\text{FP}}(\Lambda^{n,*}) &\leq (n+1)^{|\mathcal{X}|} \max_{P \in \bar{\Lambda}^{n,*}} P_X(\mathcal{T}(P)) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \bar{\Lambda}^{n,*}} \mathcal{D}(P\|P_X)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\left(\lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\right)} \\ &= 2^{-n\left(\lambda - 2|\mathcal{X}|\frac{\log(n+1)}{n}\right)}, \end{aligned} \tag{3.6}$$

proving the first part of the lemma with $\delta_n = 2|\mathcal{X}|\frac{\log(n+1)}{n}$ and where the last inequality derives from (3.4).

We now pass to the second part of the lemma. Let Λ^n be in $\mathcal{S}_D$ and let P be in $\bar{\Lambda}^n$. Then we have (see Section 2.4.1 for a justification of the last inequality):

$$\begin{aligned} 2^{-\lambda n} &\geq P_X(\bar{\Lambda}^n) \\ &\geq P_X(\mathcal{T}(P)) \\ &\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P\|P_X)}, \end{aligned} \tag{3.7}$$

which, by taking the logarithm of both sides, proves that $P \in \bar{\Lambda}^{n,*}$. □

The first relation proved in Lemma 3.1 says that, asymptotically, $\Lambda^{n,*}$ defines a valid strategy for $\mathscr{D}$, while the second one implies the optimality of $\Lambda^{n,*}$. In fact, if for a certain strategy of $\mathscr{A}$ we have that $P \notin \bar{\Lambda}^{n,*}$, *a fortiori* we have that $P \notin \bar{\Lambda}^n$ for any other choice of $\bar{\Lambda}^n$ hence resulting in a larger false negative error probability.

Some interesting consequences of the lemma are the following. The optimal strategy for $\mathscr{D}$ does not depend on the strategy chosen by $\mathscr{A}$. By adopting a game-theoretic terminology this means that the best defence strategy is a *dominant* one. As a further consequence, the optimal defence strategy does not depend on P_Y, meaning that the optimal strategy is *universal* with respect to Y in $\mathcal{C}$, i.e., it is optimal across all the sources under the alternative hypothesis (H_1). As we anticipated, this result makes the assumption that $\mathscr{D}$ knows P_Y irrelevant. In the same way, it is not necessary for $\mathscr{D}$ to know the probability distribution of the attacked sequences.

The result stated in Lemma 3.1 corresponds to the well known Hoeffding test for the non-adversarial case [69].

We now pass to the determination of the optimal strategy of $\mathscr{A}$. The existence of a dominant strategy for $\mathscr{D}$ significantly simplifies the search for the optimal attack strategy. In fact, since a rationale Defender will surely play the dominant strategy $\Lambda^{n,*}$, $\mathscr{A}$ can choose her strategy by assuming that $\Lambda^n = \Lambda^{n,*}$. In this way, the derivation of the optimal attack becomes an easy task. By observing that the goal of $\mathscr{A}$ is to maximize P_{FN}, we argue that such a goal is obtained by trying to bring the sequences produced by Y within $\Lambda^{n,*}$, i.e., by trying to reach the condition:

$$\mathcal{D}(P_{g(y^n)} \| P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n}. \tag{3.8}$$

In doing so $\mathscr{A}$ must only respect the constraint that $d(y^n, g(y^n)) \leq nL$. The optimal strategy for $\mathscr{A}$ can then be expressed as follows:[3]

$$g^*(y^n) = \arg \min_{z^n:\, d(z^n, y^n) \leq nL} \mathcal{D}(P_{z^n} \| P_X). \tag{3.9}$$

Together with Lemma 3.1, the above observation permits to state our first fundamental result, summarized in the following theorem.

[3] In general, the minimization in (3.9) may have multiple solutions, all of them equivalent, i.e., leading to the same value of the payoff.

Theorem 3.2 (Equilibrium Point of the DG-KS Game)**.** The DG-KS game is a dominance solvable game and the profile $(\Lambda^{n,*}, g^*)$ is a rationalizable equilibrium of the game.

Proof. Lemma 3.1 asserts that $\Lambda^{n,*}$ is a strictly dominant strategy for $\mathscr{D}$, thus permitting us to eliminate all the other strategies in $\mathcal{S}_{\mathscr{D}}$ (since they are strictly dominated by $\Lambda^{n,*}$). The theorem, then, follows by observing that g^* satisfies

$$-u(\Lambda^{n,*}, g^*) \geq -u(\Lambda^{n,*}, g) \quad \forall g \in \mathcal{S}_{\mathscr{A}}, \tag{3.10}$$

that is, g^* maximizes the false negative error probability for a fixed $\Lambda^{n,*}$. In fact, for any to-be-attacked sequence y^n, whenever the minimum in (3.9) is not lower than the acceptance threshold, no other strategy will succeed in bringing y^n inside the acceptance region; hence, $\mathscr{A}$ maximizes the false negative probability, namely $P_Y(g(y^n) \in \Lambda^{n,*})$, by playing g^*. □

3.2.1 Characterization of the Game by means of Transportation Theory

In this section we show that, thanks to the permutation invariance ensured by additive distortion measures, we can consider an interesting reformulation of the game. More specifically, we can look at the Attacker's strategy from a different perspective, by drawing a parallelism with *transportation theory* (see Section 2.3), which permits to derive a very intuitive and insightful interpretation of the optimal Attacker's strategy, opening the way to the analysis of the source distinguishability performed in Chapter 4.

Given a sequence y^n drawn from Y, the goal of $\mathscr{A}$ is to transform it into a sequence z^n belonging to the acceptance region chosen by $\mathscr{D}$. Let us denote by $n(i,j)$ the number of times that the i-th symbol of the alphabet is transformed into the j-th one as a consequence of the attack. Similarly, we denote by $S^n_{YZ}(i,j) = n(i,j)/n$ the fraction of times the i-th symbol of the alphabet is transformed into the j-th one. In the following we will refer to $S^n_{YZ}(i,j)$ as *transportation map*. The fact that S^n_{YZ} refers to n-long sequences is explicitly indicated by adding the superscript n.

For any additive distortion measure d, we have

$$d(y^n, z^n) = \sum_i d(y_i, z_i) = \sum_{i,j} n(i,j) d(i,j), \tag{3.11}$$

where $d(i,j)$ is the distortion introduced when the symbol i is transformed into the symbol j. Hence, d is permutation-invariant (see Definition 2.1), and the average per-symbol distortion between y^n and z^n can be expressed in terms of S^n_{YZ} as

$$d(y^n, z^n)/n = \sum_{i,j} S^n_{YZ}(i,j) d(i,j). \tag{3.12}$$

The map S^n_{YZ} determines also the empirical distribution (i.e., the type) of the attacked sequence. In fact, by denoting with $P_{z^n}(j)$ the relative frequency of symbol j within z^n, we have

$$P_{z^n}(j) = \sum_i S^n_{YZ}(i,j) \triangleq S^n_Z(j). \tag{3.13}$$

Since $\mathscr{A}$ can not change more symbols than there are in the sequence y^n, a map S^n_{YZ} can be applied to a sequence y^n only if $S^n_Y(i) \triangleq \sum_j S^n_{YZ}(i,j) = P_{y^n}(i)$. Accordingly, S^n_{YZ} can be interpreted as the *joint empirical pmf* (i.e., the joint type) of the sequences y^n and z^n. In the same way, S^n_Y and S^n_Z correspond, respectively, to the empirical pmf's of y^n and z^n.

By remembering that Λ^n depends only on the empirical pmf of the test sequence, and given that the empirical pmf of the attacked sequence depends on S^n_Z only through S^n_{YZ}, we can define the action of the Attacker as the choice of a transportation map among all *admissible* maps, a map being admissible if:

$$\begin{cases} S^n_Y = P_{y^n} \\ \sum_{i,j} S^n_{YZ}(i,j) d(i,j) \le L, \end{cases} \tag{3.14}$$

which is a set of linear constraints in S^n_{YZ}. The set of the admissible maps is denoted by $\mathcal{A}^n(L, P_{y^n})$.

Given the above, the space of strategies of the Attacker can be seen as the set of all the possible ways of associating an admissible

transformation map to the to-be-attacked sequence. In the following, we will refer to the result of such an association as $S^n_{YZ}(y^n)$, or $S^n_{YZ}(i,j;y^n)$, when we need to refer explicitly to the relative frequency with which the symbol i is transformed into the symbol j. In the same way, $S^n_Z(j;y^n)$ denotes the output marginal of $S^n_{YZ}(i,j;y^n)$. With regard to the input marginal, we always have $S^n_Y(i;y^n) = P_{y^n}(i)$. Similarly, we use the notation $S^n_Y(y^n)$ to denote the pmf P_{y^n}. By adopting this symbolism, the space of strategies of $\mathscr{A}$ can be redefined as:

$$\mathcal{S}_{\mathscr{A}} = \{S^n_{YZ}(y^n)\colon S^n_{YZ}(i,j) \in \mathcal{A}^n(L, P_{y^n})\}. \tag{3.15}$$

We can also rewrite the payoff function as follows

$$u(\Lambda^n, S^n_{YZ}) = - \sum_{y^n\colon\, S^n_Z(y^n)\in\Lambda^n} P_Y(y^n). \tag{3.16}$$

By adopting the above transportation theory perspective, Theorem 3.2 can be rephrased as follows.

Corollary 3.3 (Equilibrium Point of the DG-KS Game)**.** Let

$$\Lambda^{n,*} = \left\{P \in \mathcal{P}_n\colon \mathcal{D}(P\|P_X) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\right\}, \tag{3.17}$$

and

$$S^{n,*}_{YZ}(y^n) = \arg \min_{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})} \mathcal{D}(S^n_Z\|P_X). \tag{3.18}$$

Then $\Lambda^{n,*}$ is a dominant equilibrium for $\mathscr{D}$ and the profile $(\Lambda^{n,*}, S^{n,*}_{YZ}(y^n))$ is the only rationalizable equilibrium of the DG-KS game, which, then, is a dominance solvable game.

While the formula defining the optimum acceptance region in (3.17) can be easily implemented by the Defender, the task of the Attacker is more complex due to the necessity of solving the minimization problem in (3.18). However, we notice that the number of variables in the minimization in (3.18) is quadratic in $|\mathcal{X}|$, thus representing a dramatic improvement with respect to the minimization in (3.9) where the number of variables involved in the minimization is n. By further inspecting (3.18), we see that such a minimization resembles an *optimal transport* problem, however it departs from it since the divergence

between a source pmf and a target one is minimized in (3.18), subject to a distortion constraint, whereas, OT faces with the somewhat-dual problem of minimizing the distortion needed to make the two pmf's equal (we will also adopt this perspective in Chapter 4 when we will analyze the limiting performance of the game). Since $S^n_{YZ}(i,j) \in \mathbb{Q}_n$ (i.e., $n(i,j) = n \cdot S^n_{YZ}(i,j) \in \mathbb{N}$), the problem in (3.17) is an integer minimization problem. By observing that the divergence term $\mathcal{D}(S^n_Z \| P_X)$ is convex as a function of the transportation map S^n_{YZ}, see [58, Theorem 2.7.2] (the dependence of S^n_Z in S^n_{YZ} is linear and the divergence is convex), if the admissibility set is defined by convex constraints in S^n_{YZ}, as it is the case with any additive distortion, the minimization problem in (3.18) is a convex integer optimization problem, for which a unique *global* optimal solution exists. Such a solution can be found by using common optimization algorithms implemented by existing solvers for convex MINLP (Mixed Integer Nonlinear Problems) [70], [71], see [72].

3.3 Analysis of the Payoff at the Equilibrium

The next step of our analysis focuses on the computation of the payoff at the equilibrium. Specifically, given the asymptotic nature of the game, we will evaluate the error exponent of the false negative error probability at the equilibrium, i.e., ε^* (see Section 2.5). As a result, at the equilibrium, P_{FN} will either tend to 0 or not for $n \to \infty$ depending on the relationship between the maximum allowed distortion and the KL-divergence between P_X and P_Y. As to the false positive error exponent, namely η^*, in the setup defined by the DG-KS game we always have $\eta^* \geq \lambda$ (see (3.1)).

To start with, let Γ^n be the set of sequences generated by Y that can be moved into $\Lambda^{n,*}$ as a consequence of the attack, that is

$$\Gamma^n(P_X, \lambda, L) = \{y^n \colon \exists\ z^n \in \Lambda^{n,*}(P_X, \lambda) \text{ s.t. } d(y^n, z^n) \leq nL\}\,. \quad (3.19)$$

Accordingly, the false negative error probability is equal to the probability that the sequence y^n belongs to this set, that is $P_{\text{FN}} = P_Y(y^n \in \Gamma^n)$.

Proposition 3.1. The set $\Gamma^n(P_X, \lambda, L)$ defined in (3.19) is a union of type classes for any permutation invariant distance-measure.

The above proposition can be easily proven by observing that $\Lambda^{n,*}$ depends on the observed sequence only via the type class and that, whenever the distance measure is permutation invariant, the action of $\mathscr{A}$ is equivalent to the application of a transportation map $S_{YZ}^{n,*}(y^n)$.

The set in (3.19) can then be easily redefined in terms of types instead of sequences:[4]

$$\Gamma^n(P_X, \lambda, L) = \{P \in \mathcal{P}_n \colon \exists\ S_{PV}^n \in \mathcal{A}^n(L, P) \text{ s.t. } V \in \Lambda^{n,*}(P_X, \lambda)\}. \tag{3.20}$$

The above region defines all the type classes (with denominator n) whose sequences can be moved within $\Lambda^{n,*}$ by $\mathscr{A}$. In order to decide whether the sequences generated by two generic sources (not necessarily belonging to $\mathcal{P}_n$) can be eventually distinguished as n tends to infinity, we now investigate the asymptotic behavior of P_{FN} at the equilibrium.

To do so, it is convenient to introduce the asymptotic version of $\Gamma^n(P_X, \lambda, L)$, defined as follows:

$$\Gamma(P_X, \lambda, L) = \{P \in \mathcal{P} \colon \exists\ S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Lambda^*(P_X, \lambda)\}, \tag{3.21}$$

where

$$\Lambda^*(P_X, \lambda) = \{P \in \mathcal{P} \colon \mathcal{D}(P\|P_X) \leq \lambda\}. \tag{3.22}$$

In the same way, the definitions of $S_{PV}(i, j)$ and $\mathcal{A}(L, P)$ are obtained immediately from those of $S_{PV}^n(i, j)$ and $\mathcal{A}^n(L, P)$, by relaxing the requirement that $S_{PV}(i, j)$ and $P(i)$ are rational numbers with denominator n.

We now have all the necessary tools to prove the following theorem.

Theorem 3.4 (Asymptotic Payoff of the DG-KS Game)**.** For the DG-KS game, the error exponent of the false negative error probability at the equilibrium is given by:[5]

$$\varepsilon^* = \min_{P \in \Gamma(P_X, \lambda, L)} \mathcal{D}(P\|P_Y), \tag{3.23}$$

[4] With a slight abuse of notation, we denote with S_{PV}^n the transportation map from a pmf $P \in \mathcal{P}_n$ to another pmf $V \in \mathcal{P}_n$, when the sequences that induce the pmf's and their underlying sources are not specified. The same notation is used in other parts of the monograph.

[5] The use of the minimum instead of the infimum is justified by the compactness of $\Gamma(P_X, \lambda, L)$ which is demonstrated within the proof itself.

leading to the following cases:

1. $\varepsilon^* = 0$, if $P_Y \in \Gamma(P_X, \lambda, L)$;
2. $\varepsilon^* \neq 0$, if $P_Y \notin \Gamma(P_X, \lambda, L)$.

Proof. In order to derive the error exponent of the false negative probability, we must evaluate the following limit:

$$\varepsilon^* = -\lim_{n\to\infty} \frac{1}{n} \log(P_Y(P_n \in \Gamma^n)), \tag{3.24}$$

namely, the error exponent of the probability of the sequence of sets Γ^n (we use lim – instead of the lim sup – because, as we will show, such limit exists). The computation of the above limit can be carried out by applying the generalization of Sanov's theorem reported in Section 2.4.2. In order to apply the theorem to this case, it is sufficient to show that, for a given distance measure d: $\mathcal{P} \times \mathcal{P} \to \mathbb{R}^+$, Γ^n tends to Γ in the Hausdorff metric δ_H, that is, $\Gamma^n \xrightarrow{H} \Gamma$ (see Corollary 2.2 in Section 2.4.2).[6]

Due to the convexity and continuity of the divergence function w.r.t. its arguments, and the density of rational numbers into the real ones, the Hausdorff distance between $\Lambda^{n,*}$ and Λ^* gets smaller as n increases, meaning that $\delta_H(\Lambda^{n,*}, \Lambda^*) \to 0$ as $n \to \infty$ (and hence, $\Lambda^{n,*} \xrightarrow{H} \Lambda^*$). We now show that such a property can be extended to the sets Γ^n and Γ. To this purpose, it is convenient to rewrite Γ and Γ^n in a slightly different manner, by considering the *inverse transportation map* that moves a distribution out of the acceptance region, that is

$$\Gamma(P_X, \lambda, L) = \{P \in \mathcal{P}\colon \exists\, S_{VP} \in \mathcal{A}(L, V), \text{for some } V \in \Lambda^*(P_X, \lambda)\}. \tag{3.25}$$

The equivalence of definitions (3.25) and (3.21) follows from the fact that for any map S_{PV} that moves P into V, the inverse map S_{VP} moves V into P by introducing the same distortion.[7] A similar equivalence holds for

[6] We remind that, for computing the Hausdorff distance, the distance measure d between pmf's must be such that $\mathcal{P}$ endowed with d is bounded (see discussion in Section 2.4.2).

[7] We are implicitly assuming that the element-wise distortion $d(i, j)$ is symmetric, i.e., $d(i, j) = d(j, i)\ \forall (i, j)$, which holds in all the cases considered in this monograph.

the set $\Gamma^n(P_X, \lambda, L)$. Being $\Gamma^n \subseteq \Gamma$ (which is obvious from the definition of Γ^n and Γ), any pmf P in Γ^n also belongs to Γ, and hence $\delta_\Gamma(\Gamma^n) = \sup_{P\in\Gamma^n} \inf_{P'\in\Gamma} d(P', P) = 0$. In order to show that $\delta_H(\Gamma^n, \Gamma) \to 0$ as $n \to \infty$ we must prove that $\delta_{\Gamma^n}(\Gamma) = \sup_{P\in\Gamma} \inf_{P'\in\Gamma^n} d(P, P') \to 0$ as $n \to \infty$.

Let us fix $P_1 \in \Gamma$. Let V_1 be a pmf in $\Lambda^*(P_X, \lambda)$ such that $S_{V_1P_1} \in \mathcal{A}(L, V_1)$. We can choose a point $V_2 \in \Lambda^{n,*}(P_X, \lambda)$ such that $d(V_1, V_2) \leq \delta_H(\Lambda^{n,*}, \Lambda^*)$. By exploiting the fact that $\delta_H(\Lambda^{n,*}, \Lambda^*)$ tends to zero as $n \to \infty$, V_2 can be taken arbitrarily close to V_1 for large enough n. According to Theorem A.2 (Appendix A), it is possible to move V_2 into a pmf P_2 close to P_1 with a map in $\mathcal{A}^n(L, V_2)$; by construction, $P_2 \in \Gamma^n$. Specifically (see the proof of Theorem A.2), for any given P_1 and $S_{V_1P_1}$, the map $S^n_{P_2V_2} \in \mathcal{A}^n(L, V_2)$ can be chosen in such a way that $P_2 \in \mathcal{B}(P_1, e_n)$[8] with $e_n = (2/n + \delta_H(\Lambda^{n,*}, \Lambda^*)) \cdot |\mathcal{X}|^2$. Accordingly, $\inf_{P\in\Gamma^n} d(P, P_1) \leq d(P_2, P_1) \leq e_n$, $\forall P_1$. Then, $\delta_{\Gamma^n}(\Gamma) = \sup_{P'\in\Gamma} \inf_{P\in\Gamma^n} d(P, P') \leq e_n$ which tends to zero, as $n \to \infty$, thus concluding the proof. □

The main consequence of Theorem 3.4 is that, given P_X, L and λ, the set of sources P_Y can be split into two distinct regions: the subset for which, as a consequence of the attack, the false negative error exponent is 0 ($P_Y \in \Gamma(P_X, \lambda, L)$) and the subset for which the false negative error exponent is positive and then the false negative probability P_{FN} tends to zero exponentially fast ($P_Y \in \bar{\Gamma}(P_X, \lambda, L)$). Stated in another way, given two pmf's P_X and P_Y, a maximum attack distortion L and the desired false positive error exponent λ, Theorem 3.4 permits to understand whether $\mathscr{D}$ may succeed in making the false negative error probability tend to zero exponentially fast. In the following we will refer to such a case by saying that $\mathscr{D}$ *wins* the game. When this is not possible, the false negative error probability may either tend to a finite value strictly larger than 0 (possibly 1) or tend to zero at a sub-exponential rate. In both cases, we will say that $\mathscr{A}$ *wins* the game and that the source X and Y can not be distinguished reliably.

[8] For any point $P \in \mathcal{P}$, $\mathcal{B}(P, \tau)$ denote the neighborhood of P of radius τ, according to the metric d.

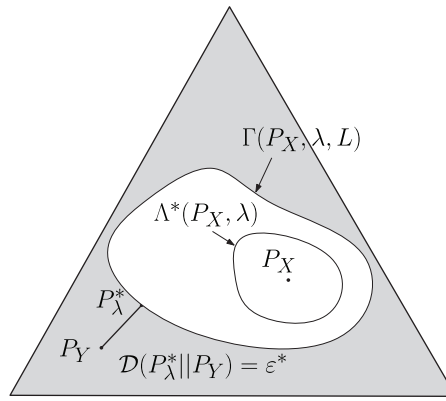


Figure 3.1: Geometric interpretation of $\Gamma(P_X, \lambda, L)$ and $\Lambda^*(P_X, \lambda)$ in the probability simplex by the light of Theorem 3.4.

With the above ideas in mind, $\Gamma(P_X, \lambda, L)$ can be interpreted as the region with the sources that cannot be *reliably distinguished* from P_X guaranteeing a false positive error exponent at least equal to λ in the presence of an adversary with allowed distortion L. Accordingly, we say that $\Gamma(P_X, \lambda, L)$ represents the *indistinguishability region* of the adversarial detection test in the DG-KS setup. A geometric interpretation of Theorem 3.4 is given in Figure 3.1.

In general, the expression of Γ does not allow an analytic computation of the pmf's P_Y that $\mathscr{D}$ is not able to distinguish from P_X. In the next section, we consider a simple case in which a closed-form expression can be found for Γ.

3.3.1 Bernoulli Sources

To exemplify the results of Theorem 3.4, we consider the particular case in which the distortion constraint is expressed in terms of the Hamming distance and we specialize the expression of Γ to such a case (it is easy to see that the Hamming distance satisfies the conditions under which the theorems in the previous sections have been proved). In such a case, in fact, a closed-form expression can be found for Γ thus greatly simplifying the analysis. The simplification relies on the following lemma.

Lemma 3.5. If $d(y^n, z^n) = d_H(y^n, z^n)$, the set Γ^n can be expressed as:

$$\Gamma^{n,*}(P_X, \lambda, L) = \{P \in \mathcal{P}_n \colon \exists P' \in \Lambda^{n,*}(P_X, \lambda) \text{ s.t. } d_{L_1}(P, P') \leq 2L\} \tag{3.26}$$

where d_{L_1} is the L_1 distance between P and P' (sometimes called variational distance).

Proof. We start by proving that a sequence whose type has a L_1 distance larger than $2L$ from all the types in $\Lambda^{n,*}$ cannot belong to Γ^n_H. Let y^n and z^n be two sequences, and let P_{y^n} and P_{z^n} be their types. The distance between P_{y^n} and P_{z^n} can be rewritten as follows:

$$\begin{aligned} d_{L_1}(P_{y^n}, P_{z^n}) &= \sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)] \\ &\quad + \sum_{a \in \mathcal{X}^-} [P_{z^n}(a) - P_{y^n}(a)] \\ &= 2 \sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)], \end{aligned} \tag{3.27}$$

where $\mathcal{X}^+$ (res. $\mathcal{X}^-$, $\mathcal{X}^=$) denotes the set of a's for which $P_{y^n}(a) > P_{z^n}(a)$ (res. $P_{y^n}(a) < P_{z^n}(a)$, $P_{y^n}(a) = P_{z^n}(a)$), and where the last equality follows from the observation that:

$$\sum_{a \in \mathcal{X}^-} P_{y^n}(a) = 1 - \sum_{a \in \mathcal{X}^+} P_{y^n}(a) - \sum_{a \in \mathcal{X}^=} P_{y^n}(a). \tag{3.28}$$

Let us consider now the Hamming distance between the sequences y^n and z^n. By considering $\mathcal{X}^+$, we see that $d_H(y^n, z^n)$ is larger than or equal to $\sum_{a \in \mathcal{X}^+} n[P_{y^n}(a) - P_{z^n}(a)]$. In fact, for each $a \in \mathcal{X}^+$, there must be at least $n[P_{y^n}(a) - P_{z^n}(a)]$ positions in which the sequences y^n and z^n differ, so to justify the presence of $n[P_{y^n}(a) - P_{z^n}(a)]$ more a's in y^n than in z^n, thus yielding:

$$d_{L_1}(P_{y^n}, P_{z^n}) \leq \frac{2 d_H(y^n, z^n)}{n}. \tag{3.29}$$

For the sequences y^n whose type does not satisfy (3.26), we have $d_{L_1}(P_{y^n}, P_{z^n}) > 2L \ \forall z^n \in \Lambda^{n,*}$, yielding

$$2L < d_{L_1}(P_{y^n}, P_{z^n}) \leq \frac{2 d_H(y^n, z^n)}{n}, \tag{3.30}$$

showing that $\Gamma^n \subseteq \Gamma^{n,*}$.

We now show that $\Gamma^{n,*} \subseteq \Gamma^n$. Let P be a type in $\Gamma^{n,*}$. Then there exists a type $P' \in \Lambda^{n,*}$ whose distance from P is lower than or equal to $2L$. Let y^n be a sequence belonging to $T(P)$, the type class of P. Starting form y^n we can easily build a new sequence z^n whose type is equal to P' by proceeding as follows. Let $\mathcal{X}^+$ be the set of a's for which $P_{y^n}(a) > P'(a)$. For each $a \in \mathcal{X}^+$ we take $n[P_{y^n}(a) - P'(a)]$ positions where $y_i = a$, and replace a with a value $b \in \mathcal{X}^-$, in such a way that at the end we have $P_{z^n}(a) = P'(a)\ \forall a \in \mathcal{X}$. Note that this is always possible as we have

$$\sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P'(a)] = \sum_{b \in \mathcal{X}^-} [P'(b) - P_{y^n}(b)]. \tag{3.31}$$

Since to pass from y^n to z^n we modified only $\sum_{a \in \mathcal{X}^+} n[P_{y^n}(a) - P'(a)]$ positions of y^n we have:

$$\begin{aligned} d_H(y^n, z^n) &= \sum_{a \in \mathcal{X}^+} n[P_{y^n}(a) - P'(a)] \\ &= \frac{n d_{L_1}(P_{y^n}, P')}{2} \\ &\leq nL, \end{aligned} \tag{3.32}$$

showing that $y^n \in \Gamma^n$, and hence $\Gamma^{n,*} \subseteq \Gamma^n$, thus concluding the proof of the lemma. □

Lemma 3.5 permits to rewrite the expression for the indistinguishability region in a simpler form:

$$\Gamma^* = \left\{ P \in \mathcal{P} \colon \exists\ P' \in \Lambda_0^*(P_X) \text{ s.t. } d_{L_1}(P, P') \leq 2L \right\}. \tag{3.33}$$

The relation between Hamming distance and L_1 distance, investigated in the proof of the Lemma 3.5, will be exploited in other parts of the monograph (for instance in Chapter 6).

We now apply the general expression found above to the case of two Bernoulli sources. For the sequences emitted by these sources, the Hamming distance is a natural choice to define the distortion constraint, thus permitting to adopt the simplified definition of Γ given in (3.33).

Let then X and Y be two Bernoulli sources with parameters $p = P_X(1)$ and $q = P_Y(1)$ respectively. In this case the acceptance region

for H_0 assumes a very simple form. In fact, the KL-divergence between P_{x^n} and P_X depends only on the number of 1's in x^n, the divergence being a monotonic increasing[9] function of $|\nu_x(1) - p|$, where we denoted by $\nu_x(1)$ the relative frequency of 1's in x^n. When seen as an union of types, the acceptance region may be defined in terms of $P(1)$ (the probability of 1 under P) only:

$$\Lambda^{n,*}(p,\lambda) = \{P \in \mathcal{P}_n \colon P(1) \in (\nu_{inf}(\lambda), \nu_{sup}(\lambda))\}\,, \tag{3.34}$$

where $\nu_{inf}(\lambda)$ and $\nu_{sup}(\lambda)$ derive from the equality

$$\mathcal{D}(P\|P_X) = \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}. \tag{3.35}$$

Note that in some cases we may have $\nu_{inf} = 0$ and/or $\nu_{sup} = 1$, since Equation (3.35) may admit a solution only for $P(1) > p$, $P(1) < p$, or no solution at all.

The optimal strategy of $\mathscr{A}$ is also easy to define. Given the monotonic nature of the KL-divergence, $\mathscr{A}$ will increase (decrease) the number of 1's in y^n to make the relative frequency of 1's in z^n as close as possible to p. $\mathscr{A}$ will succeed in inducing a decision error if the relative frequency of ones in z^n belongs to the interval (ν_{inf}, ν_{sup}). Since the distortion constraint states that $d_H(y^n, z^n) \leq nL$, we clearly have:

$$\Gamma^n(p,\lambda,L) = \{P \in \mathcal{P}_n \colon P(1) \in (\nu_{inf}(\lambda) - L, \nu_{sup}(\lambda) + L)\}\,, \tag{3.36}$$

with the boundaries of the interval truncated to 0 or 1 when needed. For the computation of the error exponent of P_{FN} at the equilibrium, we first introduce the asymptotic version of $\Lambda^{n,*}$ and Γ^n:

$$\Lambda^*(p,\lambda) = \{P \in \mathcal{P} \colon P(1) \in (\nu^\infty_{inf}(\lambda), \nu^\infty_{sup}(\lambda))\}, \tag{3.37}$$

where ν^∞_{inf} and ν^∞_{sup} are now derived from the equality

$$\mathcal{D}(P\|P_X) = \lambda. \tag{3.38}$$

Then the indistinguishability region is

$$\Gamma(p,\lambda,L) = \{P \in \mathcal{P} \colon P(1) \in [\nu^\infty_{inf}(\lambda) - L, \nu^\infty_{sup}(\lambda) + L]\}. \tag{3.39}$$

[9] Actually the KL-divergence may have an asymmetric behavior for $n_x(1) < np$ and $n_x(1) > np$ however this asymmetry does not have any impact on our analysis.

As stated by Theorem 3.4, we can distinguish two cases:

$$\begin{aligned} q = P_Y(1) &\in [\nu_{inf}^{\infty}(\lambda) - L, \nu_{sup}^{\infty}(\lambda) + L] \\ q = P_Y(1) &\notin [\nu_{inf}^{\infty}(\lambda) - L, \nu_{sup}^{\infty}(\lambda) + L]. \end{aligned} \tag{3.40}$$

In the first case $\varepsilon^* = 0$. In the second case P_{FN} tends to zero exponentially fast for $n \to \infty$ and the error exponent can be computed by resorting to Equations (3.23) and (3.33). Let us suppose for instance that $q > \nu_{sup}^{\infty} + L$. The type in $\Gamma(p, \lambda, L)$ closest to P_Y in divergence is a Bernoulli source with parameter $p^* = \nu_{sup}^{\infty} + L$, and hence the error exponent will be $\varepsilon^* = \mathcal{D}(p^* \| q)$.

3.4 Numerical Analysis: A Case Study

In this section, we resort to numerical analysis to get some insights into the best achievable performance of the game between $\mathscr{D}$ and $\mathscr{A}$ for a close-to-reality situation related to a class of camera identification problems in the field of image forensics. Let us assume that two signal sources X and Y differ by the *noisiness* of the signals they produce. In order to test the hypothesis that a signal has been generated by X, $\mathscr{D}$ applies a wavelet decomposition to the signal and considers the statistics of the DWT (Discrete Wavelet Transform) coefficients at a certain decomposition level [73]. The Defender knows that the DWT coefficients are independent and follow a Laplacian distribution $P_X(x) = \frac{\gamma}{2} e^{-\gamma |x|}$. The DWT coefficients of the signal produced by the source Y also follow a Laplacian distribution but with a different decay parameter ω. In order to distinguish between images acquired by the two sources (cameras), $\mathscr{D}$ identifies a flat region of the image and analyzes how the pixel grey levels are distributed around the mean value of the area. $\mathscr{D}$ knows that if the image has been produced by the first camera the pixels follow a Laplacian distribution with decay parameter γ, while for images acquired by the second camera, the decay parameter is equal to ω. Given a sequence y^n of DWT coefficients (or pixel gray levels) produced by Y and a distortion constraint nL, by exploiting the results of this chapter, we can derive the optimal attack strategy. Moreover, we can also investigate whether $\mathscr{D}$ can effectively distinguish between sequences (images) generated by X and Y by ensuring that the false

positive error probability tends to zero exponentially fast with error exponent at least equal to λ.

The analysis of the previous section applies to DMS sources and not to continuous sources as those considered here. We get around this problem by quantizing the continuous probability density functions (pdf's). If the quantization step is small enough, the analysis of the discrete case will provide useful indications about the continuous problem. We then quantize the Laplacian pdf's onto the set of integers by restricting the pdf to values that have a non-negligible probability of appearing in a sequence of a certain length. Specifically, the probability $P_X(i)$ is computed as:

$$P_X(i) = \int_{i-1/2}^{i+1/2} \frac{\gamma}{2} e^{-\gamma|x|} dx. \tag{3.41}$$

For the values of n, γ and ω used in our simulations, it is enough to consider values until $i = \pm 20$ since the probability that a value outside the interval $[-20, 20]$ shows up is significantly lower than one.[10] For instance, with $n = 10^6$ and $\gamma = 1$, such a probability is about 0.002. A similar procedure is adopted to discretize P_Y. Let us call $\hat{P}_X$ and $\hat{P}_Y$ the discretized versions of P_X and P_Y.

We use numerical analysis through Monte Carlo simulations, by working as follows: we generate a large number of sequences according to $\hat{P}_Y$ and perform numerical optimization to move them within $\Lambda^{n,*}$ implementing the attack in (3.18). To do so, we use a solver for convex MINLP [71] (see a discussion in the end of Section 3.2), that works by solving a relaxed version of the minimization problem in which $n(i,j)$'s are not required to be integer, that is, the quantities $S^n_{YZ}(i,j)$ are not required to be in $\mathbb{Q}_n$, but they are assumed to be continuous variables ($S^n_{YZ}(i,j) \in \mathbb{R} \cap [0,1]$). Given the convexity of the objective function, the relaxed problem can be solved efficiently by resorting to the steepest gradient descent method [74]. Once the relaxed solution has been obtained, the optimum integer solution is found by searching in the neighborhood of the relaxed minimum.

An estimate of the false negative error probability can be obtained by measuring the success rate. Figure 3.2 shows results obtained by applying

[10] For $i = 20$ we let $P_X(i) = \int_{i-1/2}^{\infty} \frac{\gamma}{2} e^{-\gamma|x|} dx$. Similarly for $i = -20$.

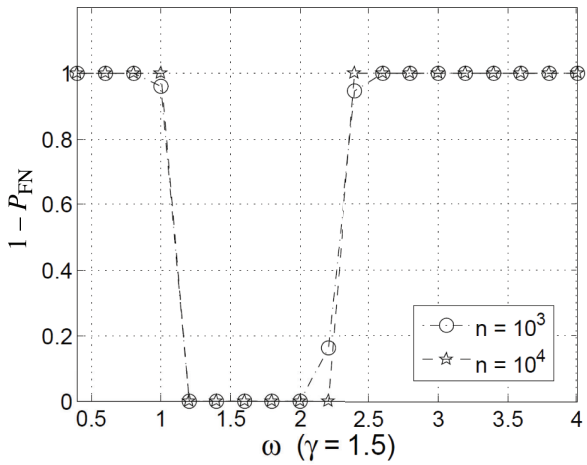


Figure 3.2: False negative error probability obtained through Monte Carlo simulations (1000 random sequences), for $L = 0.01$, $\lambda = 0.06$.

the above procedure for $\gamma = 1.5$, various values of ω, $L = 1/100$, $\lambda = 0.06$ and two different values of n ($n = 10^3$ and $n = 10^4$). Each point of the curve is obtained by generating 10^3 sequences according to $\hat{P}_Y$. The behavior of P_{FN} agrees with the insights provided by Theorem 3.4: the values of ω can be split into two main classes, those for which the false negative error probability approaches zero and those for which the false negative probability tends to 1. Of course, the former class corresponds to the cases for which ω is further from γ thus easing the job of the Defender. Such a dual behavior is more apparent for large values of n since for such values the numerical analysis gets closer to the asymptotic conditions underlying the theoretical analysis. The numerical analysis then permits, for each value of λ (and for a fixed L), to compute the minimum and maximum values of ω for which $\mathscr{D}$ is going to fail. An example of this kind of analysis is given in Figure 3.3, where the range of ω for which $\mathscr{D}$ fails is plotted as a function of λ.

3.5 DG-KS Game under Maximum Distortion-Limited Attack

We now extend the analysis of Sections 3.2 and 3.3 to the case in which the distortion measure constraining the Attacker is expressed in terms

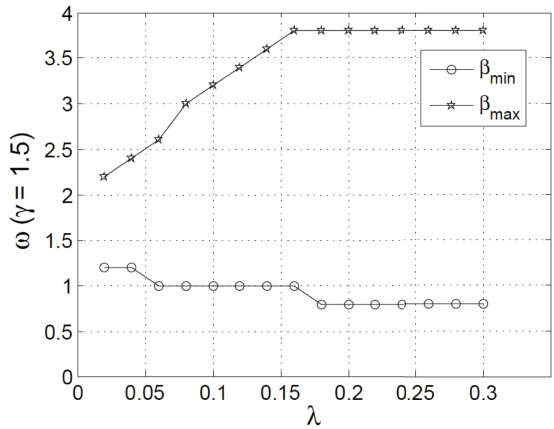


Figure 3.3: Range of ω for which increasing n results in a false negative error probability tending to 1, for $\gamma = 1.5$ and $n = 10^4$.

of the maximum absolute distance between the samples of y^n and z^n, that is, to the case in which the distortion is measured by relying on the L_∞ distance.

The particular interest in this scenario is justified by the fact that, in many practical applications, the distortion constraint must be satisfied locally, thus requiring that the maximum absolute distance between y^n and z^n is limited, rather than its average across the whole sequences. This is the case, for instance, of biomedical and remote sensing image compression, for which the maximum error introduced at each pixel location must be strictly controlled [75]. Other examples, where the use of the L_∞ distance is recommended, come from image processing applications, when it must be ensured that two versions of the same image, an original and a processed one, are visually indistinguishable [76]. In such a case, it is necessary that the absolute difference between the two images is lower than the visibility threshold (often referred to as just noticeable distortion (JND) [77]) at each pixel location.

It is easy to see that the L_∞ distance measure is a permutation invariant measure, and then it is possible to express the distortion constraint the Attacker is subject to by limiting the set of transportation maps S^n_{YZ} he can choose from, that is, by defining the set of admissible

maps similarly to (3.14). More specifically, we observe that the maximum distance between the sequences y^n and z^n can be rewritten as follows:

$$d_{L_\infty}(y^n, z^n) = \max_k |z_k - y_k| = \max_{(i,j):\, S^n_{YZ}(i,j)\neq 0} |i-j|. \tag{3.42}$$

The set of strategies of the Attacker can be defined as in (3.15), where the set of the admissible maps $\mathcal{A}^n_{L_\infty}(L, P_{y^n})$ is given by

$$\begin{cases} S^n_Y = P_{y^n} \\ \max\limits_{(i,j):\, S^n_{YZ}(i,j)\neq 0} |i-j| \leq L. \end{cases} \tag{3.43}$$

We observe that now the distortion constraint is imposed on a per-letter basis and not only on the average, and then L is the maximum allowed per-symbol distortion level.

Passing to the analysis of the indistinguishability region, it is straightforward to see that all the previous definitions continue to hold by replacing $\mathcal{A}^n(L, P_{y^n})$ with $\mathcal{A}^n_{L_\infty}(L, P_{y^n})$. In fact, the dominant strategy for $\mathscr{D}$ does not depend on the set of strategies available to $\mathscr{A}$. Let $\Gamma^n_{L_\infty}(P_X, \lambda, L)$ denote the set of the types for which $\mathscr{D}$ decides in favor of H_0 as a consequence of the attack. The asymptotic version of $\Gamma^n_{L_\infty}(P_X, \lambda, L)$ is defined as in (3.21)

$$\Gamma_{L_\infty}(P_X, \lambda, L) = \{P \in \mathcal{P}\colon \exists\ S_{YZ} \in \mathcal{A}_{L_\infty}(L, P) \text{ s.t. } S_Z \in \Lambda^*(P_X, \lambda)\}, \tag{3.44}$$

where $\mathcal{A}_{L_\infty}(L, P)$ is the asymptotic counterpart of $\mathcal{A}^n_{L_\infty}(L, P)$.

By observing that the maximum distortion constraint can be equivalently rewritten as a collection of linear constraints in S^n_{YZ}, that is

$$\max_{(i,j):\, S^n_{YZ}(i,j)\neq 0} |i-j| \leq L \longleftrightarrow S^n_{YZ}(i,j) = 0, \quad \forall i, j\colon |i-j| \leq L, \tag{3.45}$$

we deduce that the admissible set in (3.43) is a linear set. Accordingly, Theorem 3.4 also holds in the L_∞ case and the asymptotic payoff can be computed as in (3.23), with the indistinguishability region given by (3.44).

4

Limit Performance and Source Distinguishability

A drawback with the analysis carried out in Chapter 3 is the asymmetric role of the false positive and false negative error exponents, namely η and ε, which derives from the adoption of the Neyman–Pearson approach in the definition of the game. With such an approach, in fact, the Defender aims at ensuring a given value for η, namely λ, but is satisfied with any strictly positive ε. In the analysis of this chapter, we make a more reasonable assumption and say that the Defender succeeds, i.e., he is able to distinguish between X and Y despite the presence of the adversary, if – at the equilibrium – both error probabilities tend to zero exponentially fast, regardless of the particular values assumed by the error exponents. More precisely, by mimicking Chernoff-Stein's lemma [58, Section 12.8] for the non adversarial version of the test, we analyze the behavior of the indistinguishability regions of the test, namely $\Gamma(P_X, \lambda, L)$, when the false positive decay rate λ approaches 0, to see whether, given a maximum allowable distortion L, it is possible for $\mathscr{D}$ to simultaneously attain strictly positive error exponents for the two kinds of error, hence permitting to reliably distinguish between P_X and P_Y. Doing so permits to study the *ultimate achievable performance* of the detection in adversarial setting.

By exploiting the parallelism with optimal transport, we introduce the concept of Security Margin ($\mathcal{SM}$), defined as the maximum distortion introduced by the Attacker for which the two sources can be distinguished by the Defender ensuring arbitrarily small, yet positive, error exponent for Type I and II error probabilities. The $\mathcal{SM}$ is a powerful concept that permits to summarize in a single quantity the distinguishability of two sources X and Y in the adversarial setting.

4.1 Characterization of the Indistinguishability Region using OT

By adopting an optimal transport perspective, we can rewrite the indistinguishability region in (3.21) in a more compact and easier-to-interpret way, as follows

$$\Gamma(P_X, \lambda, L) = \{P \in \mathcal{P}\colon \exists\, Q \in \Lambda^*(P_X, \lambda) \text{ s.t } EMD(P, Q) \leq L\}, \quad (4.1)$$

where EMD denote the Earth Mover Distance defined in Section 2.3, Equation (2.18), where the cost function corresponds here to the distortion $d(\cdot, \cdot)$ used to constraint the strategies available to the Attacker.

Such insightful rewriting of the indistinguishability region is useful in the subsequent analysis.

4.2 Best Achievable Performance in the DG-KS Setup

We now consider the behavior of the DG-KS game as function of λ; in particular, we study the behavior of $\Gamma(P_X, \lambda, L)$ when $\lambda \to 0$. Such analysis permits to investigate whether two sources X and Y are *ultimately* distinguishable in the setting defined by the DG-KS game. The rationale behind such an analysis stems directly from the definition of the acceptance region. In fact, from the definition of $\mathcal{S}_{\mathscr{D}}$, it is easy to see that a smaller λ leads to a more favorable game for the Defender, since he can adopt a smaller acceptance region and then obtain a larger payoff. Stated in another way, from $\mathscr{D}$'s perspective, evaluating the behavior of

the game for $\lambda \to 0$ corresponds to exploring the *best achievable* false negative error exponent, when P_{FP} tends to zero exponentially fast.

More formally, we start by proving the following property.

Proposition 4.1. For any two values λ_1 and λ_2 such that $\lambda_2 < \lambda_1$, $\Gamma(P_X, \lambda_2, L) \subseteq \Gamma(P_X, \lambda_1, L)$.

Proof. The proposition follows immediately from (4.1) by observing that $\Gamma(P_X, \lambda, L)$ depends on λ only through the acceptance region $\Lambda(P_X, \lambda)$, for which we obviously have $\Lambda^*(P_X, \lambda_2) \subseteq \Lambda^*(P_X, \lambda_1)$ whenever $\lambda_2 < \lambda_1$. □

Thanks to Proposition 4.1, we can compute the limit of the false negative error exponent when λ tends to zero, as summarized in the following theorem, extending the Chernoff-Stein's Lemma to the adversarial setup considered. Let us first give the following definition:

$$\Gamma(P_X, L) = \{P \in \mathcal{P}\colon EMD(P, P_X) \leq L\}. \tag{4.2}$$

Theorem 4.1. Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum average per-letter distortion L, the maximum achievable false negative error exponent for the DG-KS game is

$$\lim_{\lambda \to 0} \lim_{n \to \infty} -\frac{1}{n} \log P_{\text{FN}} = \min_{P \in \Gamma(P_X, L)} \mathcal{D}(P \| P_Y). \tag{4.3}$$

Proof. The innermost limit in the left-hand side of (4.3) defines the error exponent for a fixed λ, say it $\varepsilon(\lambda)$. From Theorem 3.4, we know that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_{\text{FN}} = \varepsilon(\lambda) = \min_{P \in \Gamma(P_X, \lambda, L)} \mathcal{D}(P \| P_Y). \tag{4.4}$$

Then, according to Proposition 4.1, the sequence $\varepsilon(\lambda)$ is monotonically non decreasing as λ decreases. In addition, since $\Gamma(P_X, L) \subseteq \Gamma(P_X, \lambda, L)$ $\forall \lambda$, for any $\lambda > 0$, we have:

$$\varepsilon(\lambda) \leq \min_{P \in \Gamma(P_X, L)} \mathcal{D}(P \| P_Y). \tag{4.5}$$

Being $\varepsilon(\lambda)$ bounded from above and non-decreasing, the limit for $\lambda \to 0$ exists and is finite. We must now prove that the limit is indeed equal to

$\min_{P\in\Gamma(P_X,L)} \mathcal{D}(P\|P_Y)$. Let P_0^* be the point achieving the minimum in (4.3) and P_λ^* the point achieving the minimum in the set $\Gamma(P_X,\lambda,L)$, i.e., the point achieving the minimum in Equation (3.23) (see Figure 3.1 for a pictorial representation of P_λ^*). Due to Lemma B.1 (Appendix B.1), for any arbitrarily small τ, we can choose a small enough λ such that, for any P in $\Gamma(P_X,\lambda,L)$, a pmf P' in $\Gamma(P_X,L)$ exists whose distance from P is lower than τ. By taking $P=P_\lambda^*$ and exploiting the continuity of the $\mathcal{D}$ function, we have

$$\mathcal{D}(P'\|P_Y) \leq \min_{P\in\Gamma(P_X,\lambda,L)} \mathcal{D}(P\|P_Y) + \delta(\tau), \tag{4.6}$$

for some $P'\in\Gamma(P_X,L)$ and some value $\delta(\tau)$ such that $\delta(\tau)\to 0$ as $\tau\to 0$. A fortiori, relation (4.6) holds for $P'=P_0^*$ and then we can write

$$\varepsilon(\lambda) \geq \min_{P\in\Gamma(P_X,L)} \mathcal{D}(P\|P_Y) - \delta(\tau), \tag{4.7}$$

where $\delta(\tau)$ can be made arbitrarily small by decreasing λ.

Equation (4.7), together with Equation (4.5), shows that we can get arbitrarily close to $\min_{P\in\Gamma(P_X,L)} \mathcal{D}(P\|P_Y)$, by making λ small enough, hence proving that the right-hand side of (4.3) is the limit of the sequence $\varepsilon(\lambda)$ as $\lambda\to 0$. □

Figure 4.1 gives a geometric interpretation of Theorem 4.1. The figure is obtained from Figure 3.1 in Chapter 3 by observing that when $\lambda\to 0$ the optimal acceptance region collapses into the single pmf P_X, i.e., $\Lambda^*=\{P_X\}$.

By the light of Theorem 4.1, $\Gamma(P_X,L)$ is the *smallest indistinguishability region* for the DG-KS game. Moreover, from Equation (4.2), we see that the distinguishability of two pmf's ultimately depends on their *EMD*. In fact, if $EMD(P_Y,P_X)>L$, $\mathscr{D}$ is able to distinguish X from Y by adopting a sufficiently small λ. On the contrary, if $EMD(P_Y,P_X)\leq L$, there is no positive value of λ for which the sequences emitted by the two sources can be asymptotically distinguished.

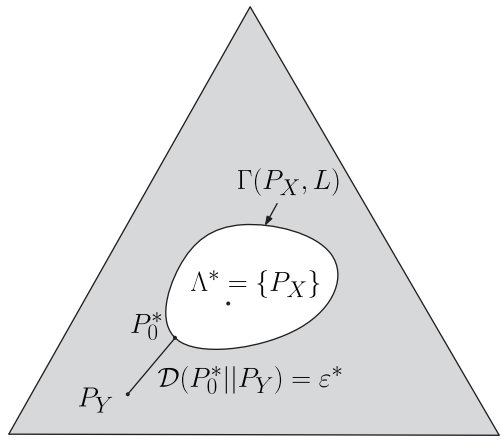


Figure 4.1: Geometric interpretation of set $\Gamma(P_X, L)$ and point P_0^* in the probability simplex by the light of Theorem 4.1.

4.3 Security Margin in the DG-KS Setup

By adopting a different perspective, given two sources X and Y, one may ask which is the maximum attack distortion for which $\mathscr{D}$ can distinguish X and Y. The answer to this question follows immediately from Theorem 4.1 and leads naturally to the following definition.

Definition 4.1 (Security Margin in the DG-KS Setup)**.** Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum average per-letter distortion for which the two sources can be reliably distinguished in the DG-KS setup is called Security Margin and is given by

$$\mathcal{SM}(P_Y, P_X) = EMD(P_Y, P_X). \tag{4.8}$$

Since the *EMD* is a symmetric function of P_X and P_Y [51], the Security Margin does not depend on the role of X and Y in the test, i.e., $\mathcal{SM}(P_X, P_Y) = \mathcal{SM}(P_Y, P_X)$.

The Security Margin is a powerful measure summarizing in a single quantity *how securely* two sources can be distinguished in an adversarial setup.

It is worth remarking that the Security Margin between two sources pertains to the *security* of the hypothesis test behind the binary detection problem and not to its *robustness*, since it is derived from the

performance at the equilibrium of the game, i.e., by assuming that both the players of the game make best choices in a strategic fashion. To exemplify the above concept, let us consider the simple case of two binary sources. Specifically, let X and Y be two Bernoulli sources with parameters $p = P_X(1)$ and $q = P_Y(1)$ respectively. Let also assume that the distortion constraint is expressed in terms of the Hamming distance between the sequences, that is $d(i, j) = 0$ when $i = j$ and 1 otherwise. Without loss of generality let $p > q$. The distortion associated to a transportation map S_{XY} can be written as:

$$\sum_{i,j} S_{YX}(i, j)d(i, j) = S_{YX}(0, 1) + S_{YX}(1, 0). \tag{4.9}$$

Since $p > q$, it is easy to conclude that the minimum of the above expression is obtained when $S_{YX}(1, 0) = 0$ (intuitively, if the source X outputs more 1's than Y, it does not make any sense to turn the 1's emitted by Y into 0's). As a consequence, to satisfy the constraint $S_X(1) = p$ we must let $S_{YX}(0, 1) = p-q$, yielding $\mathcal{SM}(P_Y, P_X) = p-q$, or more generally

$$\mathcal{SM}(P_Y, P_X) = |p - q|. \tag{4.10}$$

We can conclude that if the Attacker is allowed to introduce an average Hamming distortion larger or equal than $|p - q|$, then there is no way for the Defender to distinguish between the two sources. This is not the case if the output of the source Y passes through a binary symmetric channel with crossover probability equal to $|p - q|$, since the output of the channel would still be distinguishable from the sequences emitted by X. Consider, for example, a simple case where $q = 1/2$ and $p > 1/2$. Regardless of the crossover probability, the output of the channel will always be a binary source with equiprobable symbols, which is distinguishable from X given that $p > 1/2$. In other words, in the setup defined by the DG-KS game, the two Bernoulli sources cannot be distinguished securely in the presence of an attacker introducing a distortion equal to $|p - q|$, while they can be distinguished if the output of the source Y passes through a noisy channel introducing the same average distortion introduced by the Attacker.

4.4 Security Margin Computation

Given two discrete sources $X \sim P_X$ and $Y \sim P_Y$, the computation of the Security Margin requires the evaluation of $EMD(P_X, P_Y)$. In general, the *EMD* between two sources can be computed by resorting to numerical analysis, and, due to its wide use as a similarity measure in computer vision applications, several efficient algorithms have been proposed for that (see for example [78]). We know from Section 2.3.2 that, when the distortion (cost) function has the general form $d(i,j) = |i-j|^p$, with $p \geq 1$, we can resort to a fast iterative algorithm for the computation of the *EMD*, i.e., the Hoffman algorithm, known as *NWC* rule. A case of great interest is $p = 1$ and $p = 2$, according to which the distortion between y^n and the attacked sequence z^n corresponds, respectively, to the L_1 and L_2^2 distortion.

In some simple, yet insightful, cases, a closed form solution can be found, as detailed below.

4.4.1 Uniform Sources with Different Cardinalities

Let $X \sim P_X$ and $Y \sim P_Y$ be two uniform pmf's with alphabet sets $\mathcal{X}$ and $\mathcal{Y}$ such that $|\mathcal{X}| = \alpha|\mathcal{Y}|$, with $\alpha \in \mathbb{N}$. In this case, thanks to Hoffman's algorithm, for any L_p^p distortion, the *EMD*, and then the Security Margin, can be expressed as:

$$\mathcal{SM}_{L_p^p}(P_X, P_Y) = \frac{1}{|\mathcal{Y}|} \sum_{i=0}^{|\mathcal{X}|-1} \sum_{j=0}^{\alpha-1} (|i^{low} - j^{low}| - j - (\alpha-1)i)^p. \quad (4.11)$$

The formula implicitly assumes that $j^{low} > i^{low}$, the extension to the case in which such a relationship does not hold being immediate.

4.4.2 Security Margin under the Hamming Distance

A case of interest in which the *EMD* assumes a particularly simple form is when the Hamming distance d_H is considered. Specifically, in this case, the Security Margin between two sources $X \sim P_X$ and $Y \sim P_Y$

with source alphabet $\mathcal{X}$ can be expressed as:[1]

$$\begin{aligned}\mathcal{SM}_{d_H}(P_X, P_Y) &= \min_{S_{XY}:\sum_x S_{XY}=P_Y,\sum_y S_{XY}=P_X} \sum_{i,j} S_{XY}(i,j) d_H(i,j) \\ &= \sum_i \left(\sum_{j,j\neq i} S'_{XY}(i,j) \right) \\ &= \sum_{i\in\mathcal{X}} [P_X(i) - P_Y(i)]_+, \end{aligned} \tag{4.12}$$

where, in the second equality, S'_{XY} is any pmf that satisfies $S_{XY}(i, j = i) = \min\{P_X(i), P_Y(j = i)\}$, $\forall i \in \mathcal{X}$. Then, by construction, for any $i \in \mathcal{X}$, $\sum_{j\neq i} S'_{XY}(i,j) = [P_X(i) - P_Y(j = i)]_+$.[2]

4.4.3 Security Margin Under the L_1 Distance

Let $X \sim P_X$ and $Y \sim P_Y$ be two sources with alphabet $\mathcal{X}$. If the distortion function corresponds to the L_1 distance, the *EMD*, and hence the Security Margin, assumes a particularly simple form. Specifically, the $\mathcal{SM}$ between P_X and P_Y can be expressed in closed form as follows:

$$\mathcal{SM}_{L_1}(P_X, P_Y) = \sum_{i\in\mathcal{X}} \left| \sum_{s=1}^{i} (P_X(s) - P_Y(s)) \right|. \tag{4.13}$$

The above expression can be derived by rephrasing the *EMD* computation as a minimum cost flow problem [50, Section 1.2] and applying the flow decomposition principle [79] to the solution of the TP problem provided by the Hoffman's algorithm, i.e., by the NWC rule. For sake of completeness, we provide the proof in Appendix C.

4.5 Source Distinguishability Under Maximum Distortion-Limited Attack

So far, we have considered the case of additive distortion measures. In this section, we extend the definition of the Security Margin to the

[1] For a given quantity s, $[s]_+ \triangleq \max\{s, 0\}$. Equivalently, $[s]_+ = s$ if $s \geq 0$ and zero otherwise.

[2] According to the transportation perspective, the optimum map is any map that, given the source pile P_X, leaves in place as much mass as possible and moves the remaining (surplus) mass to fill the sink pile P_Y in an arbitrary way.

case in which the distortion introduced by the Attacker is measured by relying on the L_∞ distance.

By following the same steps of Sections 4.2 and 4.3, we study the behavior of the indistinguishability region of the test when $\lambda \to 0$ to determine the smallest indistinguishability region. It is interesting to notice that, even if the adoption of the L_∞ distance prevents a direct formulation of the problem in terms of mass transport, the distinguishability between two sources X and Y is still closely related to the optimal transportation map between P_X and P_Y. The basis for such a connection is rooted in the following lemma.

Lemma 4.2. Given two distributions P and Q, the transportation map S_{PQ}^{NWC} obtained by applying the NWC rule to P and Q is a solution of the problem

$$\min_{S_{PQ}:\, S_P=P, S_Q=Q} \Big(\max_{(i,j):\, S_{PQ}(i,j)\neq 0} |i-j| \Big). \tag{4.14}$$

Proof. Let $S^* \neq S_{PQ}^{NWC}$ be a generic transformation that maps P into Q. Given that $S^* \neq S_{PQ}^{NWC}$ there exists at least one quadruple of bins (t, r, v, s), with $t < r$ and $v < s$, for which, $S^*(t, s) > 0$ and $S^*(r, v) > 0$. Let us assume, without loss of generality, that $S^*(t, s) \leq S^*(r, v)$. We now define a new map S' which is obtained from S^* by letting:

$$\begin{aligned} S'(t, v) &= S^*(t, v) + S^*(t, s) \\ S'(t, s) &= 0 \\ S'(r, v) &= S^*(r, v) - S^*(t, s) \\ S'(r, s) &= S^*(r, s) + S^*(t, s). \end{aligned} \tag{4.15}$$

Since $\max\{|t-s|, |r-v|\} > \max\{|t-v|, |r-s|\}$, the maximum distortion introduced by S' is lower than or equal to that introduced by S^*, that is:

$$\max_{(i,j):\, S^*(i,j)\neq 0} |i-j| \geq \max_{(i,j):\, S'(i,j)\neq 0} |i-j|. \tag{4.16}$$

We now inspect S', if there is another quadruple of bins (t', r', v', s') satisfying the same properties of (t, r, v, s), we let $S^* = S'$ and iterate the above procedure. The process ends when no quadruple of bins with the required properties exists and hence when $S' = S_{PQ}^{NWC}$. Since at

each step the distortion introduced by the new map does not increase, the above procedure proves that S_{PQ}^{NWC} introduces a distortion lower than or equal to that introduced by any other S^* mapping P into Q, thus proving that S_{PQ}^{NWC} achieves the minimum in (4.14). □

Thanks to Lemma 4.2, the set $\Gamma_{L_\infty}(P_X, \lambda, L)$ in (3.44) can be rewritten as follows:

$$\Gamma_{L_\infty}(P_X, \lambda, L) = \left\{ P \in \mathcal{P}\colon \exists\, Q \in \Lambda^*(P_X, \lambda) \text{ s.t} \max_{(i,j)\colon S_{PQ}^{NWC}(i,j)\neq 0} |i-j| \leq L \right\}. \tag{4.17}$$

By letting λ tend to 0, we obtain the smallest indistinguishability region, thus extending Theorem 4.1 to the DG-KS game with L_∞ distortion. Let us define[3]

$$\Gamma_{L_\infty}(P_X, L) = \left\{ P \in \mathcal{P}\colon \max_{(i,j)\colon S_{P_X P}^{NWC}\neq 0} |i-j| \leq L \right\}. \tag{4.18}$$

We can prove the following theorem.

Theorem 4.3. Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum allowable per-letter distortion L, the maximum achievable false negative error exponent for the DG-KS game with L_∞ distortion is

$$\lim_{\lambda\to 0} \lim_{n\to\infty} -\frac{1}{n} \log P_{\text{FN}} = \min_{P\in\Gamma_{L_\infty}(P_X,L)} \mathcal{D}(P\|P_Y). \tag{4.19}$$

Proof. The proof relies on the extension of Proposition 4.1 and Lemma B.1 to the L_∞ case. The extension of Proposition 4.1 is immediate since, even in this case the indistinguishability region depends on λ only through $\Lambda^*(P_X, \lambda)$, whose form does not depend on the particular norm adopted to express the distortion constraint. The extension of Lemma B.1 requires some more care and is proven in Appendix B.2 (Lemma B.2). For the rest, the theorem can be proven by reasoning as in the proof of Theorem 4.1. □

[3] We exploit the fact that, by symmetry, $\max_{(i,j)\colon S_{PP_X}^{NWC}\neq 0} |i-j| = \max_{(i,j)\colon S_{P_X P}^{NWC}\neq 0} |i-j|$.

As a consequence of Theorem 4.3, the distinguishability of two sources depends again on the optimum transportation map between the pmf's of the two sources. Specifically, given the sources X and Y, the Defender is able to distinguish between them if and only if

$$\max_{(i,j)\in S_{XY}^{NWC}(i,j)\neq 0} |i-j| > L. \tag{4.20}$$

Condition (4.20) can be used to determine the maximum attack distortion for which $\mathscr{D}$ is able to distinguish the two sources X and Y, i.e., the Security Margin.

Definition 4.2 (Security Margin in the DG-KS Setup with L_∞ Distortion)**.** Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum distortion for which the two sources can be reliably distinguished in the DG-KS setting with L_∞ distortion is given by

$$\mathcal{SM}_{L_\infty}(P_X, P_Y) = \max_{(i,j):\, S_{XY}^{NWC}(i,j)\neq 0} |i-j|, \tag{4.21}$$

where S_{XY}^{NWC} is obtained by applying the NWC rule to map P_X into P_Y.

5

Binary Detection Game with Training Data

In this chapter, we consider a more close-to-reality scenario and study the case in which the sources are not fully known to Defender and Attacker.

The analysis is motivated by the fact that the assumption of full knowledge of the sources, made in Chapter 3,[1] is rarely met in real applications, where the statistical model of the sources is often not available to the Defender. In this scenario, it is likely that the Defender can build a suitable model to characterize H_0 by relying on a number of samples drawn from the sources.

For the above reasons, in this section we remove the assumption that P_X and P_Y are known and study the detection game when training data is available to the two players. Specifically, we first formally define the binary detection game with training data, then we solve the game by determining the equilibrium point in the case in which equal training sequences are available to the players. The payoff at the equilibrium of the game is computed and the performance are compared with those achieved by the game with known sources. The case of different training

[1] We remind that, in the asymptotic case, only the knowledge of P_X is required.

sequences available to the players is also addressed at the end of the chapter.

5.1 Detection Game with Training Data (DG-TR)

5.1.1 Problem Definition

By sticking to the notation introduced in Section 2.1, let $\mathcal{C}$ be the class of discrete memoryless sources with alphabet $\mathcal{X}$, and let $X \sim P_X$ be a source in $\mathcal{C}$ characterizing H_0. As for the DG-KS game, the purpose of the Defender is to decide whether a test sequence z^n was drawn from X or not. To make his decision, $\mathscr{D}$ relies on the knowledge of a training sequence of a given length N, namely t_D^N, drawn from X. On his side, the Attacker takes a sequence y^n emitted by another source $Y \sim P_Y$ still belonging to $\mathcal{C}$ and tries to modify it in such a way that $\mathscr{D}$ thinks that the modified sequence was generated by the same source that generated t_D^N. As usual, $\mathscr{A}$ must satisfy a distortion constraint stating that the distance between the modified sequence and y^n must be lower than a threshold. Like the Defender, the Attacker derives his knowledge about the statistics of the sequences generated under H_0 through a training sequence t_A^K drawn from P_X, that in general may not coincide with t_D^N. We assume that t_D^N, t_A^K, and y^n, as well as the observed sequence under H_0, i.e., x^n, are generated independently. With regard to P_Y, we could also assume that it is known through two training sequences, one available to $\mathscr{A}$ and one to $\mathscr{D}$, however we will see that – as for the case of known sources, and, at least asymptotically – such an assumption is not necessary, and hence we make the simplifying assumption that P_Y is known to neither $\mathscr{D}$ nor $\mathscr{A}$.

In the above framework, H_0 is equivalent to the hypothesis that the test sequence has been generated by the same source that generated t_D^N. We denote with Λ_{tr}^n the acceptance region for H_0.[2] Throughout this chapter, we find convenient to think of Λ_{tr}^n as a subset of $\mathcal{X}^n \times \mathcal{X}^N$,

[2]For the sake of clarity, we add the subscript "ks" to denote the quantities Λ, Γ and ε for the game with known sources and the subscript "tr" for the game with training data.

i.e., as the set of all the pairs of sequences (z^n, t_D^N) that the Defender considers to be drawn from the same, unknown, source.

5.1.2 DG-TR with Independent Training Data (DG-TRa)

In order to confine the analysis to a case in which the analysis of the game is tractable, we follow the same approach adopted for the known sources case and consider a version of the game in which $\mathscr{D}$ bases his decision on a limited set of statistics computed on the test and training sequences: specifically, we require that $\mathscr{D}$ relies only on the relative frequencies with which the symbols in $\mathcal{X}$ appear in z^n and t^N, i.e., P_{z^n} and P_{t^N}. Note that, as in the KS case, P_{z^n} and P_{t^N} are not sufficient statistics for $\mathscr{D}$, since even if Y is a memoryless source, the Attacker could introduce some memory within the sequence as a result of the attack. In the same way, he could introduce some dependencies between the attacked sequence z^n and t^N. It is then necessary to treat the assumption that $\mathscr{D}$ relies only on P_{z^n} and P_{t^N} as an explicit requirement.

As a consequence of the limited resources assumption, the acceptance region Λ_{tr}^n can only be a union of Cartesian products of pairs of type classes, i.e., if the pair of sequences (z^n, t^N) belongs to Λ_{tr}^n, then any pair of sequences belonging to the Cartesian product $\mathcal{T}(P_{z^n}) \times \mathcal{T}(P_{t^N})$ will also be contained in Λ_{tr}^n.[3] Since a type class is univocally defined by the empirical pmf of the sequences contained in it, we can redefine Λ_{tr}^n as a union of pairs of types (P, Q) with $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}_N$. In the following, we will use the two interpretations of Λ_{tr}^n (as a set of pairs of sequences or pairs of types) interchangeably, the exact meaning being always recoverable from the context.

As in the previous case, we are interested in studying the asymptotic behavior of the game when n, N, and K tend to infinity. Rather than considering the limits with n, N, and K tending to infinity independently, we will express N and K as a function of n, and study what happens when n tends to infinity. In this way, the exponents of the Type I and II error probability are still defined as in (2.5) (Section 1.2).

[3] Strictly speaking, Λ_{tr}^n should depend on both n and N: however, in the following we will express N as a function of n, thus making the dependence on N implicit.

With the above ideas in mind, we are now ready to define a first version of the binary decision game with training sequences. We will do so by directly rewriting the set of strategies for the Attacker in terms of transportation maps. As done in Chapter 3, in fact, we assume that the distance measure $d(\cdot,\cdot)$ defining the distortion introduced by the Attacker is additive, and hence we can adopt the transportation theoretic formalism introduced in Section 3.2.1.

Definition 5.1. The DG-TRa $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game is a zero-sum, strategic, game played by $\mathscr{D}$ and $\mathscr{A}$, defined by the following strategies and payoff.

- *Defender's strategies.* The set of strategies $\mathscr{D}$ can choose from is the set of acceptance regions for which the maximum false positive probability across all possible $P_X \in \mathcal{P}$ is lower than a given threshold:[4]

$$\mathcal{S}_{\mathscr{D}} = \left\{ \Lambda_{tr}^n \subset \mathcal{P}_n \times \mathcal{P}_N \colon \max_{P_X \in \mathcal{P}} P_X((z^n, t_D^N) \notin \Lambda_{tr}^n) \leq 2^{-\lambda n} \right\}, \quad (5.1)$$

  where the quantity $P_X((z^n, t_D^N) \notin \Lambda_{tr}^n)$ is the false positive error probability, that is the probability that two independent sequences generated by X do not belong to Λ_{tr}^n.

- *Attacker's strategies.* The set of strategies $\mathscr{A}$ can choose from consists of all the possible ways of choosing an admissible transportation map to transform y^n into z^n:

$$\mathcal{S}_{\mathscr{A}} = \{ S_{YZ}^n(y^n, t_A^K) \colon S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n}) \}, \quad (5.2)$$

  where $\mathcal{A}^n(L, P_{y^n})$ is the set of admissible maps given the maximum allowed per-letter distortion L, and where we have explicitly indicated that the choice of the map depends on t_A^K, since, when performing the attack, $\mathscr{A}$ can exploit the knowledge of his training sequence.

- *The payoff function.* Adopting again the Neyman–Pearson approach, the payoff is defined in terms of the false negative error

[4] To simplify the notation, when it is not strictly necessary, we will omit to indicate explicitly the dependence of N, res. K, on n.

probability, that is:

$$u(\Lambda_{tr}^n, S_{YZ}^n) = - \sum_{\substack{t_D^N \in \mathcal{X}^N,\ t_A^K \in \mathcal{X}^K \\ y^n:\ (S_Z^n(y^n, t_A^K), t_D^N) \in \Lambda_{tr}^n}} P_Y(y^n) P_X(t_D^N) P_X(t_A^K), \tag{5.3}$$

where the error probability is averaged across all possible y^n and training sequences and where we have exploited the independence of y^n, t_D^N and t_A^K. Once again, we adopted $\mathscr{D}$'s perspective in the definition of the payoff.

We stress that the (apparently weird) dependence of the false positive and false negative probabilities on the training sequence t_D^N is because $\mathscr{D}$ bases the decision on both the observation and the training sequences, the decision made by $\mathscr{D}$ being on whether test and training sequences are generated by the same source or not.

Before going on with the analysis, we pause to discuss some of the choices we implicitly made in the above definition. A first observation regards the payoff function. As a matter of fact, the expression in (5.3) looks problematic, since its evaluation requires that the pmf's P_X and P_Y are known, however this is not the case in our scenario since we have assumed that P_X is known only through t_D^N and t_A^K, and that P_Y is not known at all. As a consequence it may seem that the players of the game are not able to compute the payoff associated to a given profile and hence they have no arguments upon which they can base their choice. While this is indeed a problem in a generic setup, we will show later on that asymptotically (when n, N and K tend to infinity) the optimal strategies of $\mathscr{D}$ and $\mathscr{A}$ are uniformly optimum across all P_X and P_Y and hence the ignorance of P_X and P_Y is not a problem. One may wonder why we did not define the payoff under a worst case assumption (from $\mathscr{D}$'s perspective) on P_X and/or P_Y. The reason is that doing so would result in a meaningless game since the worst case for $\mathscr{D}$ would always correspond to $P_Y = P_X$ for which no decision is possible.

As a second remark, we stress that, as in the DG-KS case, limiting the strategies of the Attacker to deterministic mapping is not a restrictive

choice since, at least asymptotically, the optimal strategy of $\mathscr{D}$ depends neither on the strategy chosen by $\mathscr{A}$ (hence on t_A^K) nor on P_Y.

5.1.3 DG-TR with Identical Training Data (DG-TRb)

An interesting variant of the DG-TRa game is obtained by assuming that the training sequence available to $\mathscr{A}$ is equal to that available to $\mathscr{D}$, leading to the following definition.

Definition 5.2. The DG-TRb $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game is a zero-sum, strategic, game defined as the DG-TRa with the only difference that $K = N$ and $t_A^K = t_D^N$ (simply referred to as t^N in the following)

$$\mathcal{S}_{\mathscr{D}} = \left\{\Lambda_{tr}^n \subset \mathcal{P}_n \times \mathcal{P}_N \colon \max_{P_X \in \mathcal{P}} P_X\{(z^n, t^N) \notin \Lambda_{tr}^n\} \leq 2^{-\lambda n}\right\}, \tag{5.4}$$

$$\mathcal{S}_{\mathscr{A}} = \{S_{YZ}^n(y^n, t^N) \colon S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})\}, \tag{5.5}$$

$$u(\Lambda_{tr}^n, S_{YZ}^n) = - \sum_{\substack{(y^n, t^N) \in \mathcal{X}^n \times \mathcal{X}^N \colon \\ (S_Z^n(y^n, t^N), t^N) \in \Lambda_{tr}^n}} P_Y(y^n) P_X(t^N). \tag{5.6}$$

The set of strategies of $\mathscr{D}$ and $\mathscr{A}$ are the same as in the DG-TRa game, and only the payoff is defined differently.

Due to its simplicity, in the rest of the section we will first focus on version b of the game, and then extend our results so to cover version a as well.

5.2 Asymptotic Equilibrium of the DG-TRb Game

We start the analysis by determining the optimal acceptance region for $\mathscr{D}$. The derivation of the optimal strategy for $\mathscr{D}$ passes through the definition of the generalized log-likelihood ratio function $h(z^n, t^N)$ ([80, Chapter 24], [81, p. 403]).

Given the test and training sequences z^n and t^N, that may or may not come from the same source, the generalized log-likelihood ratio function is defined as:[5]

$$h(z^n, t^N) = \mathcal{D}(P_{z^n} \| P_{r^{n+N}}) + \frac{N}{n} \mathcal{D}(P_{t^N} \| P_{r^{n+N}}), \tag{5.7}$$

[5] We observe that the h function resembles the Jensen–Shannon divergence (JSD) [82], where the two divergence terms in (5.7) are taken with equal weights.

where $P_{r^{n+N}}$ denotes the empirical pmf of the sequence r^{n+N}, obtained by concatenating z^n and t^N, i.e.,

$$r_i = \begin{cases} z_i & i \leq n \\ t_{i-n} & n < i \leq n+N. \end{cases} \tag{5.8}$$

By observing that $h(z^n, t^N)$ depends on the test and the training sequences only through their empirical pmf, we can also use the notation $h(P_{z^n}, P_{t^N})$. The study of the equilibrium for the DG-TRb passes through the following lemma.

Lemma 5.1. For any P_X we have:

$$n\mathcal{D}(P_{z^n}\|P_{r^{n+N}}) + N\mathcal{D}(P_{t^N}\|P_{r^{n+N}}) \leq n\mathcal{D}(P_{z^n}\|P_X) + N\mathcal{D}(P_{t^N}\|P_X), \tag{5.9}$$

where equality holds if and only if $P_X = P_{r^{n+N}}$.

Proof. We rewrite (5.9) by moving all the non-zero terms to the left-hand side:

$$\begin{aligned} &n\mathcal{D}(P_{z^n}\|P_{r^{n+N}}) + N\mathcal{D}(P_{t^N}\|P_{r^{n+N}}) \\ &\quad - n\mathcal{D}(P_{z^n}\|P_X) - N\mathcal{D}(P_{t^N}\|P_X) \leq 0. \end{aligned} \tag{5.10}$$

By using the definition of the empirical KL divergence and grouping the first term with the third and the second with the fourth, the left hand side of (5.10) is equivalent to

$$n\sum_{a\in\mathcal{X}} P_{z^n}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)} + N\sum_{a\in\mathcal{X}} P_{t^N}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)}. \tag{5.11}$$

Being r^{n+N} the concatenation of z^n and t^N, we argue that $nP_{z^n}(a) + NP_{t^N}(a) = (n+N)P_{r^{n+N}}(a)\ \forall a \in \mathcal{X}$, which permits to rewrite the sum in (5.11) as follows:

$$(n+N)\sum_{a\in\mathcal{X}} P_{r^{n+N}}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)} = -(n+N)\mathcal{D}(P_{r^{n+N}}\|P_X). \tag{5.12}$$

Hence, the proof of relation (5.9) follows from the positivity of the divergence function, which equals zero if and only if $P_X = P_{r^{n+N}}$.

In hindsight, relation (5.9) derives from the property that the empirical probability distribution $P_{r^{n+N}}$ maximizes the probability that

a source outputs the concatenation of x^n and t^N, i.e., $P_X(r^{n+N}) \leq P_{r^{n+N}}(r^{n+N})\ \forall P_X$. To show this, from (5.12) we write:

$$\sum_{a\in\mathcal{X}} N_{r^{n+N}}(a) \log \frac{P_X(a)}{P_{r^{n+N}}(a)} \leq 0. \tag{5.13}$$

By exploiting the properties of the logarithm, Equation (5.13) can be rewritten as follows:

$$\log \prod_{a\in\mathcal{X}} P_X(a)^{N_{r^{n+N}}(a)} \leq \log \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)}, \tag{5.14}$$

which implies

$$P_X(r^{n+N}) \leq \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)} = P_{r^{n+N}}(r^{n+N}). \tag{5.15}$$

□

Given the above, we are now ready to prove the following result.

Lemma 5.2. Let $\Lambda_{tr}^{n,*}$ be defined as follows:

$$\Lambda_{tr}^{n,*} = \left\{(P_{z^n}, P_{t^N})\colon h(P_{z^n}, P_{t^N}) < \lambda - |\mathcal{X}|\frac{\log(n+1)(N+1)}{n}\right\}, \tag{5.16}$$

with

$$\lim_{n\to\infty} \frac{\log(N(n)+1)}{n} = 0. \tag{5.17}$$

Then:

1. $\max_{P_X} P_X\{(z^n, t^N) \notin \Lambda_{tr}^{n,*}\} \leq 2^{-n(\lambda-\nu_n)}$, with $\lim_{n\to\infty} \nu_n = 0$,
2. $\forall \Lambda_{tr}^n \in \mathcal{S}_{\mathscr{D}}$, we have $\bar{\Lambda}_{tr}^n \subseteq \bar{\Lambda}_{tr}^{n,*}$.

Proof. Being $\Lambda_{tr}^{n,*}$ a union of pairs of types (or, equivalently, a union of Cartesian products of type classes), we have:

$$\begin{aligned}\max_{P_X} P_{\mathrm{FP}} &= \max_{P_X} \sum_{(z^n,t^N)\in\bar{\Lambda}_{tr}^{n,*}} P_X(z^n, t^N) \\ &= \max_{P_X} \sum_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} P_X(\mathcal{T}(P_{z^n}) \times \mathcal{T}(P_{t^N})).\end{aligned} \tag{5.18}$$

For the class of discrete memoryless sources, the number of types with denominators n and N is bounded by $(n+1)^{|\mathcal{X}|}$ and $(N+1)^{|\mathcal{X}|}$ respectively (see Section 2.4.1), so we can write:

$$\begin{aligned}\max_{P_X} P_{\text{FP}} &\leq \max_{P_X} \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{n,*}} \\ &\quad [(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}P_X(\mathcal{T}(P_{z^n})\times\mathcal{T}(P_{t^N}))] \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\cdot\max_{P_X} \\ &\quad \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} 2^{-n[\mathcal{D}(P_{z^n}\|P_X)+\frac{N}{n}\mathcal{D}(P_{t^N}\|P_X)]}, \end{aligned} \tag{5.19}$$

where in the second inequality we have exploited the independence of z^n and t^N and the property of types according to which for any sequence z^n we have $P_X(\mathcal{T}(P_{z^n})) \leq 2^{-n\mathcal{D}(P_{z^n}\|P_X)}$ (see Section 2.4.1). By exploiting Lemma 5.1, we can write:

$$\begin{aligned}\max_{P_X} P_{\text{FP}} &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\cdot \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} \\ &\quad 2^{-n[\mathcal{D}(P_{z^n}\|P_{r^{n+N}})+\frac{N}{n}\mathcal{D}(P_{t^N}\|P_{r^{n+N}})]} \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\ 2^{-n\left(\lambda-|\mathcal{X}|\frac{\log(n+1)(N+1)}{n}\right)} \\ &= 2^{-n\left(\lambda-2|\mathcal{X}|\frac{\log(n+1)(N+1)}{n}\right)}, \end{aligned} \tag{5.20}$$

where the last inequality derives from the definition of $\Lambda_{tr}^{n,*}$. Together with (5.17), Equation (5.20) proves the first part of the lemma with $\nu_n = 2|\mathcal{X}|\frac{\log(n+1)(N+1)}{n}$.[6]

For any $\Lambda_{tr}^{n} \in \mathcal{S}_{\mathscr{D}}$, let (z^n, t^N) be a generic pair of sequences contained in $\bar{\Lambda}_{tr}^{n}$. Due to the limited resources assumption the Cartesian product between $\mathcal{T}(P_{z^n})$ and $\mathcal{T}(P_{t^N})$ will be entirely contained in $\bar{\Lambda}_{tr}^{n}$. Then we have:

$$\begin{aligned}2^{-\lambda n} &\geq \max_{P_X} P_X(\bar{\Lambda}) \\ &\overset{(a)}{\geq} \max_{P_X} P_X(\mathcal{T}(P_{z^n})\times\mathcal{T}(P_{t^N}))\end{aligned}$$

[6] We notice that $\nu_n \to 0$ as $n \to \infty$ thanks to the condition in (5.17).

$$\begin{aligned} &\overset{(b)}{\geq} \max_{P_X} \frac{2^{-n[\mathcal{D}(P_{z^n}\|P_X)+\frac{N}{n}\mathcal{D}(P_{t^N}\|P_X)]}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}} \\ &\overset{(c)}{=} \frac{2^{-n[\mathcal{D}(P_{z^n}\|P_{r^{n+N}})+\frac{N}{n}\mathcal{D}(P_{t^N}\|P_{r^{n+N}})]}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}}, \end{aligned} \tag{5.21}$$

where (a) is due to the limited resources assumption, (b) follows from the independence of z^n and t^N and $\underline{a}$ the lower bound on the probability of a pair of type classes (see Section 2.4.1), and (c) derives from Lemma 5.1. By taking the logarithm of both sides we find that $(z^n, t^N) \in \bar{\Lambda}_{tr}^{n,*}$, thus completing the proof. □

The first part of Lemma 5.2 shows that, at least asymptotically, $\Lambda_{tr}^{n,*}$ is an admissible strategy for the Defender; in fact, the constraint in (5.4) is fulfilled asymptotically and then $\Lambda_{tr}^{n,*}$ belongs to $\mathcal{S}_{\mathscr{D}}$ for sufficiently large n. Then, the optimality of $\Lambda_{tr}^{n,*}$ follows from the second part of the lemma.

An important observation is that the optimal strategy of $\mathscr{D}$ is univocally determined by the false positive constraint. This solves the apparent problem that we pointed out when defining the payoff of the game, namely that the payoff depends on P_X and P_Y and hence it is not fully known to $\mathscr{D}$. According to the lemma, the optimal strategy of $\mathscr{D}$ does not depend on the strategy chosen by $\mathscr{A}$ (then, neither on the training sequence available to him), that is $\Lambda_{tr}^{n,*}$ is a *strictly dominant strategy* for $\mathscr{D}$. As a consequence, $\Lambda_{tr}^{n,*}$ is the optimal Defender's strategy even for version a of the DG-TR game.

As it happened for the DG-KS game, due to the existence of a dominant strategy for the Defender, the derivation of the optimal attack strategy is an easy task. We only need to observe that the goal of $\mathscr{A}$ is to take a sequence y^n drawn from Y and modify it by applying an admissible transportation map, trying to reach the condition

$$h(S_Z^n(y^n, t^N), P_{t^N}) < \lambda - |\mathcal{X}|\frac{\log(n+1)(N+1)}{n}. \tag{5.22}$$

The optimal attack strategy, then, can be expressed as a minimization problem, i.e.,

$$S_{YZ}^{n,*}(y^n, t^N) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} h(S_Z^n, P_{t^N}). \tag{5.23}$$

Note that to implement this strategy $\mathscr{A}$ needs to know t^N, i.e., (5.23) determines the optimal strategy only for version b of the game. Since $S_{YZ}^{n,*}(y^n, t^N)$ (res. $S_Z^n(y^n, t^N)$) depends on the sequences y^n and t^N only through their empirical pmf, we can also use the notation $S_{YZ}^{n,*}(P_{y^n}, P_{t^N})$ (res. $S_Z^n(P_{y^n}, P_{t^N})$).

Finally, we observe that the optimization problem the Attacker must solve is the same as for the KS case with the only difference that the objective function is the h function instead of $\mathcal{D}$, the convexity of the h function in the $n(i, j)$ variables following by the same arguments.

Having determined the optimal strategies for $\mathscr{D}$ and $\mathscr{A}$, we can state the first main result of this chapter.

Theorem 5.3 (Equilibrium Point of DG-TRb Game)**.** The DG-TRb game is a dominance solvable game and the profile $(\Lambda_{tr}^{n,*}, S_{YZ}^{n,*}(y^n, t^N))$ is the only rationalizable equilibrium.

5.2.1 Comparison between the KS and TR Setups

To get a better insight into the meaning of the equilibrium point of the DG-TRb game, it is instructive to compare it with the equilibrium of the corresponding game with known sources, namely the DG-KS.

To start with, we observe that the use of the h function instead of the divergence $\mathcal{D}$ derives from the fact that, for the DG-TR case, the Defender must ensure that the false positive probability stays below the desired threshold for all possible discrete memoryless sources. To do so, he has to estimate the pmf that *better explains the evidence* provided by z^n and t^N, that is the pmf maximizing the probability of observing z^n and t^N. We know (see relation (5.15)) that such a maximizing pmf corresponds to the empirical pmf of the concatenation of z^n and t^N, i.e., $P_{r^{n+N}}(r^{n+N})$, and the generalized log-likelihood function corresponds to 1 over n the log of the (asymptotic) probability that a source with pmf equal to $P_{r^{n+N}}$ outputs the sequences z^n and t^N. A geometric illustration of the difference between the $\mathcal{D}$ and the h functions is given in Figure 5.1. Another observation regards the optimal strategy of the Attacker. As a matter of fact, the functions $h(P_{z^n}, P_{t^N})$ and $\mathcal{D}(P_{z^n} \| P_{t^N})$ share a similar behavior: they are both positive and convex functions

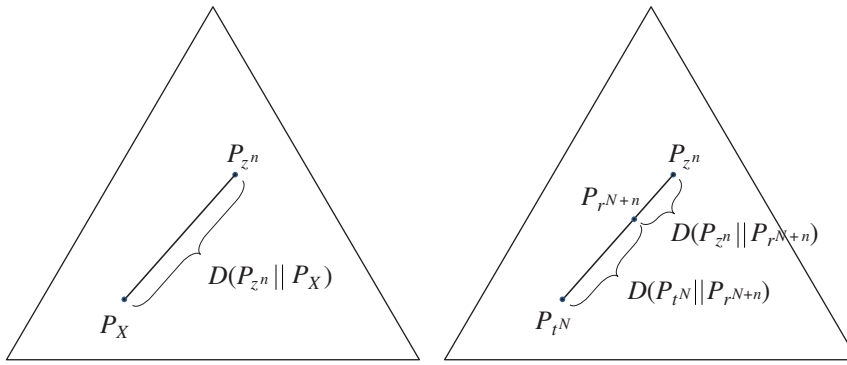


Figure 5.1: Geometric interpretation of the difference between $\mathcal{D}$ (left) and h (right) functions. The position of $P_{r^{N+n}}$ in the segment joining P_{z^n} to P_{t^N} depends on the ratio between N and n.

achieving the absolute minimum when $P_{z^n} = P_{t^N}$,[7] so one may be tempted to think that from the Attacker's point of view minimizing $\mathcal{D}(P_{z^n} \| P_{t^N})$ is equivalent to minimizing $h(P_{z^n}, P_{t^N})$. While this is the case in some situations, essentially, when the absolute minimum can be reached, in general the two minimization problems yield different solutions.

To further compare the DG-TRb and the DG-KS games, it is useful to rewrite the generalized log-likelihood function in a more convenient way. By applying some algebra, it is easy to prove the following equivalent expression for h:

$$h(P_{z^n}, P_{t^N}) = \mathcal{D}(P_{z^n} \| P_{t^N}) - \frac{N+n}{n} \mathcal{D}(P_{r^{n+N}} \| P_{t^N}), \tag{5.24}$$

showing that $h(P_{z^n}, P_{t^N}) \leq \mathcal{D}(P_{z^n} \| P_{t^N})$, with the equality holding only in the trivial case $P_{z^n} = P_{t^N}$. This suggests that, at least for large n, it should be easier for the Attacker to bring a sequence generated by Y within $\Lambda_{tr}^{n,*}$ than to bring it within $\Lambda_{ks}^{n,*}$. This is indeed the case, as it will be shown in Section 5.3.1, where we will provide a rigorous proof that the DG-TRb game is actually more favorable to the Attacker than the DG-KS game.

[7] Since h is the difference of two divergence functions with the same absolute minimum, the convexity of h directly follows from the convexity of $\mathcal{D}$.

We conclude this section by investigating the behavior of the optimal acceptance strategy for different values of the ratio $\frac{N}{n}$. To do so, we introduce the two quantities $c_z = \frac{n}{n+N}$ and $c_t = \frac{N}{n+N}$, representing the weights of the sequences z^n and t^N in r^{n+N}. It is easy to show, in fact, that

$$P_{r^{n+N}} = c_z P_{z^n} + c_t P_{t^N}. \tag{5.25}$$

In the simplest case, n and N tends to infinity with the same speed, hence we can assume that the ratio between N and n is fixed, namely, $\frac{N}{n} = c \neq 0$ (we obviously have $c_z = \frac{1}{1+c}$ and $c_t = \frac{c}{1+c}$). Under this assumption, the decision of the Defender is dictated by (5.16) and no particular behavior can be noticed. This is not the case when N/n tends to zero or ∞.

If $N/n \to 0$, then $P_{r^{n+N}} \to P_{z^n}$ and $h(P_{z^n}, P_{t^N}) \to 0$. This means that the Defender will always decide in favor of H_0. This makes sense since when the test sequence is infinitely longer than the training sequence, the evidence provided by the training sequence is not strong enough to let the Defender reject hypothesis 0.

If $N/n \to \infty$, the analysis is slightly more involved. In this case $c_t \to 1$ and $P_{r^{n+N}} \to P_{t^N}$, hence the first term in (5.7) tends to $\mathcal{D}(P_{z^n} \| P_{t^N})$. To understand the behavior of the second term of (5.7) when $n \to \infty$, we can use the Taylor expansion of $\mathcal{D}(P\|Q)$ when P approaches Q (see [83, Chapter 4]), which applied to the second term of the h function yields:

$$\begin{aligned}
\frac{N}{n} \cdot \mathcal{D}(P_{t^N} \| P_{r^{n+N}}) &\approx \frac{N}{2n} \cdot \sum_x \frac{(P_{t^N}(x) - P_{r^{n+N}}(x))^2}{P_{r^{n+N}}(x)} \\
&= \frac{N}{2n} \cdot \sum_x \frac{(c_x P_{t^N}(x) - c_x P_{z^n}(x))^2}{P_{r^{n+N}}(x)} \\
&= \frac{\frac{n}{N}}{2(\frac{n}{N}+1)^2} \sum_x \frac{(P_{t^N}(x) - P_{z^n}(x))^2}{P_{r^{n+N}}(x)}.
\end{aligned} \tag{5.26}$$

When $N/n \to \infty$, the above expression clearly tends to zero, and hence $h(P_{z^n}, P_{t^N}) \to \mathcal{D}(P_{z^n} \| P_{t^N})$. In other words, the optimal acceptance region tends to be equal to the one obtained for the case of know sources with P_X replaced by P_{t^N}. This is also an intuitively reasonable result: when the training sequence is much longer than the test sequence, the

empirical pmf of the training sequence provides such a reliable estimate of P_X that the Defender can treat it as the "true" pmf.

In the following we will always assume that $N/n = c$, since from the above analysis this turns out to be the most interesting case.

5.3 Analysis of the Payoff at the Equilibrium

Having derived the equilibrium point of the DG-TRb game, we are ready to analyze the payoff at the equilibrium to understand who, between the Defender and the Attacker is going to *win* the game. Our aim is to derive a result similar to the one derived in Chapter 3, so that given two pmf's P_X and P_Y, a false positive error exponent λ and a distortion constraint L, we can derive the *ultimate achievable* false negative error exponent at the equilibrium $\varepsilon^*_{tr,b}$.[8] Specifically, we would like to know whether it is possible for $\mathscr{D}$ to obtain a strictly positive value of $\varepsilon^*_{tr,b}$, thus ensuring that the false negative error probability tends to zero exponentially fast for increasing values of n.

From the knowledge of the equilibrium point, we can define the set $\Gamma^n_{tr,b}$ containing all the pairs of sequences (y^n, t^N), for which $\mathscr{A}$ is able to bring y^n within $\Lambda^{n,*}_{tr}$. By adopting the transportation formulation of the attack strategy, $\Gamma^n_{tr,b}$ can be expressed as a set of pairs of pmf's or types (P_{y^n}, P_{t^N}), that is:

$$\Gamma^n_{tr,b}(\lambda, L) = \{(P,Q) \in \mathcal{P}_n \times \mathcal{P}_N: \\ \exists\, S^n_{PV} \in \mathcal{A}^n(L,P) \text{ s.t. } (V,Q) \in \Lambda^{n,*}_{tr}(\lambda)\}. \quad (5.27)$$

We will find it convenient to fix the type Q and consider the set of types P_{z^n} for which (P_{z^n}, Q) belongs to the sets $\Lambda^{n,*}_{tr}$ and $\Gamma^n_{tr,b}$, that is:

$$\Lambda^{n,*}_{tr}(Q,\lambda) = \{P \in \mathcal{P}_n\colon (P,Q) \in \Lambda^{n,*}_{tr}(\lambda)\}, \quad (5.28)$$

$$\Gamma^n_{tr,b}(Q,\lambda,L) = \{P \in \mathcal{P}_n\colon \exists\, S_{PV} \in \mathcal{A}^n(L,P) \text{ s.t. } V \in \Lambda^{n,*}_{tr}(Q,\lambda)\}. \quad (5.29)$$

To go on, we need to generalize the above sets. To start with, we generalize the h function so that it can be applied to pmf's that do not

[8] For the sake of clarity, we specify the version of the game (i.e., b) in the subscript of the exponent, since this quantity will take a different value in the various setups. We will do the same for the set Γ.

necessarily belong to $\mathcal{P}_n$ or $\mathcal{P}_N$. By remembering that $N/n = c$, we introduce the following definition:

$$h_c(P, Q) = \mathcal{D}(P\|U) + c\mathcal{D}(Q\|U), \tag{5.30}$$

with

$$U = \frac{1}{1+c}P + \frac{c}{1+c}Q. \tag{5.31}$$

Note that when $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}_N$, the above definition is equivalent to (5.7). By using h_c instead of h, we can extend the definitions (5.29) and (5.28) to a generic pmf Q in $\mathcal{P}$.

The derivation of the false negative error exponent at the equilibrium passes through the asymptotic extensions of the sets:

$$\Gamma_{tr,b}(Q, \lambda, L) = \{P \in \mathcal{P}\colon \exists\, S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Lambda^*_{tr}(Q, \lambda)\}, \tag{5.32}$$

where

$$\Lambda^*_{tr}(Q, \lambda) = \{P\colon h_c(P, Q) < \lambda\}. \tag{5.33}$$

Of course, when P and Q are not empirical pmf's, the meaning of $\Lambda^{n,*}_{tr}$ as the acceptance region for H_0 (and that of $\Gamma_{tr,b}(Q, \lambda, L)$ as the set of points that can be moved inside the acceptance region by the Attacker) is lost.

The importance of the above definition is that for any source P_X, decay rate λ and maximum allowed per-letter distortion L, the set $\Gamma_{tr,b}(Q, \lambda, L)$, evaluated for $Q = P_X$, corresponds to the *indistinguishability region* of the DG-TRb game, i.e., the set of all the pmf's for which $\mathscr{D}$ does not succeed in distinguishing between H_0 and H_1 ensuring a false negative error probability that tends to zero exponentially fast. Equivalently, if $P_Y \in \Gamma_{tr,b}(P_X, \lambda, L)$, no strictly positive false negative error exponent can be achieved by $\mathscr{D}$. The above conclusions follow from the following theorem:

Theorem 5.4 (Asymptotic Payoff of the DG-TRb Game)**.** For the DG-TRb game, with $N/n = c$, the false negative error exponent at the equilibrium is given by

$$\varepsilon^*_{tr,b}(\lambda) = \min_Q \left[c \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \right], \tag{5.34}$$

leading to the following cases:

1. $\varepsilon^*_{tr,b} = 0$, if $P_Y \in \Gamma_{tr,b}(P_X, \lambda, L)$,
2. $\varepsilon^*_{tr,b} > 0$, if $P_Y \notin \Gamma_{tr,b}(P_X, \lambda, L)$.

Proof. The theorem is an application of the generalized Sanov's theorem (Section 2.4.2). The false negative error probability at the equilibrium, for a given n, can be written as

$$\begin{aligned} P_{\text{FN}} &= \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) P_Y(\Gamma^n_{tr,b}(Q, \lambda, L)) \\ &= \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) \sum_{P \in \Gamma^n_{tr,b}(Q,\lambda,L)} P_Y(\mathcal{T}(P)). \end{aligned} \tag{5.35}$$

We start by deriving an upper bound of the false negative error probability. By exploiting the usual bounds on the probability of a type class and the number of types in $\mathcal{P}_n$ (see Section 2.4.1), we can write:

$$\begin{aligned} P_{\text{FN}} &\leq \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) \sum_{P \in \Gamma^n_{tr,b}(Q,\lambda,L)} 2^{-n\mathcal{D}(P\|P_Y)} \\ &\leq \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P \in \Gamma^n_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y)} \\ &\leq \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y)} \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \cdot 2^{-n \min\limits_{Q \in \mathcal{P}_N} [\frac{N}{n}\mathcal{D}(Q\|P_X) + \min\limits_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y)]} \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \cdot 2^{-n \min\limits_{Q} [c\mathcal{D}(Q\|P_X) + \min\limits_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y)]}, \end{aligned} \tag{5.36}$$

where the last inequality is obtained by minimizing over Q without requiring that $Q \in \mathcal{P}_N$ and where the use of the minimum instead of the infimum is justified by the fact that $\Gamma^n_{tr,b}(Q, \lambda, L)$ and $\Gamma_{tr,b}(Q, \lambda, L)$ are compact sets. By taking the log and dividing by n we find:

$$-\frac{\log P_{\text{FN}}}{n} \geq \min_{Q \in \mathcal{C}} \left[c\mathcal{D}(Q\|P_X) + \min_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \right] + \alpha_n, \tag{5.37}$$

with $\alpha_n = |\mathcal{X}| \frac{\log(n+1)(N+1)}{n}$ tending to 0 when n tends to infinity.

We now turn to the analysis of a lower bound for P_{FN}. Let Q^* be the pmf achieving the minimum in (5.34). Due to the density of rational numbers within real numbers, we can find a sequence of pmf's $Q_n \in \mathcal{P}_n$ that tends to Q^* when n tends to infinity. By remembering that $N = nc$, the subsequence $Q_N = Q_{nc}$ also tends to Q^* when n (and hence N) tends to infinity.[9] We can write:

$$\begin{aligned} P_{\text{FN}} &= \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) P_Y(\Gamma^n_{tr,b}(Q, \lambda, L)) \\ &\geq P_X(\mathcal{T}(Q_N)) P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L)), \\ &\geq \frac{2^{-N\mathcal{D}(Q_N \| P_X)}}{(N+1)^{|\mathcal{X}|}} P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L)), \end{aligned} \tag{5.38}$$

where in the first inequality we have replaced the sum with the single element of the subsequence Q_N defined previously, and the second inequality derives from the usual lower bound on the probability of a type class (see Section 2.4.1). From (5.38), by taking the log and dividing by n, we obtain

$$-\frac{\log P_{\text{FN}}}{n} \leq c\mathcal{D}(Q_N \| P_X) - \frac{1}{n} \log P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L)) + \alpha'_n, \tag{5.39}$$

where, as in (5.37), $\alpha'_n = |\mathcal{X}| \frac{\log(N+1)}{n}$ tends to zero when n tends to infinity.

We now apply the generalized Sanov limit (see Theorem 2.1 in Section 2.4.2) for computing the term $P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L))$ in (5.39). To do so, we must show that $\Gamma^n_{tr,b}(Q_N, \lambda, L) \to \Gamma_{tr,b}(Q^*, \lambda, L)$, in the Hausdorff sense. This can be done by reasoning as in the proof of Theorem 3.4 (where we proved that $\Gamma^n_{ks}(P_X, \lambda, L) \xrightarrow{H} \Gamma_{ks}(P_X, \lambda, L)$). The only difference with respect to that case is the form of the acceptance region and its asymptotic counterpart. However, since the generalized test function h_c has a similar behavior to $\mathcal{D}$ and Q_N tends to Q^* as $n \to \infty$, it easy to see that $\delta_H(\Lambda^{n,*}_{tr}(Q^N), \Lambda^*_{tr}(Q^*)) \to 0$. Hence, the proof of the Hausdorff convergence of $\Gamma^n_{tr,b}$ to set $\Gamma_{tr,b}$ follows from the same arguments used for the know sources case.

[9] In order to simplify the analysis, we assume that c is a non-null integer value, the extension of the proof to non-integer values of c is tedious but straightforward.

Then, from the generalized Sanov's theorem, we get:

$$-\lim_{n\to\infty}\frac{1}{n}\log P_Y(\Gamma^n_{tr,b}(Q_N,\lambda,L)) = \min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)}\mathcal{D}(P\|P_Y). \quad (5.40)$$

Hence, by exploiting the continuity of the divergence function, for n large enough, we can write

$$-\frac{\log P_{\text{FN}}}{n} \le c\mathcal{D}(Q^*\|P_X)+\beta'_n+\min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)}\mathcal{D}(P\|P_Y)+\beta''_n+\alpha'_n, \quad (5.41)$$

where all the sequences α'_n, β'_n and β''_n tend to zero when n tends to infinity. By coupling Equations (5.37) and (5.41) and by letting $n\to\infty$, we eventually obtain:

$$-\lim_{n\to\infty}\frac{\log P_{\text{FN}}}{n} = \min_{Q}[c\cdot\mathcal{D}(Q\|P_X)+\min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)}\mathcal{D}(P\|P_Y)], \quad (5.42)$$

thus proving the theorem. □

According to Theorem 5.4, we can distinguish two cases depending on the relationship between P_X and P_Y. In the former case, for which the minimum in (5.34) is obtained by letting $Q=P_X$, it is not possible for $\mathscr{D}$ to obtain a strictly positive false negative error exponent while ensuring that the false positive error exponent is at least equal to λ. In the latter case, the two divergences in (5.34) can not be made simultaneously equal to zero, hence P_{FN} tends to zero exponentially fast. In other words, given λ and L, the condition $P_Y\notin\Gamma_{tr,b}(P_X,\lambda,L)$ ensures that the distance between P_Y and P_X is large enough to allow a reliable discrimination between H_0 and H_1, notwithstanding the presence of the adversary. As anticipated, then, $\Gamma_{tr,b}(P_X,\lambda,L)$ is the indistinguishability region of the DG-TRb game. A pictorial representation of the sets $\Lambda^{n,*}_{tr}$ and $\Gamma_{tr,b}$ is given in Figure 5.2.

5.3.1 DG-KS vs. DG-TRb

In this section we compare the performance achievable by $\mathscr{D}$ for the DG-KS and DG-TRb games. We start the analysis by comparing the indistinguishability regions of the two games, namely $\Gamma_{ks}(P_X,\lambda,L)$ and

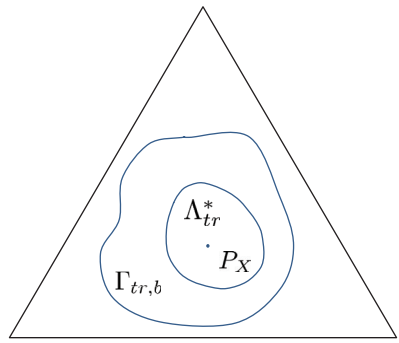


Figure 5.2: Geometric interpretation of Λ^*_{tr} and $\Gamma_{tr,b}$. When $P_Y \in \Gamma_{tr,b}$, a reliable distinction between H_0 and H_1 is not possible and the Attacker *wins* the game.

$\Gamma_{tr,b}(P_X, \lambda, L)$, reported below

$$\Gamma_{ks}(P_X, \lambda, L) = \{P \in \mathcal{P}\colon \exists\ S_{PV} \in \mathcal{A}(L, P),\ \text{s.t. } V \in \Lambda^*_{ks}(P_X, \lambda)\}, \tag{5.43}$$

with

$$\Lambda^*_{ks}(P_X, \lambda) = \{P \in \mathcal{P}\colon \mathcal{D}(P\|P_X) \leq \lambda\}; \tag{5.44}$$

and

$$\Gamma_{tr,b}(P_X, \lambda, L) = \{P \in \mathcal{P}\colon \exists S_{PV} \in \mathcal{A}(L, P),\ \text{s.t. } V \in \Lambda^*_{tr}(P_X, \lambda)\}, \tag{5.45}$$

with

$$\Lambda^*_{tr}(P_X, \lambda) = \{P \in \mathcal{P}\colon h_c(P, P_X) \leq \lambda\}. \tag{5.46}$$

We observe that the comparison between the two regions relies on the comparison between the divergence and the generalized log-likelihood function stated by the following:

Lemma 5.5 (Relationship between h_c and $\mathcal{D}$)**.** Let $N/n = c$, with $c \neq 0$, $c \neq \infty$, for any $P \neq P_X$ we have,

$$h_c(P, P_X) < \mathcal{D}(P\|P_X). \tag{5.47}$$

Proof. By rewriting $h_c(P, P_X)$ as in (5.24), we have:

$$h_c(P, P_X) = \mathcal{D}(P\|P_X) - (1 + c)\mathcal{D}(U\|P_X), \tag{5.48}$$

with $U = P/(1+c) + cP_X/(1+c)$, which is equal to P_X if and only if $P = P_X$, when we have $\mathcal{D}(U\|P_X) = 0$ thus yielding $h_c(P, P_X) = \mathcal{D}(P\|P_X) = 0$. □

From the above lemma, the strict inclusion between the acceptance regions, that is $\Lambda^*_{ks}(P_X, \lambda) \subset \Lambda^*_{tr}(P_X, \lambda)$, follows immediately.

From Lemma 5.5, we can prove the following theorem.

Theorem 5.6 (DG-TRb vs. DG-KS)**.** For any finite, non-null value of c, any P_X, $\lambda > 0$ and L, the following results hold:

- For any pmf P belonging to the boundary of $\Gamma_{tr,b}(P_X, \lambda, L)$ there exists a positive value τ such that $\mathcal{B}(P, \tau) \subset \overline{\Gamma_{ks}(P_X, \lambda, L)}$.
- For any pmf P belonging to the boundary of $\Gamma_{ks}(P_X, \lambda, L)$ there exists a positive value τ such that $\mathcal{B}(P, \tau) \subset \Gamma_{tr,b}(P_X, \lambda, L)$.
- $\Gamma_{ks}(P_X, \lambda, L) \subset \Gamma_{tr,b}(P_X, \lambda, L)$.

Proof. As an immediate consequence of Lemma 5.5, we have:

$$\Gamma_{ks}(P_X, \lambda, L) \subseteq \Gamma_{tr,b}(P_X, \lambda, L). \tag{5.49}$$

Point 1.
Let P' be a point on the boundary of $\Gamma_{tr,b}(P_X, \lambda, L)$. Since $\Gamma_{ks}(P_X, \lambda, L)$ is a closed set, we can prove that $\mathcal{B}(P', \tau) \subset \overline{\Gamma_{ks}(P_X, \lambda, L)}$ for some $\tau > 0$, by showing that $P' \in \overline{\Gamma_{ks}(P_X, \lambda, L)}$.

Let us assume, by contradiction, that $P' \in \Gamma_{ks}$ (be it inside or on the boundary). Then, we have that $\mathcal{D}(R'\|P_X) \leq \lambda$ for some map $S_{P'R'} \in \mathcal{A}(L, P')$; consequently, from Lemma 5.5, $h_c(R', P_X) < \lambda$, that is, R' is an internal point of Λ^*_{tr}. Let δ be such that $\mathcal{B}(R', \delta) \subset \Lambda^*_{tr}$. By exploiting Theorem A.2 in Appendix A, it is possible to fix $\tau > 0$ such that for any $P \in \mathcal{B}(P', \tau)$ a map $S_{PR} \in \mathcal{A}(L, P)$ exists such that $R \in \mathcal{B}(R', \delta)$ (specifically, we can choose $\tau = \delta/|\mathcal{X}|^2$). Then, by construction, $\mathcal{B}(P', \tau) \subset \Gamma_{tr,b}$, that is, P' is an internal point of $\Gamma_{tr,b}$, thus raising the absurd.

Point 2.
The proof of this point follows straightforwardly from Point 1. In fact, having proved that any point on the boundary of $\Gamma_{tr,b}$ lies outside Γ_{ks}, as a consequence, any point on the boundary of Γ_{ks} is an internal point of $\Gamma_{tr,b}$.[10] By definition of internal point, there exists $\tau > 0$ such that $\mathcal{B}(P, \tau) \subset \Gamma_{tr,b}$, thus concluding the proof.
Point 3.
From the above points, it follows that there is at least one point (in fact an infinite number of points) that belongs to $\Gamma_{tr,b}$ but not to Γ_{ks}, thus proving that the inclusion relation in (5.49) is a strict one. □

Theorem 5.6 has the following corollary.

Corollary 5.7. Let ε^*_{ks} and $\varepsilon^*_{tr,b}$ denote the error exponents at the equilibrium for the DG-KS and DG-TRb games. Then we have:

$$\varepsilon^*_{tr,b} \leq \varepsilon^*_{ks}, \tag{5.50}$$

where the equality holds if and only if $P_Y \in \Gamma_{ks}(P_X, \lambda, L)$, in which case both error exponents are equal to 0.

Proof. The corollary is obvious when $P_Y \in \Gamma_{tr,b}(P_X, \lambda, L)$, since in this case $\varepsilon^*_{tr,b} = 0$ while ε^*_{ks} is equal to zero if $P_Y \in \Gamma_{ks}(P_X, \lambda, L)$ and nonzero otherwise. When $P_Y \notin \Gamma_{tr,b}(P_X, \lambda, L)$, by considering the expression of the error exponent for the DG-TRb game we have:

$$\begin{aligned}
\varepsilon^*_{tr,b} &= \min_{Q} \Big[c \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \Big] \\
&\leq c\mathcal{D}(P_X\|P_X) + \min_{P \in \Gamma_{tr,b}(P_X,\lambda,L)} \mathcal{D}(P\|P_Y) \\
&\overset{(a)}{=} \min_{P \in \Gamma_{tr,b}(P_X,\lambda,L)} \mathcal{D}(P\|P_Y) \\
&< \min_{P \in \Gamma_{ks}(P_X,\lambda,L)} \mathcal{D}(P\|P_Y) = \varepsilon^*_{ks},
\end{aligned} \tag{5.51}$$

where the last strict inequality follows by the fact that the absolute minimum of $\mathcal{D}(P\|P_Y)$ is obtained for $P = P_Y$ which we have assumed to lie outside $\Gamma_{tr,b}(P_X, \lambda, L)$ and hence, due to the convexity of $\mathcal{D}$, the

[10] Note that this is true since $\Gamma_{ks} \subseteq \Gamma_{tr,b}$.

value of P satisfying the minimization on the right-hand side of equality (a) belongs to the boundary of $\Gamma_{tr,b}(P_X, \lambda, L)$ that, by Theorem 5.6, lies outside the closed set $\Gamma_{ks}(P_X, \lambda, L)$. □

Theorem 5.6 and Corollary 5.7 permit to conclude that binary detection with training data is more favorable to the Attacker than binary detection with known sources. The reason behind such a result is the use of the h function instead of the divergence, which in turn stems from the need for the Defender to ensure that the constraint on the false positive error probability is satisfied for all $P_X \in \mathcal{P}$. It is such a worst case assumption that ultimately favors the Attacker in the DG-TRb setup.

5.4 Game with Independent Training Sequences (DG-TRa)

We now pass to the analysis of version a of the detection game with training data (Definition 5.1). In this case, $\mathscr{D}$ and $\mathscr{A}$ rely on independent training sequences, namely t_D^N and t_A^K. Similarly to version b, we assume that both N and K grow linearly with n and that the asymptotic analysis is carried out by letting n go to infinity. As a matter of fact, assuming that K grows faster than N with respect to n is not reasonable in practical applications, since usually the Defender has a better knowledge of the system than the Attacker. On the contrary, one could consider the case where K grows less than linearly with n, thus considering a situation which is more favorable to the Defender.

Given the above, in the following, we assume that $N = cn$ and $K = dn$. As we already noted in Section 5.2, the strategy $\Lambda_{tr}^{n,*}$ identified by Lemma 5.2 is optimum regardless of the relationship between t_D^N and t_A^K, hence the only difference between versions a and b of the game is in the strategy of the Attacker. In fact, now the Attacker does not have a perfect knowledge of the acceptance region adopted by the Defender, since such a region depends on the empirical pmf of t_D^N which the Attacker does not know. In this case, finding the optimal attack strategy is a difficult task.

A reasonable strategy for the Attacker could be to use the empirical pmf of t_A^K, in place of the one derived from t_D^N. More precisely, by

using the notation introduced in Section 5.3 (Equation (5.28)), the Attacker could try to move y^n into $\Lambda_{tr}^{n,*}(P_{t_A^K})$, while the acceptance region adopted by the Defender is $\Lambda_{tr}^{n,*}(P_{t_D^N})$. Given that t_D^N and t_A^K are generated by the same source, their empirical pmf's will both tend to P_X when n goes to infinity, and hence using $\Lambda_{tr}^{n,*}(P_{t_A^K})$ should be *in some way* equivalent to using $\Lambda_{tr}^{n,*}(P_{t_D^N})$. In fact, in the following we will show that, given P_X, L and λ, the indistinguishability region for version a of the game, let us call it $\Gamma_{tr,a}(P_X, \lambda, L)$, is identical to the indistinguishability region of version b. Of course, this does not mean that the achievable payoff for the DG-TRa game is equal to that of the DG-TRb, since outside the indistinguishability region, the false negative error exponent for case a may be different (actually larger) than that of case b.

5.4.1 Training Sequences of the Same Length

We start our analysis by assuming that $c = d$ (and hence $N = K$), i.e., the training sequences available to the Defender and the Attacker have the same length.

Our goal is to investigate the asymptotic behavior of the payoff of the DG-TRa game for the profile $(\Lambda_{tr}^{n,*}(P_{t_D^N}), \tilde{S}_{YZ}^n)$, where the, not necessarily optimum, strategy $\tilde{S}_{YZ}^n(y^n, t_A^N)$ played by the Attacker is defined as:

$$\tilde{S}_{YZ}^n(P_{y^n}, P_{t_A^N}) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} h(S_Z^n, P_{t_A^N}). \tag{5.52}$$

We will use the map $\tilde{S}_{YZ}^n$ to bound the false negative error exponent and show that, even if the DG-TRa game is less favorable to the Attacker than the DG-TRb game, the two games have the same indistinguishability region.

By following the same flow of ideas used in Section 5.3, we consider the set of pmf's for which the Attacker is able to move P_{y^n} within the acceptance region, that is

$$\tilde{\Gamma}_{tr,a}^n(\lambda, L) = \{(P_{y^n}, P_{t_D^N}, P_{t_A^N}) \colon (\tilde{S}_Z^n(P_{y^n}, P_{t_A^N}), P_{t_D^N}) \in \Lambda_{tr}^{n,*}(\lambda)\}. \tag{5.53}$$

Similarly to version b of the game, we find it useful to introduce the following definition:

$$\tilde{\Gamma}^n_{tr,a}(P_{t_D^N}, P_{t_A^N}, \lambda, L) = \{P_{y^n} \in \mathcal{P}_n\colon \tilde{S}^n_Z(P_{y^n}, P_{t_A^N}) \in \Lambda^{n,*}_{tr}(P_{t_D^N}, \lambda)\}. \tag{5.54}$$

By using the generalized function h_c instead of h in the definition of the acceptance region, we can apply the above definition to any pair of pmf's. Specifically, given two pmf's Q and R, we define:

$$\tilde{\Gamma}^n_{tr,a}(Q, R, \lambda, L) = \{P \in \mathcal{P}_n\colon \tilde{S}^n_Z(P, R) \in \Lambda^{n,*}_{tr}(Q, \lambda)\}. \tag{5.55}$$

It is easy to see that:

$$\begin{aligned} \tilde{\Gamma}^n_{tr,a}(Q, R, \lambda, L) &\subseteq \tilde{\Gamma}^n_{tr,a}(Q, Q, \lambda, L) \\ \tilde{\Gamma}^n_{tr,a}(Q, Q, \lambda, L) &= \Gamma^n_{tr,b}(Q, \lambda, L), \end{aligned} \tag{5.56}$$

since when (and only when) $Q = R$, $\mathscr{A}$ performs his attack by using exactly the same acceptance region adopted by $\mathscr{D}$, while in all the other cases he can rely only on an estimate based on his own training sequence. Paralleling the analysis of the DG-TRb case, we introduce the asymptotic set

$$\tilde{\Gamma}_{tr,a}(Q, R, \lambda, L) = \{P \in \mathcal{P}\colon \tilde{S}_Z(P, R) \in \Lambda^*_{tr}(Q, \lambda)\}, \tag{5.57}$$

where $\Lambda^*_{tr}(Q, \lambda)$ is the same set defined in (5.33). Straightforwardly, the relations in (5.56) also hold for $\tilde{\Gamma}_{tr,a}$.

We are now ready to prove the following result.

Theorem 5.8 (Asymptotic Payoff of the DG-TRa Game)**.** The error exponent of the payoff associated to the profile $(\Lambda^{*,n}_{tr}(P_{t_D^N}), \tilde{S}^n_Z(P_{y^n}, P_{t_A^N}))$ can be lower and upper bounded as follows[11]

$$\tilde{\varepsilon}_{tr,a} \geq \min_{Q,R} \Big\{ c[\mathcal{D}(Q\|P_X) + \mathcal{D}(R\|P_X)] + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,R,\lambda,L)} \mathcal{D}(P\|P_Y) \Big\}, \tag{5.58}$$

$$\tilde{\varepsilon}_{tr,a} \leq \min_{Q} \Big[2c \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P\|P_Y) \Big]. \tag{5.59}$$

[11] Here we use $\tilde{\varepsilon}_{tr,a} = -\limsup_{n\to\infty} \frac{1}{n} \log(P_{\mathrm{FN}})$, since the limit may not exist.

Proof. The proof is similar to the proof of Theorem 5.4, with the noticeable difference that now the lower and upper bounds are different, hence preventing us to derive a precise expression for the error exponent. Let us start with the lower bound. By recalling the definition of the false negative error probability, for any n we can write:

$$
\begin{aligned}
P_{\mathrm{FN}} &= \sum_{t_D^N} \sum_{t_A^N} P_X(t_D^N) P_X(t_A^N) P_Y(\tilde{\Gamma}^n_{tr,a}(P_{t_D^N}, P_{t_A^N}, \lambda, L)) \\
&= \sum_{t_D^N} \sum_{t_A^N} P_X(t_D^N) P_X(t_A^N) \sum_{P \in \tilde{\Gamma}^n_{tr,a}(P_{t_D^N}, P_{t_A^N}, \lambda, L)} P_Y(\mathcal{T}(P)) \\
&= \sum_{Q \in \mathcal{P}_N} \sum_{R \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) P_X(\mathcal{T}(R)) \sum_{P \in \tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)} P_Y(\mathcal{T}(P)) \\
&\leq \sum_{Q \in \mathcal{P}_N} \sum_{R \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) P_X(\mathcal{T}(R)) \cdot (n+1)^{|\mathcal{X}|} \\
&\qquad \cdot 2^{-n \min\limits_{P \in \tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)} \mathcal{D}(P\|P_Y)} \\
&\leq \sum_{Q \in \mathcal{P}_N} P_X(\mathcal{T}(Q)) \cdot (n+1)^{|\mathcal{X}|} (N+1)^{|\mathcal{X}|} \\
&\qquad \cdot 2^{-n \min\limits_{R \in \mathcal{P}_N} \left[c\mathcal{D}(R\|P_X) + \min\limits_{P \in \tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)} \mathcal{D}(P\|P_Y) \right]} \\
&\leq (n+1)^{|\mathcal{X}|} (N+1)^{2|\mathcal{X}|} \\
&\qquad \cdot 2^{-n \min\limits_{Q,R} \left[c\mathcal{D}(Q\|P_X) + c\mathcal{D}(R\|P_X) + \min\limits_{P \in \tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)} \mathcal{D}(P\|P_Y)) \right]},
\end{aligned}
\tag{5.60}
$$

where the use of the minimum instead of the infimum is justified by the compactness of the involved sets, and where in the last inequality we replaced the minimization over all Q and R in $\mathcal{P}_N$, with a minimization over the entire space of pmf's. By taking the logarithm of both sides and letting n tend to infinity, we get the lower bound in (5.58).

We now turn the attention to the upper bound. To do so, let Q^* be the pmf achieving the minimum in (5.59). Due to the density of rational numbers within real numbers, we can find two sequences of pmf's Q_n and R_n that tend to Q^* when n tends to infinity, and such that $Q_n \in \mathcal{P}_n$, $R_n \in \mathcal{P}_n, \forall n$. By remembering that $N = nc$, we can

say that the subsequences $Q_N = Q_{nc}$ and $R_N = R_{nc}$ also tend to Q^* when n (and hence N) tends to infinity. We can, then, consider the subsequences Q_N and R_N and write the following chain of inequalities:

$$\begin{aligned} P_{\text{FN}} &= \sum_{Q\in\mathcal{P}_N}\sum_{R\in\mathcal{P}_N} P_X(\mathcal{T}(Q))P_X(\mathcal{T}(R))P_Y(\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)) \\ &\geq P_X(\mathcal{T}(Q_N))P_X(\mathcal{T}(R_N))P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L)) \\ &\geq \frac{2^{-N(\mathcal{D}(Q_N\|P_X)+\mathcal{D}(R_N\|P_X))}}{(N+1)^{2|\mathcal{X}|}}\cdot P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L)), \end{aligned} \tag{5.61}$$

where the first inequality has been obtained by replacing each summation with a single element of the sum (two elements of the sequences Q_N and R_N), and the second relies on the usual lower bound on the probability of a type class (see Section 2.4.1). By taking the logarithm of each side in (5.61) and dividing by n, we get:

$$\begin{aligned} -1/n\log(P_{\text{FN}}) \leq\ & c\mathcal{D}(Q_N\|P_X) + c\mathcal{D}(R_N\|P_X) \\ & - 1/n\log(P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L))) - \beta_n, \end{aligned} \tag{5.62}$$

with $\beta_n = 2|\mathcal{X}|\log(N+1)$ tending to 0 for $n\to\infty$.

In order to apply the generalized Sanov's theorem for evaluating the probability term in (5.62), we need to prove that

$$\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L) \overset{H}{\to} \Gamma_{tr,b}(Q^*,\lambda,L). \tag{5.63}$$

We observe that the proof of such convergence is more involved with respect to the similar proofs in Theorems 3.4 and 5.4, due to the complicated expression of $\tilde{\Gamma}^n_{tr,a}$. In fact, this time, the Attacker does not know the exact form of the acceptance region adopted by $\mathscr{D}$, i.e., $\Lambda^{n,*}_{tr}(Q_N)$, and considers the estimated version $\Lambda^{n,*}_{tr}(R_N)$ to carry out the minimization. Accordingly, set $\tilde{\Gamma}^n_{tr,a}$ can not be written in a form similar to (5.29) (and (3.20)), thus preventing us from directly using the same arguments used therein. Instead, we will prove the Hausdorff convergence of $\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L)$ to $\Gamma_{tr,b}(Q^*,\lambda,L)$ by resorting to the definition of an auxiliary set.

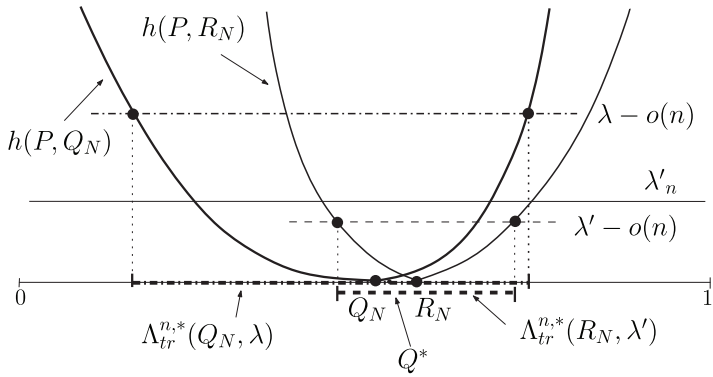


Figure 5.3: Geometric construction of $\Lambda_{tr}^{n,*}(R_N, \lambda')$. For ease of graphical representation, the case with $|\mathcal{X}| = 2$ is depicted. $o(n) = (|\mathcal{X}| \log(n+1)(N+1))/n$.

Let $\lambda'_n = \max\{\lambda' \colon \Lambda_{tr}^{n,*}(R_N, \lambda') \subseteq \Lambda_{tr}^{n,*}(Q_N, \lambda)\}$.[12] We can define the following set:

$$\dot{\Gamma}_{tr,a}^{n}(Q_N, R_N, \lambda, L) = \{P \in \mathcal{P}_n \colon \exists\, S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Lambda_{tr}^{n,*}(R_N, \lambda'_n)\}\,. \tag{5.64}$$

By the definition of λ'_n, it is easy to see that the above set is contained in $\tilde{\Gamma}_{tr,a}(Q_N, R_N, \lambda, L)$. Then, the following chain of inclusions holds:

$$\dot{\Gamma}_{tr,a}^{n}(Q_N, R_N, \lambda, L) \subseteq \tilde{\Gamma}_{tr,a}^{n}(Q_N, R_N, \lambda, L) \subseteq \Gamma_{tr,b}^{n}(Q_N, \lambda, L).$$

Since $\Gamma_{tr,b}^{n}(Q_N, \lambda, L) \xrightarrow{H} \Gamma_{tr,b}(Q^*, \lambda, L)$ (see the proof of Theorem 5.4), by applying the squeeze theorem, (5.63) is proven if we show that[13]

$$\dot{\Gamma}_{tr,a}^{n}(Q_N, R_N, \lambda, L) \xrightarrow{H} \Gamma_{tr,b}(Q^*, \lambda, L).$$

By reasoning as in the proof of Theorems 3.4 and 5.4, the above relation follows by proving that $\delta_H(\Lambda_{tr}^{n,*}(R_N, \lambda'_n), \Lambda^*(Q^*, \lambda)) \to 0$ as $n \to \infty$, which derives easily from the density of rational numbers into real ones, the continuity of the h_c function and the fact that $R_N \to Q^*$ and $\lambda'_n \to \lambda$ as n tends to infinity.

[12] Notice that, since Q_N and R_N tend to the same pmf Q^* as n tends to infinity, and $\lambda > 0$, if n is sufficiently large, the set is non-empty (see Figure 5.3).

[13] The squeeze theorem (known also as sandwich theorem) also holds in the case of Hausdorff convergence [84].

The assumptions of the generalized Sanov's theorem are then satisfied and we can write:

$$-\lim_{n\to\infty} 1/n \log(P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N, R_N, \lambda, L)) = \min_{P\in\tilde{\Gamma}_{tr,a}(Q^*,Q^*,\lambda,L)} \mathcal{D}(P\|P_Y). \tag{5.65}$$

Therefore, by going on from (5.62), letting $n \to \infty$ and exploiting the continuity of $\mathcal{D}$ with respect to its arguments, we have

$$\tilde{\varepsilon}_{tr,a} \leq c\mathcal{D}(Q^*\|P_X) + c\mathcal{D}(Q^*\|P_X) + \min_{P\in\tilde{\Gamma}_{tr,a}(Q^*,Q^*\lambda,L)} \mathcal{D}(P\|P_Y). \tag{5.66}$$

By recalling that

$$Q^* = \arg\min_{Q} \left[2c \cdot \mathcal{D}(Q\|P_X) + \min_{P\in\tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P\|P_Y)\right], \tag{5.67}$$

we get the upper bound in (5.59). □

Theorem 5.8 has an important corollary.

Corollary 5.9 (Indistinguishability Region of the DG-TRa Game)**.** The false negative error exponent associated to the profile $(\Lambda^{n,*}_{tr}(P_{t^N_D}), \tilde{S}_{YZ}(\cdot, P_{t^N_A}))$ is equal to zero if and only if $P_Y \in \Gamma_{tr,b}(P_X, \lambda, L)$, and hence the indistinguishability region of the DG-TRa game is equal to that of the DG-TRb game.

Proof. From the upper bound in Theorem 5.8, it follows that $\tilde{\varepsilon}_{tr,a} = 0$ if $P_Y \in \tilde{\Gamma}_{tr,a}(P_X, P_X, \lambda, L)$, whereas from the lower bound we see that $\tilde{\varepsilon}_{tr,a} = 0$ implies $P_Y \in \tilde{\Gamma}_{tr,a}(P_X, P_X, \lambda, L)$. By remembering that $\tilde{\Gamma}_{tr,a}(P_X, P_X, \lambda, L) = \Gamma_{tr,b}(P_X, L, \lambda)$, the corollary is proven. □

Corollary 5.9 provides an interesting insight into the achievable performance of the DG-TRa game. While, in general, version a of the game is less favorable to the Attacker than version b, since in the latter case the Attacker knows exactly the acceptance region adopted by the Defender, if the Attacker adopts the strategy $\tilde{S}_{YZ}$, the indistinguishability regions of the two games are the same. Such a strategy, then, is optimal at least as far as the indistinguishability region is concerned. Outside that region, the Attacker could achieve a higher payoff (i.e., a lower

error exponent) by adopting a different strategy. On the other hand, a strategy that allows the Attacker to reach the same payoff as for version b may not exist.

5.4.2 Training Sequences with Different Length

We conclude this section by briefly discussing the case in which the training sequences t_D^N and t_A^K have different lengths, i.e., $c \neq d$. To simplify the analysis we assume that the length of t_D^N, i.e., c, is known to the Attacker; in this way $\mathscr{A}$ knows at least the form the h_c function used by $\mathscr{D}$. We focus on the following attack strategy: use the training sequence t_A^K to estimate $P_{t_D^N}$ and use such estimate to attack the sequence y^n. Specifically, the Attacker may use the following estimate of $P_{t_D^N}$:

$$\begin{aligned} \tilde{P}_{t_D^N}(i) &= \frac{1}{N}\lfloor P_{t_A^K}(i) \cdot N \rfloor \quad \forall i = 1 \dots |\mathcal{X}| - 1, \\ \tilde{P}_{t_D^N}(|\mathcal{X}|) &= 1 - \sum_{i=1}^{|\mathcal{X}|-1} \tilde{P}_{t_D^N}(i), \end{aligned} \tag{5.68}$$

to implement the attack function:

$$\tilde{S}_{YZ}^n(P_{y^n}, P_{t_A^K}) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} h_c(P_{z^n}, \tilde{P}_{t_D^N}). \tag{5.69}$$

With the above definitions, we can easily extend the analysis carried out for the case $c = d$ and obtain very similar results. Specifically, the upper bound in Theorem 5.8 can be rewritten as:

$$\tilde{\varepsilon}_{tr,a} \leq \min_Q \left[(c+d) \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P\|P_Y) \right], \tag{5.70}$$

whose proof is practically identical to the proof of Theorem 5.8 and is omitted for sake of brevity. By observing that the performance achievable by the Defender in version a of the game are at least as good as those achievable in version b (since in the latter case $\mathscr{A}$ knows exactly the acceptance region adopted by $\mathscr{D}$ and hence his attacks will surely be more effective), Equation (5.70) permits to conclude that the indistinguishability region is equal to that obtained for the case $c = d$.

5.5 Security Margin in the DG-TR Setup

As we did for the game with known sources, we now study the behavior of the DG-TR game when λ tends to zero in order to investigate the *best achievable* performance for the Defender and compute the Security Margin in this case. The analysis goes along the same steps as for the DG-KS case.

We derive our results by focusing on the game with equal training sequences (DG-TRb). From the analysis carried out in the previous sections, we know that, as long as the length of the sequences t_D^N and t_A^K grows linearly with n, the indistinguishability region is the same for both versions of the game. By relying on this result, it is straightforward to prove that the Security Margin is the same even for the DG-TRa case.

We start by rewriting the set $\Gamma_{tr}(Q, \lambda, L)$ in (5.32)–(5.33) by exploiting the optimal transport interpretation:[14]

$$\Gamma_{tr}(Q, \lambda, L) = \{P \in \mathcal{P}\colon \exists\ R \in \Lambda_{tr}^*(Q, \lambda) \text{ s.t. } EMD(P, R) \leq L\}, \quad (5.71)$$

where

$$\Lambda_{tr}^*(Q, \lambda) = \{P \in \mathcal{P}\colon h_c(P, Q) \leq \lambda\}. \quad (5.72)$$

From Section 5.3, we know that, when $Q = P_X$ the above set corresponds to the indistinguishability region for the DG-TR game, namely $\Gamma_{tr}(P_X, \lambda, L)$.

Then, we observe that the $\mathcal{D}$ and h_c have a similar behavior, in that both $\mathcal{D}(P\|Q)$ and $h_c(P, Q)$ are convex functions in P and are equal to zero if and only if $P = Q$. Hence, Proposition 4.1 can be extended to the set $\Gamma_{tr}(Q, \lambda, L)$, yielding the following.

Proposition 5.1. For any two values λ_1 and λ_2 such that $\lambda_2 < \lambda_1$, $\Gamma_{tr}(Q, \lambda_2, L) \subseteq \Gamma_{tr}(Q, \lambda_1, L)$.

In a similar way, Lemma B.1 (Appendix B.1) can be extended to the set $\Gamma_{tr}(Q, \lambda, L)$ (see discussion at the end of the same appendix), permitting to prove the counterpart of Theorem 4.1 for the detection game with training data.

[14] Since the indistinguishability region is the same for all the versions of the game with training data, from now on we adopt the notation Γ_{tr} instead of $\Gamma_{tr,b}$.

Theorem 5.10. Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum allowable average per-letter distortion L, the maximum achievable false negative error exponent for the DG-TRb game is:

$$\lim_{\lambda \to 0} \varepsilon^*_{tr,b}(\lambda) = \min_Q \left[c \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P\|P_Y) \right], \tag{5.73}$$

where $\Gamma(Q, L)$ is defined as in (4.2) by replacing P_X with Q.

Proof. The proof goes along the same line of the proof of Theorem 4.1 in Section 4.2. We know from Theorem 5.4 that the expression of the false negative error exponent of the DG-TRb game at the equilibrium is given by

$$\varepsilon^*_{tr,b}(\lambda) = \min_Q \left[c \cdot \mathcal{D}(Q\|P_X) + \min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \right]. \tag{5.74}$$

From Proposition 5.1, we see immediately that $\varepsilon^*_{tr,b}(\lambda)$ is non-increasing when λ decreases, since the innermost minimization in (5.74) is taken over a smaller set when λ decreases. Then, by the same token, we have:

$$\varepsilon^*_{tr,b}(\lambda) \leq \min_Q \left(c\mathcal{D}(Q\|P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P\|P_Y) \right). \tag{5.75}$$

This implies that $\lim_{\lambda \to 0} \varepsilon^*_{tr,b}(\lambda)$ exists and is finite. Given that Lemma B.1 still holds for the set $\Gamma_{tr}(Q, \lambda, L)$ $\forall Q$, we can reason as in the proof of Theorem 4.1 to conclude that:

$$\min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \geq \min_{P \in \Gamma(Q,L)} \mathcal{D}(P\|P_Y) - \delta(\tau), \tag{5.76}$$

where $\delta(\tau)$ can be made arbitrarily small by decreasing λ. By adding the term $c\mathcal{D}(Q\|P_X)$ to both sides of (5.76) and considering that the relation holds for any $Q \in \mathcal{P}$, we can write:

$$\begin{aligned} \varepsilon^*_{tr,b}(\lambda) &= \min_Q \left[c\mathcal{D}(Q\|P_X) + \min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P\|P_Y) \right] \\ &\geq \min_Q \left[c\mathcal{D}(Q\|P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P\|P_Y) \right] - \delta(\tau), \end{aligned} \tag{5.77}$$

which concludes the proof due to the arbitrariness of $\delta(\tau)$. □

A consequence of Theorem 5.10 is that $\lim_{\lambda \to 0} \varepsilon^*_{tr,b}(\lambda) = 0$ if and only if $P_Y \in \Gamma(P_X, L)$, which then can be seen as the smallest indistinguishability region for the TRb setting.

From the above theorem, we can conclude that the *smallest* indistinguishability regions for the setup with known sources and training data are the same,[15] thus implying that the Security Margin in the DG-TR setting, say $\mathcal{SM}_{tr}$, is the same of the Security Margin in the DG-KS case, that is

$$\mathcal{SM}_{tr}(P_X, P_Y) = EMD(P_X, P_Y). \tag{5.78}$$

We remark that, for any allowed distortion $L < EMD(P_X, P_Y)$, the minimum value of the false positive error exponent (λ) which allows the Defender to take a reliable decision in the DG-TR setting is lower than that in the DG-KS setting. However, as a result, the difference between the two settings regards the *decay rate* of the error probabilities and not the ultimate distinguishability of the sources.

[15] Then, when λ tends to zero, we do not need to differentiate anymore between DG-KS and DG-TR in the definition of Γ.

The version of record is available at: http://dx.doi.org/10.1561/0100000102

6

Binary Detection Games with Corrupted Training

In this chapter, we extend the analysis of the binary detection game with training data by considering a scenario in which the Attacker interferes with the learning phase by corrupting part of the training sequence.

From a theoretical point of view, this represents a major deviation from the analysis carried out in the previous chapters. The first and most important consequence of the possibility that the training sequence has been corrupted by the Attacker, is that now the attack influences also the accuracy of the decision under H_0. In other words, the action of the Attacker has an impact on both the false positive and false negative error probabilities. This was not the case in the previous setups, where the false positive error probability was independent of the strategy chosen by the Attacker. As a result, the fulfilment of the constraint on the false positive error probability requires that the possible actions of the attacker are taken into account, by adopting a worst case approach. Such a fundamental modification of the structure of the game influences all the rest of the analysis, thus calling for the adoption of more complex tools, and leading to more general results that incorporate those derived in Chapters 3–5 as limit cases, but substantially depart from them.

More specifically, by modeling the interplay between $\mathscr{A}$ and $\mathscr{D}$ as a game, the set of strategies available to the Defender corresponds to the possible detection rules he can adopt, while the Attacker must decide how to corrupt the training data, up to his maximum capacity, and the test data, subject to a distortion constraint, so to induce a decision error. After providing a rigorous definition of the game, we derive the optimal strategy for the Defender and the optimal corruption strategy for the Attacker when the length of the training and the observed sequences tend to infinity. Then, we compute the payoff at the equilibrium and analyze the best achievable performance when the Type I and II error probabilities tend to zero exponentially fast. Specifically, by mimicking and extending the analysis in Chapter 4, we study the distinguishability of any two sources as a function of the percentage of training samples corrupted by the Attacker and when the test sequence can be modified up to a certain distortion level. It turns out that the distinguishability of the sources can be summarized by two parameters, namely the Security Margin, which now depends of the corruption level of the training data, and the *blinding corruption level*, defined as the maximum portion of the training sequence corrupted by the Attacker that still allows a reliable distinction between the sources (i.e., ensuring positive error exponents for the two kinds of errors of the test).

We consider two different scenarios wherein the Attacker is allowed respectively to *add* some fake samples to the training sequence and to *replace* some samples of the training sequence with fake ones. As we will see, the second case is more favorable to the Attacker, since a lower distortion and a lower number of corrupted training samples are necessary to prevent the correct decision.

6.1 Discussion and Link with Adversarial Machine Learning

Before going on with the formalization of the game, we briefly pause to discuss the setting studied in this chapter. From a practical point of view, encompassing the case of a corrupted training sequence permits to extend the applicability of the theoretical analysis to situations in which the collection of the training data is not under the full control of the analyst. This is the case in many modern applications of machine

learning, wherein the data used in the training phase is collected in a non-controlled environment, e.g., by resorting to crowdsourcing or on-line learning with the risk that part of the data is altered with the aim of facilitating a subsequent attack [85]–[88]. In this sense, the results presented in this chapter are strongly related to adversarial machine learning [15]. Due to the natural vulnerability of machine learning systems, in fact, the Attacker may take an important advantage if no countermeasures are adopted by the Defender. The use of a training sequence to gather information about the statistical characterization of the sources can be seen as a very simple learning mechanism, therefore, the analysis of the impact that an attack carried out in such a phase has on the performance of a decision system, may help shedding new light on this important problem.

By referring to the taxonomy introduced in adversarial machine learning [15], the scenario considered in this chapter (and in the monograph) corresponds to a case of *integrity violation* attack, that is, an attack aiming at causing a false negative error at test time.

6.2 Detection Game with Corrupted Training (DG-CTR)

We now formalize the problem of binary detection in the presence of corrupted training samples.

Given a discrete and memoryless source $X \sim P_X$ and a test sequence z^n, the goal of $\mathscr{D}$ is to decide whether z^n has been drawn from X (hypothesis H_0) or not (alternative hypothesis H_1). By adopting a Neyman–Pearson perspective, we assume that $\mathscr{D}$ must ensure that the Type-I error probability is lower than a given threshold. Similarly to the previous versions of the game, we assume that $\mathscr{D}$ relies only on first order statistics to make a decision. In addition, we study the asymptotic version of the game when n tends to infinity, by requiring that P_{FP} decays exponentially fast when n increases, with an error exponent at least equal to λ, i.e., $P_{\text{FP}} \leq 2^{-n\lambda}$. On his side, the Attacker aims at inducing a Type-II error. Specifically, $\mathscr{A}$ takes a sequence y^n drawn from a source $Y \sim P_Y$ and modifies it in such a way that $\mathscr{D}$ decides that the modified sequence z^n has been generated by X. In doing so, $\mathscr{A}$ must respect a distortion constraint requiring that the average per-letter

distortion between y^n and z^n is lower than L. Both $\mathscr{A}$ and $\mathscr{D}$ know the statistics of X through a training sequence, which can be partly corrupted by $\mathscr{A}$. Depending on how the training sequence is modified by the Attacker, we can define different versions of the game. Specifically, we focus on the following two cases: in the first case, hereafter referred to as *detection game with addition of corrupted samples*, namely DG-CTRa game, the Attacker can add some fake samples to the original training sequence. This case is studied in Sections 6.3 through 6.5. In the second case, analyzed in Sections 6.6 through 6.8, the Attacker can replace some of the training samples with fake values. In the following, this case is referred to as *detection game with replacement of training samples*, namely DG-CTRr game. It is worth stressing that, even if the goal of the Attacker is to increase the false negative error probability, the training sequence is corrupted *regardless* of whether H_0 or H_1 holds, hence, in general, this part of the attack also affects the false positive error probability. As it will be clear later on, this forces the Defender to adopt a worst case perspective to ensure that P_{FP} is surely lower than $2^{-\lambda n}$.

As to Y, we assume that the Attacker knows P_Y exactly. For a proper definition of the payoff of the game, we also assume that the Defender knows P_Y, with the understanding that, as it was the case with the previous games, the optimal strategy of $\mathscr{D}$ does not depend on P_Y thus allowing us to relax the assumption that the Defender knows P_Y.

6.3 The DG-CTRa Game

6.3.1 The Adversarial Setup

A schematic representation of the scenario addressed in this section is given in Figure 6.1.

Let τ^{m_1} be a sequence drawn from X. We assume that τ^{m_1} is accessible to $\mathscr{A}$, who corrupts it by concatenating to it a sequence of fake samples τ^{m_2}. Then $\mathscr{A}$ reorders the overall sequence in a random way in order to hide the position of the fake samples. Note that reordering does not alter the statistics of the training sequence since the sequence is supposed to be generated from a memoryless source. In the following,

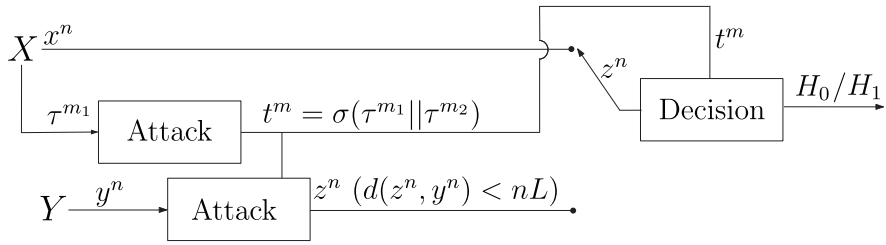


Figure 6.1: Schematic representation of the DG-CTRa setup. The Attacker corrupts both the training and the test sequence. The symbol $\|$ denotes sequence concatenation, while σ denotes a random permutation of the sequence samples.

we denote by m the final length of the training sequence ($m = m_1 + m_2$), and by $\alpha = \frac{m_2}{m_1+m_2}$ the fraction of fake samples within it. The corrupted training sequence observed by $\mathscr{D}$ is denoted by t^m. Eventually, we hypothesize a linear relationship between the lengths of the test and the corrupted training sequence, i.e., $m = cn$, for some constant value c. Since we are interested in studying the equilibrium point of the game when the length of the test and training sequences tend to infinity, strictly speaking, we should ensure that when n grows, all the quantities m, m_1 and m_2 are integer numbers for the given c and α. In practice, we will neglect such an issue, since when n grows the ratios m/n and $m_2/(m_1 + m_2)$ can approximate any real values c and α. More rigorously, we could consider only rational values of c and α, and focus on subsequences of n including only those values for which $m/n = c$ and $m_2/(m_1 + m_2) = \alpha$.

The goal of $\mathscr{D}$ is to decide if an observed sequence has been drawn from the same source that generated t^m (H_0) or not (H_1). We assume that $\mathscr{D}$ knows that a certain percentage of samples in the training sequence has been corrupted, but he has no clue about the position of the corrupted samples. The Attacker can also modify the sequence generated by Y so to induce a decision error. The possibly corrupted sequence observed by $\mathscr{D}$ is denoted by z^n. With regard to the two phases of the attack, we assume that $\mathscr{A}$ first corrupts the training sequence, then he modifies the sequence y^n. This means that, in general, z^n will depend both on y^n and t^m, while t^m (noticeably τ^{m_2}) does not depend on y^n. Stated in another way, the corruption of the training sequence

can be seen as a preparatory part of the attack, whose goal is to ease the subsequent camouflage of y^n.

6.3.2 Definition of the DG-CTRa Game

The DG-CTRa $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game is a zero-sum game defined by the set of strategies available to $\mathscr{D}$ and $\mathscr{A}$, respectively $\mathcal{S}_{\mathscr{D}}$ and $\mathcal{S}_{\mathscr{A}}$, and their corresponding opposite payoffs, as detailed in the following.

Defender's Strategies

Since $\mathscr{D}$ relies only on the first order statistics of z^n and t^m, the acceptance region of hypothesis H_0, hereafter referred to as $\Lambda^{n\times m}$, is a union of pairs of type classes,[1] or equivalently, pairs of types (P, R), where $P \in \mathcal{P}_n$ and $R \in \mathcal{P}_m$. As in the previous games, $\mathscr{D}$ follows a Neyman–Pearson approach, requiring that the false positive error probability tends to zero exponentially fast with a decay rate at least equal to λ. Given that the pmf P_X ruling the emission of sequences under H_0 is not known and given that the corruption of the training sequence is going to impair $\mathscr{D}$'s decision under H_0, we adopt a *worst case approach* and require that the constraint on the false positive error probability holds for all possible P_X and for all the possible strategies available to the Attacker. In this way, the space of strategies available to $\mathscr{D}$ is defined as follows:

$$\mathcal{S}_{\mathscr{D}} = \left\{ \Lambda^{n\times m} \subset \mathcal{P}_n \times \mathcal{P}_m \colon \max_{P_X \in \mathcal{P}} \max_{s \in \mathcal{S}_{\mathscr{A}}} P_{\mathrm{FP}} \leq 2^{-\lambda n} \right\}, \tag{6.1}$$

where $P_{\mathrm{FP}} = P_{XY}((z^n, t^m) \notin \Lambda^{n\times m})$.[2]

We will refine this definition at the end of the next section, after the exact definition of the space of strategies the Attacker can choose from.

[1] We use the superscript $n \times m$ to indicate explicitly that $\Lambda^{n\times m}$ refers to n-long test sequences and m-long training sequences (with $m = cn$).

[2] As it will be more clear in the sequel, the dependence of this probability term on Y holds in the case of targeted attack (DG-CTRat setup), where the corruption of the training sequence is targeted to the counterfeiting of the sequence y^n.

Attacker's Strategies

The attack carried out by $\mathscr{A}$ consists of two parts. Given an original training sequence τ^{m_1}, the Attacker first generates a sequence of fake samples τ^{m_2} and mixes them up with those in τ^{m_1}, producing the training sequence t^m observed by $\mathscr{D}$. Then, given a sequence y^n drawn from P_Y, he transforms y^n into z^n, eventually trying to generate a pair of sequences (z^n, t^m)[3] whose types belong to $\Lambda^{n\times m}$. In doing so, he must ensure that $d(y^n, z^n) \leq nL$ for some additive distortion function d.

Let us consider the corruption of the training sequence first. Given that $\mathscr{D}$ makes his decision by relying only on the type of t^m, we are interested in the effect that the addition of the fake samples has on P_{t^m}. By considering the different lengths of τ^{m_1} and τ^{m_2}, we have:

$$P_{t^m} = \alpha P_{\tau^{m_2}} + (1-\alpha) P_{\tau^{m_1}}, \tag{6.2}$$

where $P_{t^m} \in \mathcal{P}_m$, $P_{\tau^{m_1}} \in \mathcal{P}_{m_1}$ and $P_{\tau^{m_2}} \in \mathcal{P}_{m_2}$. The first part of the attack, then, is equivalent to choosing a pmf in $\mathcal{P}_{m_2}$ and mixing it up with $P_{\tau^{m_1}}$. By the same token, the choice of the Attacker depends only on $P_{\tau^{m_1}}$ rather than on the single sequence τ^{m_1}. Arguably, the best choice of the pmf in $\mathcal{P}_{m_2}$ will depend on P_Y, since the corruption of the training sequence is instrumental to let the Defender think that a sequence generated by Y has been drawn by the same source that generated t^m.

Regarding the second phase of the attack, that is, the attack applied to the test sequence, we define the attack as the choice of a transportation map $S^n_{YZ}(y^n, t^m)$ among all the *admissible* maps $\mathcal{A}^n(L, P_{y^n})$ (clearly, the choice of the transportation map will depend on both y^n and t^m). As already noticed in Section 5.2, $S^n_{YZ}(y^n, t^m)$ depends on the sequences through their empirical pmf, then, in the following, we will use the notation $S^n_{YZ}(P_{y^n}, P_{t^m})$.

[3] While reordering is essential to hide the position of fake samples to $\mathscr{D}$, it does not have any impact on the position of (z^n, t^m) with respect to $\Lambda^{n\times m}$, since we assumed that the Defender bases his decision only on the first order statistic of the observed sequences. For this reason, we omit to indicate the reordering operator σ in the attack procedure.

With the above ideas in mind, the set of strategies of the Attacker can be defined as follows:

$$\mathcal{S}_{\mathscr{A}} = \mathcal{S}_{\mathscr{A},T} \times \mathcal{S}_{\mathscr{A},O}, \tag{6.3}$$

where $\mathcal{S}_{\mathscr{A},T}$ and $\mathcal{S}_{\mathscr{A},O}$ indicate, respectively, the part of the attack affecting the training sequence and the observed sequence, and are defined as:

$$\mathcal{S}_{\mathscr{A},T} = \{Q(P_{\tau^{m_1}})\colon \mathcal{P}_{m_1} \to \mathcal{P}_{m_2}\}, \tag{6.4}$$

$$\mathcal{S}_{\mathscr{A},O} = \{S^n_{YZ}(P_{y^n}, P_{t^m})\colon \mathcal{P}_n \times \mathcal{P}_m \to \mathcal{A}^n(L, P_{y^n})\}. \tag{6.5}$$

Note that the first part of the attack ($\mathcal{S}_{\mathscr{A},T}$) is applied regardless of whether H_0 or H_1 holds, while the second part ($\mathcal{S}_{\mathscr{A},O}$) is applied only under H_1. We also stress that the choice of $Q(P_{\tau^{m_1}})$ depends only on the training sequence τ^{m_1}, while the transportation map used in the second phase of the attack depends on both y^n and τ^{m_1} (through t^m).

Given the above definitions, the set of strategies of the Defender can be redefined by explicitly indicating that the constraint on the false positive error probability must be verified for all possible choices of $Q(\cdot) \in \mathcal{S}_{\mathscr{A},T}$, since this is the only part of the attack affecting P_{FP}. Specifically, we can rewrite (6.1) as follows:

$$\mathcal{S}_{\mathscr{D}} = \left\{\Lambda^{n\times m} \subset \mathcal{P}_n \times \mathcal{P}_m\colon \max_{P_X} \max_{Q(\cdot)\in\mathcal{S}_{\mathscr{A},T}} P_{\text{FP}} \leq 2^{-\lambda n}\right\}, \tag{6.6}$$

where $P_{\text{FP}} = P_X((z^n, \tau^{m_1})\colon (P_{z^n}, \alpha Q(P_{\tau^{m_1}}) + (1-\alpha)P_{\tau^{m_1}}) \notin \Lambda^{n\times m})$.

Payoff

The payoff of the game is defined in terms of the false negative error probability, namely:

$$u(\Lambda^{n\times m}, (Q(\cdot),\ S^n_{YZ}(\cdot,\cdot))) = -P_{\text{FN}}, \tag{6.7}$$

where

$$P_{\text{FN}} = P_{XY}((y^n, \tau^{m_1})\colon (S^n_Z(P_{y^n}, P_{t^m}), \alpha Q(P_{\tau^{m_1}}) + (1-\alpha)P_{\tau^{m_1}}) \in \Lambda^{n\times m}) \tag{6.8}$$

and, as usual, the Defender's perspective is adopted, so that $\mathscr{D}$ aims at maximising u while $\mathscr{A}$ wants to minimize it.

6.3.3 The DG-CTRa Game with Targeted Corruption (DG-CTRat)

The DG-CTRa game is difficult to solve directly, because of the 2-step attack strategy. We will work around this difficulty by tackling first with a slightly different version of the game, namely the detection game with targeted corruption of the training sequence, DG-CTRat, depicted in Figure 6.2. According to the taxonomy introduced in [15], the attack in this case is a *targeted attack*. While the strategies available to the Defender remain the same, for the Attacker, the choice of $Q(\cdot)$ is targeted to the counterfeiting of a given sequence y^n. In other words, we will assume that the Attacker corrupts the training sequence τ^{m_1} to ease the counterfeiting of a specific sequence y^n rather than to increase the probability that the second part of the attack succeeds. This means that the part of the attack aiming at corrupting the training sequence also depends on y^n, that is:

$$\mathcal{S}_{\mathscr{A},T} = \{Q(P_{\tau^{m_1}}, P_{y^n})\colon \mathcal{P}_{m_1} \times \mathcal{P}_n \to \mathcal{P}_{m_2}\}. \tag{6.9}$$

Even if this setup is not very realistic and more favorable to $\mathscr{A}$, who can exploit the exact knowledge of y^n (rather than its statistical properties) also for the corruption of the training sequence, in the next section we will show that, at least for large n, the DG-CTRat game is equivalent to the non-targeted version of the game we are interested in.

With the above ideas in mind, the DG-CTRat game is formally defined as follows.

Definition 6.1. The DG-CTRat $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game is a zero-sum, strategic, game played by $\mathscr{D}$ and $\mathscr{A}$, defined by the following strategies and payoff.

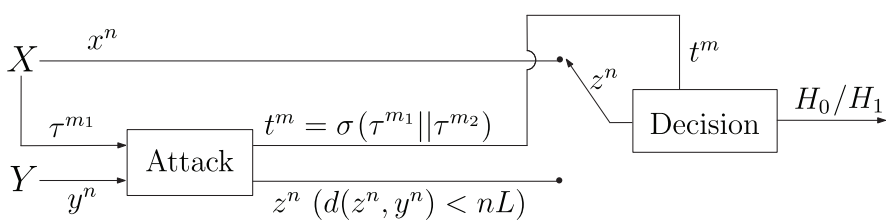


Figure 6.2: DG-CTRa game with targeted corruption of the training sequence (DG-CTRat). The difference with Figure 6.1 is that now τ^{m_2} may also depend on y^n.

- Defender's strategies.

$$\mathcal{S}_{\mathscr{D}} = \left\{ \Lambda^{n\times m} \subset \mathcal{P}_n \times \mathcal{P}_m \colon \max_{P_X} \max_{Q(\cdot,\cdot)\in\mathcal{S}_{\mathscr{A},T}} P_{\mathrm{FP}} \leq 2^{-\lambda n} \right\}, \quad (6.10)$$

  where $P_{\mathrm{FP}} = P_{XY}((z^n, \tau^{m_1}) \colon (P_{z^n}, \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha) P_{\tau^{m_1}}) \notin \Lambda^{n\times m})$.

- Attacker's strategies.

$$\mathcal{S}_{\mathscr{A}} = \mathcal{S}_{\mathscr{A},T} \times \mathcal{S}_{\mathscr{A},O}, \quad (6.11)$$

  with $\mathcal{S}_{\mathscr{A},T}$ and $\mathcal{S}_{\mathscr{A},O}$ defined as in (6.9) and (6.5) respectively.

- The payoff function. The payoff of the game is equal to the false negative error probability:

$$u(\Lambda^{n\times m}, (Q(\cdot,\cdot),\ S^n_{YZ}(\cdot,\cdot))) = -P_{\mathrm{FN}}, \quad (6.12)$$

  where P_{FN} is defined as in (6.8) with the difference that now $Q(\,)$ depends also on y^n.

An important observation regards the assumption that $\mathscr{D}$ knows α, that is the (maximum) percentage of training samples that $\mathscr{A}$ may corrupt. This is an implicit and necessary assumption in the definition of the game, since, for a proper definition, it is necessary that the players know the space of strategies of the other player. Assuming that the value of α is not known to the Defender would require that we redefine the game as a game with *incomplete information*, namely a Bayesian game. In our case, the Bayesian formulation of the game would dramatically complicate the analysis of the problem, so we decided to stick to a classical definition and interpret the value of α as a kind of worst case estimate that the Defender has on the capability of $\mathscr{A}$ to corrupt the training data. As a matter of fact, in the Neyman–Pearson setup adopted here, some estimate of the maximum percentage of samples corrupted by the Attacker is necessary, since in the absence of such an estimate the constraint on the false positive error probability could not be satisfied, given that the possibility that all the training samples have been corrupted could not be ruled out.

6.4 Solution of the DG-CTRat and DG-CTRa Games

In this section, we first derive the equilibrium solution of the DG-CTRat and DG-CTRa games and then evaluate the payoff at the equilibrium. As in the other cases, we focus on the behavior of the game when the lengths of the test and the training sequence tend to infinity.

6.4.1 Optimal Defender's Strategy

We start by deriving the asymptotically optimal strategy for $\mathscr{D}$. As for the games studied in the previous sections, a dominant and universal strategy with respect to P_Y exists for $\mathscr{D}$; hence, the optimum choice of $\mathscr{D}$ depends neither on the strategy chosen by the Attacker to corrupt the training and test sequences, nor on P_Y. In addition, since the constraint on the false positive probability must be satisfied for all Attackers' strategies, the optimal strategy for the Defender is the same for both the targeted and non-targeted versions of the game.

As a first thing, we need to search for an explicit expression of the false positive error probability. Such a probability depends on P_X and on the strategy used by $\mathscr{A}$ to corrupt the training sequence. In fact, the mapping of y^n into z^n does not have any impact on $\mathscr{D}$'s decision under H_0. We carry out our derivations by focusing on the game with targeted corruption. It will be clear from our analysis that the dependence on y^n has no impact on P_{FP}, and hence the same results hold for the game with non-targeted corruption.

For a given P_X and $Q(\cdot,\cdot)$, P_{FP} is equal to the probability that Y generates a sequence y^n and X generates two sequences x^n and τ^{m_1}, such that the pair of type classes $(P_{x^n}, \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha)P_{\tau^{m_1}})$ falls outside $\Lambda^{n\times m}$. Such a probability can be expressed as:

$$\begin{aligned} P_{\text{FP}} &= P((P_{x^n}, \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha)P_{\tau^{m_1}}) \in \bar{\Lambda}^{n\times m}) \\ &= \sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \\ &\quad \cdot \sum_{(P_{x^n},P_{tm})\in\bar{\Lambda}^{n\times m}} P_X(T(P_{x^n})) \cdot \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}_{m_1}: \\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{tm}}} P_X(T(P_{\tau^{m_1}})), \end{aligned} \tag{6.13}$$

where $\bar{\Lambda}^{n\times m}$ is the complement of $\Lambda^{n\times m}$, and where we have exploited the fact that under H_0 the training sequence τ^{m_1} and the test sequence x^n are generated independently. Given the above formulation, the set of strategies available to $\mathscr{D}$ can be rewritten as:

$$\mathcal{S}_{\mathscr{D}} = \Bigg\{ \Lambda^{n\times m}\colon \max_{P_X} \max_{Q(\cdot,\cdot)} \sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \cdot \sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} P_X(T(P_{x^n})) \cdot \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}_{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}})) \le 2^{-\lambda n} \Bigg\}. \tag{6.14}$$

We are now ready to prove the following lemma, which provides the asymptotically optimal strategy for the Defender for both versions of the game.

Lemma 6.1. Let $\Lambda^{n\times m,*}$ be defined as follows:

$$\Lambda^{n\times m,*} = \left\{ (P_{z^n}, P_{t^m})\colon \min_{Q\in\mathcal{P}_{m_2}} h\left(P_{z^n}, \frac{P_{t^m}-\alpha Q}{1-\alpha}\right) \le \lambda - \delta_n \right\} \tag{6.15}$$

with

$$\delta_n = |\mathcal{X}| \frac{\log(n+1)((1-\alpha)nc+1)}{n}, \tag{6.16}$$

where $|\mathcal{X}|$ is the cardinality of the source alphabet, $c = \frac{m}{n}$, and where the minimization over Q is limited to all the Q's such that $P_{t^m} - \alpha Q$ is nonnegative for all the symbols in $\mathcal{X}$. Then:

1. $\max_{P_X} \max_{s\in\mathcal{S}_{\mathscr{A}}} P_{\text{FP}} \le 2^{-n(\lambda-\nu_n)}$, with $\lim_{n\to\infty} \nu_n = 0$,
2. $\forall \Lambda^{n\times m} \in \mathcal{S}_{\mathscr{D}}$, we have $\bar{\Lambda}^{n\times m} \subseteq \bar{\Lambda}^{n\times m,*}$,

where ν^n is an arbitrary sequence approaching 0 when n tends to infinity.

Proof. To prove the first part of the lemma, we observe that from the false positive error probability given by (6.13), we can write:

$$\max_{P_X}\max_{Q(\cdot,\cdot)} P_{\mathrm{FP}} \tag{6.17}$$

$$\leq \max_{P_X} \sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \cdot \sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))$$

$$\cdot \max_{Q(\cdot,\cdot)} \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}_{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}})). \tag{6.18}$$

Let us consider the term within the inner summation. For each $P_{\tau^{m_1}}$ such that $\alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha)P_{\tau^{m_1}} = P_{t^m}$, we have:[4]

$$P_X(T(P_{\tau^{m_1}})) \leq \max_{Q\in\mathcal{P}_{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right), \tag{6.19}$$

with the understanding that the maximization is carried out only over the Q's such that $P_{t^m} - \alpha Q$ is nonnegative for all the symbols in $\mathcal{X}$.

Thanks to the above observation, we can upper bound the false positive error probability as follows:

$$\begin{aligned}
&\max_{P_X}\max_{Q(\cdot,\cdot)} P_{\mathrm{FP}}\\
&\leq \max_{P_X} \sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n}))\\
&\quad\cdot \sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))\cdot|\mathcal{P}_{m_1}|\cdot\max_{Q\in\mathcal{P}_{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)\\
&\overset{(a)}{=} \max_{P_X} \sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(T(P_{x^n}))|\mathcal{P}_{m_1}|\max_{Q\in\mathcal{P}_{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)\\
&\leq |\mathcal{P}_{m_1}| \sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} \max_{Q\in\mathcal{P}_{m_2}}\max_{P_X} P_X(T(P_{x^n}))P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right),
\end{aligned} \tag{6.20}$$

[4] It is easy to see that the bound (6.19) holds also for the non-targeted game, when Q depends on the training sequence only ($Q(P_{\tau^{m_1}})$).

where in (a) we exploited the fact that the second summation does not depend on P_{y^n}.

From this point, the proof goes along the same line of the proof of Lemma 5.2 in Chapter 5, by observing that $\max_{P_X} P_X(T(P_{x^n}))$ $P_X(T(\frac{P_{t^m}-\alpha Q}{1-\alpha}))$ is upper bounded by $2^{-nh(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha})}$, and that for each pair of types in $\bar{\Lambda}^{n\times m,*}$, the quantity $h(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha})$ is larger than $\lambda-\delta_n$ for every Q by the very definition of $\Lambda^{n\times m,*}$.

We now pass to the second part of the lemma. Let $\Lambda^{n\times m}$ be a strategy in $\mathcal{S}_{\mathscr{D}}$, and let (P_{x^n},P_{t^m}) be a pair of types contained in $\bar{\Lambda}^{n\times m}$. Given that $\Lambda^{n\times m}$ is an admissible decision region (see (6.10)), the probability that X emits a test sequence belonging to $T(P_{x^n})$ and a training sequence τ^{m_1} such that after the attack $(\tau^{m_1}\|\tau^{m_2})\in T(P_{t^m})$ is lower than $2^{-\lambda n}$ for all P_X and all possible attack strategies. In formula, we have

$$
\begin{aligned}
2^{-\lambda n} &> \max_{P_X}\max_{Q(\cdot,\cdot)}\sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \\
&\quad\cdot\left[P_X(T(P_{x^n}))\cdot\sum_{\substack{P_{\tau^{m_1}}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}}))\right] \\
&\overset{(a)}{=} \max_{P_X}\sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \\
&\quad\cdot\left[P_X(T(P_{x^n}))\cdot\max_{Q(\cdot,P_{y^n})}\sum_{\substack{P_{\tau^{m_1}}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(T(P_{\tau^{m_1}}))\right] \\
&\overset{(b)}{\geq} \max_{P_X}\sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n}))\cdot\Bigg[P_X(T(P_{x^n})) \\
&\quad\cdot\max_{Q(P_{\tau^{m_1}},P_{y^n})} P_X\left(T\left(\frac{P_{t^m}-\alpha Q(P_{\tau^{m_1}},P_{y^n})}{1-\alpha}\right)\right)\Bigg] \\
&\overset{(c)}{=} \max_{P_X} P_X(T(P_{x^n}))\max_{Q\in\mathcal{P}_{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right), \qquad (6.21)
\end{aligned}
$$

where (a) is obtained by replacing the maximization over all possible strategies $Q(\cdot,\cdot)$, with a maximization over $Q(\cdot, P_{y^n})$ for each specific P_{y^n}, and (b) is obtained by considering only one term $P_{\tau^{m_1}}$ of the inner summation and optimising $Q(P_{\tau^{m_1}}, P_{y^n})$ for that term. Finally, (c) follows by observing that the optimum $Q(\cdot, P_{y^n})$ is the same for all P_{y^n}. As usual, the maximization over Q in the last expression is restricted to the Q's for which $P_{t^m} - \alpha Q \geq 0$ for all the symbols in $\mathcal{X}$.[5]

By lower bounding the probability that a memoryless source X generates a sequence belonging to a certain type class, we can continue the above chain of inequalities as follows:

$$\begin{aligned} 2^{-\lambda n} &> \frac{\max_{P_X} \max_{Q\in\mathcal{P}_{m_2}} 2^{-n\left[\mathcal{D}(P_{x^n}\|P_X)+\frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\|P_X\right)\right]}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}} \\ &\geq \frac{2^{-n\min_{Q\in\mathcal{P}_{m_2}}\min_{P_X}\left[\mathcal{D}(P_{x^n}\|P_X)+\frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\|P_X\right)\right]}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}} \\ &\overset{(a)}{=} \frac{2^{-n\min_{Q\in\mathcal{P}_{m_2}} h\left(P_{x^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}}, \end{aligned} \tag{6.22}$$

where (a) derives from the minimization properties of the generalized log-likelihood ratio function $h(\,)$ (see Lemma 5.1). By taking the log of both terms we have:

$$\min_{Q\in\mathcal{P}_{m_2}} h\left(P_{x^n}, \frac{P_{t^m}-\alpha Q}{1-\alpha}\right) > \lambda - \delta_n, \tag{6.23}$$

thus completing the proof of the lemma. □

Lemma 6.1 shows that the strategy $\Lambda^{n\times m,*}$ is asymptotically admissible (point 1) and optimal (point 2), regardless of the attack; hence, it is a dominant strategy for $\mathscr{D}$. In addition, the optimal strategy is semi-universal, since it depends on P_X but not on P_Y.

At first sight, the minimization required by the optimal Defender's strategy seems computationally prohibitive, however this is not the case since the minimization can be carried out efficiently by exploiting the

[5] It is easy to see that the same lower bound can be derived also for the non-targeted case, as the optimum Q in the second to last expression does not depend on P_{y^n}.

convexity of the h function. To be more specific, since the minimisation is limited to the Q's such that $P_{t^m} - \alpha Q$ is nonnegative for all the symbols in $\mathcal{X}$, that is within the set $\{Q \in \mathcal{P}_{m_2}\colon \frac{P_{t^m} - \alpha Q}{1-\alpha} \in \mathcal{P}_{m_1}\}$, the log-sum inequality [58, Chapter 16, p. 483] can be invoked to show that $h\left(P_{x^n}, \frac{P_{t^m} - \alpha Q}{1-\alpha}\right)$ is a convex function with respect to Q. Being this set of Q's bounded (corresponding to a subset of the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$), the optimization problem in (6.15) is a *convex* MINLP [70], for which a global optimal solution exists. As we already said at the end of Section 3.2, for this kind of problems there are several efficient solvers yielding the optimal solution [72]. The number of optimization variables, which determines the computational complexity, corresponds to the cardinality of the alphabet, i.e., $|\mathcal{X}|$, and hence the minimization is viable in many practical scenarios.

From the proof of Lemma 6.1, it is clear that the same optimal strategy holds for the targeted and non-targeted versions of the game. The situation is rather different with regard to the optimal strategy for the Attacker. Despite the existence of a dominant strategy for the Defender, in fact, the identification of the optimal Attacker's strategy for the DG-CTRa game is not easy due to the 2-step nature of the attack. In the following sections, we will focus on the targeted version of the game, which is easier to study. We will then use the results obtained for the DG-CTRat game to derive the best achievable performance for the case of non-targeted attack.

6.4.2 The DG-CTRat Game: Optimal Attacker's Strategy and Equilibrium Point

Given the dominant strategy of $\mathscr{D}$, for any given τ^{m_1} and y^n, the optimal Attacker's strategy for the DG-CTRat game boils down to the following double minimization:

$$\begin{aligned}&(Q^*(P_{\tau^{m_1}}, P_{y^n}),\ S_{YZ}^{n,*}(P_{y_n}, P_{t^m}))\\ &\quad = \arg \min_{\substack{Q \in \mathcal{P}_{m_2} \\ S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})}} \left(\min_{Q'} \ h\left(P_{z^n}, \frac{(1-\alpha)P_{\tau^{m_1}} + \alpha Q - \alpha Q'}{1-\alpha}\right)\right),\end{aligned} \tag{6.24}$$

where P_{z^n} is obtained by applying the transformation map S^n_{YZ} to P_{y^n}, and where $P_{t^m} = (1-\alpha)P_{\tau^{m_1}} + \alpha Q$. As usual, the minimization over Q' is limited to the Q' such that all the entries of the resulting pmf are nonnegative.

Remark 6.1. Under corruption of the training sequence only ($L = 0$), the optimal attack strategy for the DG-CTRat game is

$$Q^*(P_{\tau^{m_1}}, P_{y^n}) = \arg \min_{Q \in \mathcal{P}_{m_2}} \left[\min_{Q'} \; h\left(P_{y^n}, P_{\tau^{m_1}} + \frac{\alpha}{1-\alpha}(Q - Q') \right) \right], \tag{6.25}$$

while, in the game setup without corruption of the training sequence ($\alpha = 0$) we have

$$S^{n,*}_{YZ}(P_{y^n}, P_{t^m}) = \underset{S^n_{YZ} \in \mathcal{A}^n(L, P_{y^n})}{\arg\min} \; h(P_{z^n}, P_{t^m}), \tag{6.26}$$

falling back to the known case of detection with uncorrupted training, already studied in Chapter 5.

Having determined the optimal strategies of both players, it is immediate to state the following:

Theorem 6.2 (Equilibrium Point of the DG-CTRat Game)**.** The DG-CTRat game is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile $(\Lambda^{n\times m,*}, (Q^*(\cdot,\cdot),\ S^{n,*}_{YZ}(\cdot,\cdot)))$ given by Equations (6.15) and (6.24).

Proof. The theorem is an immediate consequence of the fact that $\Lambda^{n\times m,*}$ is a dominant strategy for $\mathscr{D}$. □

6.4.3 The DG-CTRat Game: Payoff at the Equilibrium

We now derive the asymptotic value of the payoff at the equilibrium, to see who and under which conditions is going to *win* the game in the DG-CTRat setup.

To start with, we identify the set of pairs $(P_{y^n}, P_{\tau^{m_1}})$ for which, as a consequence of $\mathscr{A}$'s action, $\mathscr{D}$ accepts H_0, that is

$$\begin{aligned}\Gamma^n(\lambda,\alpha,L) = \{(P_{y^n}, P_{\tau^{m_1}})\colon \exists\ (P_{z^n}, P_{t^m}) \in \Lambda^{n\times m,*} \\ \text{s.t. } P_{t^m} = (1-\alpha)P_{\tau^{m_1}} + \alpha Q \text{ and } P_{z^n} = S_Z^n \\ \text{for some } Q \in \mathcal{P}_{m_2} \text{ and } S_{YZ}^n \in \mathcal{A}(L, P_{y^n})\}. \end{aligned} \tag{6.27}$$

If we fix the type of the non-corrupted training sequence $(P_{\tau^{m_1}})$, we obtain:

$$\begin{aligned}\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L) = \{P_{y^n}\colon \exists\ P_{z^n} \in \Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}} + \alpha Q) \\ \text{s.t. } P_{z^n} = S_Z^n \\ \text{for some } Q \in \mathcal{P}_{m_2} \text{ and } S_{YZ}^n \in \mathcal{A}(L, P_{y^n})\}, \end{aligned} \tag{6.28}$$

where $\Lambda^{n,*}(P)$ denotes the acceptance region for a fixed type P of the training sequence in $\mathcal{P}_m$. It is interesting to notice that, since in the current setting $\mathscr{A}$ has two degrees of freedom, the attack has a twofold effect: the sequence y^n is modified in order to bring it inside the acceptance region $\Lambda^{n,*}(P_{t^m})$ and the acceptance region itself is modified so to facilitate the former action.

To go on, we find it convenient to rewrite the set $\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L)$ as follows:

$$\begin{aligned}&\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L) \\ &\quad = \{P_{y^n}\colon \exists\ S_{PV}^n \in \mathcal{A}(L, P_{y^n}) \text{ s.t. } S_V^n \in \Gamma_0^n(P_{\tau^{m_1}},\lambda,\alpha)\}, \end{aligned} \tag{6.29}$$

where

$$\begin{aligned}&\Gamma_0^n(P_{\tau^{m_1}},\lambda,\alpha) \\ &= \{P_{y^n}\colon \exists Q \in \mathcal{P}_{m_2} \text{ s.t. } P_{y^n} \in \Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}} + \alpha Q)\}, \end{aligned} \tag{6.30}$$

is the set containing all the test sequences (or, equivalently, test types) for which it is possible to corrupt the training set in such a way that they fall within the acceptance region. As the subscript 0 suggests, this set corresponds to the set in (6.28) when $\mathscr{A}$ cannot modify the sequence drawn from Y, i.e., $L = 0$, and then tries to hamper the decision by corrupting the training sequence only.

By considering the expression of the acceptance region, the set $\Gamma_0^n(P_{\tau^{m_1}}, \lambda, \alpha)$ can be expressed in a more explicit form as follows:

$$\Gamma_0^n(P_{\tau^{m_1}}, \lambda, \alpha) = \Big\{ P_{y^n} \colon \exists Q, Q' \in \mathcal{P}_{m_2} \text{ s.t. } h\Big(P_{y^n}, P_{\tau^{m_1}} + \frac{\alpha}{(1-\alpha)}(Q - Q') \Big) \leq \lambda - \delta_n \Big\}, \tag{6.31}$$

where the second argument of $h(\,)$ denotes a type in $\mathcal{P}_{m_1}$ obtained from the original training sequence τ^{m_1} by first adding m_2 samples and later removing (in a possibly different way) the same number of samples. Note that in this formulation Q accounts for the fake samples introduced by the Attacker and Q' for the worst case *guess*, made by the Defender, of the position of the corrupted samples. We also observe that since we are considering the DG-CTRat version, in general Q will depend on P_{y^n}. As usual, we implicitly assume that Q and Q' are chosen in such a way that $P_{\tau^{m_1}} + \frac{\alpha}{(1-\alpha)}(Q - Q')$ is nonnegative and smaller than or equal to 1 for all the alphabet symbols.

We are now ready to derive the asymptotic payoff of the game by following similar steps to that used in Sections 3.3 and 5.3 for the game with known sources and with training data, respectively. First of all, we generalize the definition of the sets $\Lambda^{n\times m,*}$, Γ^n and Γ_0^n so that they can be evaluated for a generic pmf in $\mathcal{P}$ (that is, without requiring that the pmf's are induced by sequences of finite length). This step passes through the generalization of the h function. Specifically, given any pair of pmf's $(P, P') \in \mathcal{P} \times \mathcal{P}$, we define the generalized h function as:

$$\begin{aligned} h_c(P, P') &= \mathcal{D}(P\|U) + c\mathcal{D}(P'\|U); \\ U &= \frac{1}{1+c}P + \frac{c}{1+c}P', \end{aligned} \tag{6.32}$$

where $c \in [0, 1]$. Note that when $(P, P') \in \mathcal{P}_n \times \mathcal{P}_n$, $h_c(P, P') = h(P, P')$. The asymptotic version of $\Lambda^{n\times m,*}$ is:

$$\Lambda^* = \Big\{ (P, R) \colon \min_Q \; h_c\Big(P, \; \frac{R - \alpha Q}{1-\alpha} \Big) \leq \lambda \Big\}. \tag{6.33}$$

In a similar way, we can derive the asymptotic versions of Γ^n and Γ_0^n in (6.29) and (6.30)–(6.31). To do so, we first observe that the transportation map S_{YZ}^n depends on the sources only through their pmf's.

By denoting with S^n_{PV} a transportation map from a pmf $P \in \mathcal{P}_n$ to another pmf $V \in \mathcal{P}_n$ and rewriting set Γ^n accordingly, we can easily derive the asymptotic version of Γ^n as follows:

$$\Gamma(R, \lambda, \alpha, L) = \{P \in \mathcal{P}\colon \exists S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Gamma_0(R, \lambda, \alpha)\}, \tag{6.34}$$

with

$$\begin{aligned}
&\Gamma_0(R, \lambda, \alpha) \\
&\quad = \{P \in \mathcal{P}\colon \exists\ Q \in \mathcal{P} \text{ s.t. } P \in \Lambda^*((1-\alpha)R + \alpha Q)\} \\
&\quad = \left\{P \in \mathcal{P}\colon \exists Q, Q' \in \mathcal{P} \text{ s.t. } h_c\left(P,\ R + \frac{\alpha}{(1-\alpha)}(Q - Q')\right) \le \lambda\right\},
\end{aligned} \tag{6.35}$$

where the definitions of S_{PV} and $\mathcal{A}(L, P)$ derive from those of S^n_{PV} and $\mathcal{A}^n(L, P)$ by relaxing the requirement that the terms $S_{PV}(i, j)$ and $P(i)$ are rational numbers with denominator n.

Given the above, we can prove the following theorem.

Theorem 6.3 (Asymptotic Payoff of the DG-CTRat Game)**.** The false negative error exponent of the DG-CTRat game at the equilibrium is given by

$$\varepsilon^* = \min_R \Big[(1-\alpha)c\mathcal{D}(R\|P_X) + \min_{P \in \Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)\Big]. \tag{6.36}$$

Accordingly,

1. if $P_Y \in \Gamma(P_X, \lambda, \alpha, L)$ then $\varepsilon^* = 0$;
2. if $P_Y \notin \Gamma(P_X, \lambda, \alpha, L)$ then $\varepsilon^* > 0$.

Proof. The theorem can be proven going along the same lines of the proof of Theorems 3.4 and 5.4, i.e., by applying the generalized version of Sanov's theorem (see Section 2.4.2). In particular, let us consider

$$\begin{aligned}
P_{\text{FN}} &= \sum_{(P_{y^n}, P_{\tau^{m_1}}) \in \Gamma^n(\lambda,\alpha,L)} P_X(T(P_{\tau^{m_1}}))P_Y(T(P_{y^n})) \\
&= \sum_{R \in \mathcal{P}_{m_1}} P_X(T(R)) \sum_{P \in \Gamma^n(R,\lambda,\alpha,L)} P_Y(T(P)) \\
&= \sum_{R \in \mathcal{P}_{m_1}} P_X(T(R))P_Y(\Gamma^n(R, \lambda, \alpha, L)).
\end{aligned} \tag{6.37}$$

We start by deriving an upper bound of the false negative error probability:

$$\begin{aligned}
P_{\mathrm{FN}} &\leq \sum_{R\in\mathcal{P}_{m_1}} P_X(T(R)) \sum_{P\in\Gamma^n(R,\lambda,\alpha,L)} 2^{-n\mathcal{D}(P\|P_Y)} \\
&\leq \sum_{R\in\mathcal{P}_{m_1}} P_X(T(R))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma^n(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)} \\
&\leq \sum_{R\in\mathcal{P}_{m_1}} P_X(T(R))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)} \\
&\leq (n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|} \\
&\quad \cdot 2^{-n \min\limits_{R\in\mathcal{P}_{m_1}} \left[\frac{m_1}{n}\mathcal{D}(R\|P_X) + \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)\right]} \\
&\leq (n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|} \\
&\quad \cdot 2^{-n \min\limits_{R\in\mathcal{P}} \left[(1-\alpha)c\mathcal{D}(R\|P_X) + \min\limits_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)\right]},
\end{aligned} \tag{6.38}$$

where the use of the minimum instead of the infimum is justified by the fact that $\Gamma^n(R,\lambda,\alpha,L)$ and $\Gamma(R,\lambda,\alpha,L)$ are compact sets. By taking the log and dividing by n we find:

$$\begin{aligned}
&-\frac{\log P_{\mathrm{FN}}}{n} \\
&\geq \min_{R\in\mathcal{P}} \left[(1-\alpha)c\mathcal{D}(R\|P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P\|P_Y)\right] - \beta_n,
\end{aligned} \tag{6.39}$$

where $\beta_n = |\mathcal{X}|\frac{\log(n+1)((1-\alpha)nc+1)}{n}$ tends to zero when n tends to infinity.

We now turn to the analysis of a lower bound for P_{FN}. Let R^* be the pmf achieving the minimum of the outer minimization of (6.36). Due to the density of rational numbers within real numbers, we can find a sequence of pmf's $R_{m_1} \in \mathcal{P}_{m_1}$ that tends to R^* when n (and hence m_1) tends to infinity. Then, we can write:

$$\begin{aligned}
P_{\mathrm{FN}} &= \sum_{R\in\mathcal{P}_{m_1}} P_X(T(R)) P_Y(\Gamma^n(R,\lambda,\alpha,L)) \\
&\geq P_X(T(R_{m_1})) P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)), \\
&\geq \frac{2^{-m_1\mathcal{D}(R_{m_1}\|P_X)}}{(m_1+1)^{|\mathcal{X}|}} P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)),
\end{aligned} \tag{6.40}$$

where, in the first inequality, we have replaced the sum with the single element of the subsequence R_{m_1} defined previously, and where the second inequality derives again from the well known lower bound on the probability of a type class. From (5.38), by taking the log and dividing by n, we obtain:

$$-\frac{\log P_{\mathrm{FN}}}{n} \leq (1-\alpha)c\mathcal{D}(R_{m_1}\|P_X) - \frac{1}{n}\log P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)) + \beta'_n, \quad (6.41)$$

where $\beta'_n = |\mathcal{X}|\frac{\log(m_1+1)}{n}$ tends to zero when n tends to infinity.

To assess the behavior of $P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L))$ as n tends to infinity, we resort to Corollary 2.2 of the generalized version of Sanov's theorem proven in Section 2.4.2. In order to apply the corollary, we must prove the Hausdorff convergence of $\Gamma^n(R_{m_1},\lambda,\alpha,L)$ to $\Gamma(R^*,\lambda,\alpha,L)$. First of all, we observe that by exploiting the continuity of the h_c function and the density of rational numbers into the real ones, it is easy to prove that $\Gamma_0^n(R_{m_1},\lambda,\alpha) \xrightarrow{H} \Gamma_0(R^*,\lambda,\alpha)$. Then the Hausdorff convergence of $\Gamma^n(R_{m_1},\lambda,\alpha,L)$ to $\Gamma(R^*,\lambda,\alpha,L)$ follows from the regularity properties of the set of the transportation maps stated in Appendix A. To see how, we observe that any transformation $S_{PV} \in \mathcal{A}(L,P)$ mapping P into V can be applied in the reverse direction through the transformation $S_{VP}(i,j) = S_{PV}(j,i)$. It is immediate to see that S_{VP} introduces the same distortion introduced by S_{PV}, that is $S_{VP} \in \mathcal{A}(L,V)$. Let now P be a point in $\Gamma(R^*,\lambda,\alpha,L)$. By definition, we can find a map $S_{PV} \in \mathcal{A}(L,P)$ such that $V \in \Gamma_0(R^*,\lambda,\alpha)$. Since $\Gamma_0^n(R_{m_1},\lambda,\alpha) \xrightarrow{H} \Gamma_0(R^*,\lambda,\alpha)$, for large enough n, we can find a point $V' \in \Gamma_0^n(R_{m_1},\lambda,\alpha)$ which is arbitrarily close to V. Thanks to the second part of Theorem A.2, we know that a map $S_{V'P'} \in \mathcal{A}^n(L,V')$ exists such that P' is arbitrarily close to P and $P' \in \mathcal{P}_n$. By applying the inverse map $S_{P'V'}$ to P', we see that $P' \in \Gamma^n(R_{m_1},\lambda,\alpha,L)$, thus permitting us to conclude that, when n increases, $\delta_{\Gamma(R^*,\lambda,\alpha,L)}(\Gamma^n(R_{m_1},\lambda,\alpha,L)) \to 0$. In a similar way, we can prove that $\delta_{\Gamma^n(R_{m_1},\lambda,\alpha,L)}(\Gamma(R^*,\lambda,\alpha,L)) \to 0$, hence permitting us to conclude that $\Gamma^n(R_{m_1},\lambda,\alpha,L) \xrightarrow{H} \Gamma(R^*,\lambda,\alpha,L)$.

We can now apply the generalized version of Sanov's theorem as expressed in Corollary 2.2, yielding

$$-\lim_{n\to\infty}\frac{1}{n}\log P_Y(\Gamma^n(R_{m_1},\lambda,\alpha,L)) = \min_{P\in\Gamma(R^*,\lambda,\alpha,L)}\mathcal{D}(P\|P_Y). \quad (6.42)$$

Going back to (6.41), and exploiting the continuity of the divergence function, we can conclude that for large n we have:

$$-\frac{\log P_{\mathrm{FN}}}{n} \le (1-\alpha)c\mathcal{D}(R^*\|P_X) + \min_{P\in\Gamma(R^*,\lambda,\alpha,L)}\mathcal{D}(P\|P_Y) + \nu_n, \quad (6.43)$$

where the sequence ν_n tends to zero when n tends to infinity. By coupling Equations (6.39) and (6.43) and by letting $n\to\infty$, we eventually obtain:

$$-\lim_{n\to\infty}\frac{\log P_{\mathrm{FN}}}{n} = \min_{R}\left[(1-\alpha)c\cdot\mathcal{D}(R\|P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P\|P_Y)\right], \quad (6.44)$$

thus proving the theorem. □

As an immediate consequence of Theorem 6.3, the set $\Gamma(P_X,\lambda,\alpha,L)$ defines the *indistinguishability region* of the test, that is the set of all the sources for which $\mathscr{A}$ induces $\mathscr{D}$ to decide in favor of H_0 even if H_1 holds. Moreover, by exploiting optimal transport theory, the indistinguishability region can be rewritten as:

$$\Gamma(P_X,\lambda,\alpha,L) = \{P\colon \exists\ V\in\Gamma_0(P_X,\lambda,\alpha) \text{ s.t. } EMD(P,V)\le L\}. \quad (6.45)$$

6.4.4 Analysis of the DG-CTRa Game

We now get back to the original DG-CTRa formulation. For a given choice of $Q(P_{\tau^{m_1}})\in\mathcal{S}_{\mathscr{A},T}$ (and hence t^m), given a sequence y^n, the optimal choice of the second part of the attack derives quite easily from the definition of $\Lambda^{n\times m,*}$, namely

$$S^{n,*}_{YZ}(P_{y^n},P_{t^m}) = \arg\min_{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})}\left(\min_{Q\in\mathcal{P}_{m_2}} h\left(P_{z^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right). \quad (6.46)$$

Now the point is to determine the strategy $Q(P_{\tau^{m_1}})$ which maximizes the probability that the attack in (6.46) succeeds. To this purpose, of course, the Attacker must exploit the knowledge of P_Y. Since solving such a maximization problem is not an easy task, we will proceed in a different way. We first introduce a simple (and possibly suboptimal) strategy, then we argue that such a strategy is asymptotically optimal, in that the set of the sources that cannot be distinguished from X with this choice is the same set that we have obtained for the DG-CTRat setup, which is known to be more favorable to the Attacker. More specifically, we consider the following attack.

- In the first part, $\mathscr{A}$ does not know y^n, hence he trusts the law of large numbers and optimizes $Q(P_{\tau^{m_1}})$ by using P_Y as a proxy for P_{y^n}. To do so, he applies (6.24), by replacing P_{y^n} with P_Y. Specifically, denoting by $Q^\dagger$ the resulting strategy for the first part of the attack, we have

$$Q^\dagger(P_{\tau^{m_1}}) = \arg \min_{Q \in \mathcal{P}_{m_2}} \tag{6.47}$$

$$\min_{\substack{Q' \in \mathcal{P}_{m_2} \\ S_{YZ} \in \mathcal{A}(L, P_Y)}} h_c\left(P_Z, P_{\tau^{m_1}} + \frac{\alpha}{1-\alpha}(Q - Q')\right). \tag{6.48}$$

  As a by-product of the above minimization, the Attacker also finds the map $S_{YZ}^{n,\dagger}$ representing the optimal attack when $P_{y^n} = P_Y$. Let us denote the result of the application of such a map to P_Y by $P_Z^\dagger$.

- In the second part of the attack, $\mathscr{A}$ tries to move P_{y^n} as close as possible to $P_Z^\dagger$, that is:

$$S_{YZ}^{n,\dagger}(P_{y^n}, P_{t^m}^\dagger) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} d(S_Z^n, P_Z^\dagger), \tag{6.49}$$

  where $S_{YZ}^{n,\dagger}(P_{y^n}, P_{t^m}^\dagger)$ depends upon the corrupted training sequence obtained after the application of the first part of the attack, namely $P_{t^m}^\dagger = (1-\alpha)P_{\tau^{m_1}} + \alpha Q^\dagger(P_{\tau^{m_1}})$, through $P_Z^\dagger$.

The asymptotic optimality of the strategy $(Q^\dagger(P_{\tau^{m_1}}), S_{YZ}^{n,\dagger}(P_{y^n}, P_{t^m}^\dagger))$ derives from the following theorem.

Theorem 6.4 (Indistinguishability Region of the DG-CTRa Game)**.** The indistinguishability region of DG-CTRa game is equal to that of the DG-CTRat game (see (6.34)) and is asymptotically achieved by the attack strategy $(Q^\dagger(P_{\tau^{m_1}}), S_{YZ}^{n,\dagger}(P_{y^n}, P_{t^m}^\dagger))$.

Proof (Sketch). The theorem derives from the observation that due to the law of large numbers, when n grows, P_{y^n} tends to P_Y; hence, for large enough n, optimising the first part of the attack by replacing P_{y^n} with P_Y does not introduce a significant performance loss. The rigorous proof goes along the same path of the proof of Theorem 6.3 and ultimately relies on the continuity of the h_c function and the regularity properties of the set $\mathcal{A}^n(L, P_{y^n})$. The details of the proof are omitted for sake of brevity. □

Given the asymptotic equivalence of the DG-CTRa and the DG-CTRat games, in the rest of the chapter, we will generally refer to the DG-CTRa game without specifying if we are considering the targeted or non-targeted case.

6.5 Source Distinguishability in the DG-CTRa Setup

We now study the behaviour of the DG-CTRa game when we decrease the decay rate of the false positive error probability λ (by decreasing λ, $\mathscr{D}$ can improve his payoff at the equilibrium) and derive the best achievable performance of the Defender, when it is required only that P_{FP} tends to zero exponentially fast with an arbitrarily small – yet strictly positive – error exponent, somehow extending the Chernoff-Stein lemma [58, Chapter 12.8] to the adversarial setup considered in this chapter. Afterwards, we will use such a result to derive the conditions under which a reliable distinction between two sources is possible as a function of the number of corrupted training samples α and maximum allowed distortion L.

6.5.1 Ultimate Achievable Performance of the DG-CTRa Game

The main result of this section is stated in the theorem below. Let $\Gamma(P_X, \alpha, L) = \Gamma(P_X, \lambda = 0, \alpha, L)$, that is,

$$\Gamma(P_X, \alpha, L) = \{P\colon \exists\ V \in \Gamma_0(P_X, \alpha) \text{ s.t. } EMD(P, V) \leq L\}, \quad (6.50)$$

where $\Gamma_0(P_X, \alpha) = \Gamma_0(P_X, \alpha, L = 0)$. As implied by the following theorem, $\Gamma(P_X, \alpha, L)$ is the ultimate indistinguishability region of the DG-CTRa game.

Theorem 6.5. Given two sources X and Y, a maximum allowed average per-letter distortion L and a fraction α of training samples provided by the Attacker, the maximum achievable false negative error exponent at the equilibrium for the DG-CTRa game is:

$$\begin{aligned}&\lim_{\lambda\to 0}\ \lim_{n\to\infty} -\frac{1}{n}\log\ P_{\text{FN}}\\ &\quad = \min_{R}\Big[(1-\alpha)c\mathcal{D}(R\|P_X) + \min_{P\in\Gamma(R,\alpha,L)}\mathcal{D}(P\|P_Y)\Big].\end{aligned} \quad (6.51)$$

Moreover, the ultimate indistinguishability region $\Gamma(P_X, \alpha, L)$ can be written as

$$\begin{aligned}\Gamma(P_X, \alpha, L) &= \left\{P\colon \min_{V\colon EMD(P,V)\leq L}\sum_i [V(i)-P_X(i)]^+ \leq \frac{\alpha}{(1-\alpha)}\right\}\\ &= \left\{P\colon \min_{V\colon EMD(P,V)\leq L} d_{L_1}(V, P_X) \leq \frac{2\alpha}{(1-\alpha)}\right\}.\end{aligned} \quad (6.52)$$

Proof. The proof of the first part goes along the same steps of the proof of Theorems 4.1 and 5.10 in Chapters 4 and 5 respectively, and is not repeated here. We show, instead, that $\Gamma(P_X, \alpha, L)$ can be rewritten as in (6.52).

By observing that $h_c(P, Q) = 0$ if and only if $P = Q$, it is immediate to see that the set $\Gamma_0(P_X, \lambda = 0, \alpha)$ takes the following expression:

$$\Gamma_0(P_X, \alpha) = \left\{P\colon \exists\ Q, Q' \in \mathcal{P} \text{ s.t. } P = P_X + \frac{\alpha}{(1-\alpha)}(Q - Q')\right\}. \quad (6.53)$$

Expression (6.53) can be rewritten by avoiding the introduction of the auxiliary pmf's Q and Q'. To do so, we observe that $Q(i)$ must be larger than $Q'(i)$ for all the bins i for which $P(i) > P_X(i)$ (and viceversa). In addition, Q and Q' must be valid pmf's, hence we have $\sum_i [Q(i) - Q'(i)]^+ = \sum_i [Q'(i) - Q(i)]^+ \leq 1$. Then, it is easy to see that (6.53) is equivalent to the following definition:

$$\begin{aligned}\Gamma_0(P_X, \alpha) &= \left\{ P\colon \sum_i [P(i) - P_X(i)]^+ \leq \frac{\alpha}{(1-\alpha)} \right\} \\ &= \left\{ P\colon d_{L_1}(P, P_X) \leq \frac{2\alpha}{(1-\alpha)} \right\}, \end{aligned} \tag{6.54}$$

where the second equality follows by observing that $d_{L_1}(P, P_X) = \sum_i [P(i) - P_X(i)]^+ + \sum_i [P_X(i) - P(i)]^+$. Eventually, (6.52) derives immediately from the expression of $\Gamma_0(P_X, \alpha)$ given in (6.54). □

According to Theorem 6.5, $\Gamma(P_X, \alpha, L)$ provides the *ultimate indistinguishability region* of the test, that is, the set of all the pmf's P_Y that can not be distinguished from P_X ensuring that the two types of error probabilities tend to zero exponentially fast with vanishingly small, yet positive, error exponents. These are the pmf's for which $\mathscr{A}$ wins the game (according to the meaning of "*winning the game*" stated at the end of Section 3.3).

Before going on, we pause to discuss the geometrical meaning of the set $\Gamma_0(P_X, \alpha)$ in (6.53). To do so, we introduce the set Λ_0^*, obtained from Λ^* by letting λ tends to zero:

$$\Lambda_0^* = \left\{ (P, P')\colon \exists\ Q \text{ s.t. } P' = \frac{P - \alpha Q}{(1-\alpha)} \right\}. \tag{6.55}$$

As usual, we can fix the pmf P and define:

$$\Lambda_0^*(P) = \left\{ P'\colon \exists\ Q \text{ s.t. } P' = \frac{P - \alpha Q}{(1-\alpha)} \right\}. \tag{6.56}$$

By referring to Figure 6.3 (left part), we can geometrically interpret $\Lambda_0^*(P)$ as the set of the pmf's P' such that P is a convex combination (with coefficient α) of P' with a point Q of the probability simplex. Starting from (6.35), we can then rewrite $\Gamma_0(P_X, \alpha)$ as follows:

$$\Gamma_0(P_X, \alpha) = \{P\colon \exists\ Q \in \mathcal{P} \text{ s.t. } P \in \Lambda_0^*((1-\alpha)P_X + \alpha Q)\}. \tag{6.57}$$

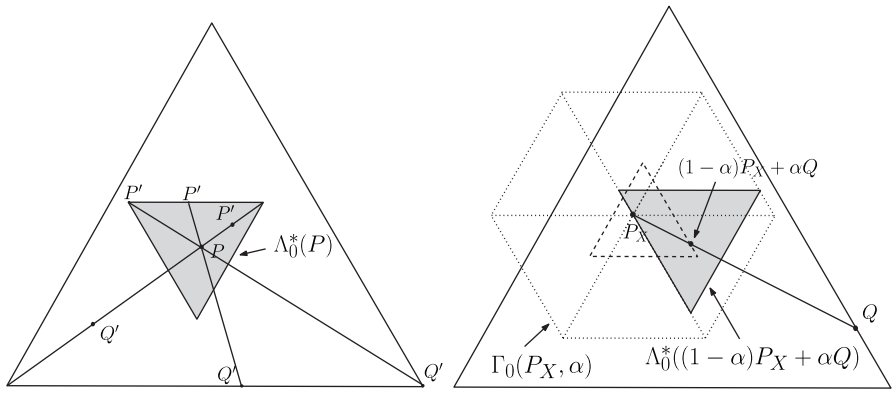


Figure 6.3: Geometrical representation of $\Lambda_0^*(P)$ (left) and construction of $\Gamma_0(P_X, \alpha)$ (right). The size of the sets are exaggerated for graphical purposes.

Accordingly, $\Gamma_0(P_X, \alpha)$ is geometrically obtained as the union of the acceptance regions built from the points that can be written as a convex combination of P_X with some point Q in the simplex. As shown in Figure 6.3 (right), such a region corresponds to a hexagon centered in P_X, which, in the probability simplex, is equivalent to the set of points whose L_1 distance from P_X is smaller than or equal to $2\alpha/(1-\alpha)$ (as stated in (6.54)). Of course, only the points of the hexagon that lie inside the simplex are valid pmf's and then must be accounted for.

A pictorial representation of the set $\Gamma(P_X, \alpha, L)$ is given in Figure 6.4.

6.5.2 Security Margin and Blinding Corruption Level

By closer inspection of the *ultimate indistinguishability region* we can derive some interesting parameters characterizing the distinguishability of two sources in adversarial setting.

Let $X \sim P_X$ and $Y \sim P_Y$ be two sources. We first focus on the case in which the Attacker can not modify the test sequence ($L = 0$). In this scenario, the ultimate indistinguishability region boils down to $\Gamma_0(P_X, \alpha)$. Then, $\mathscr{D}$ can tell the two sources apart if $d_{L_1}(P_Y, P_X) > \frac{2\alpha}{(1-\alpha)}$. On the contrary, if $d_{L_1}(P_Y, P_X) \leq \frac{2\alpha}{(1-\alpha)}$, $\mathscr{A}$ is able to make the sources indistinguishable by corrupting the training sequence. Clearly,

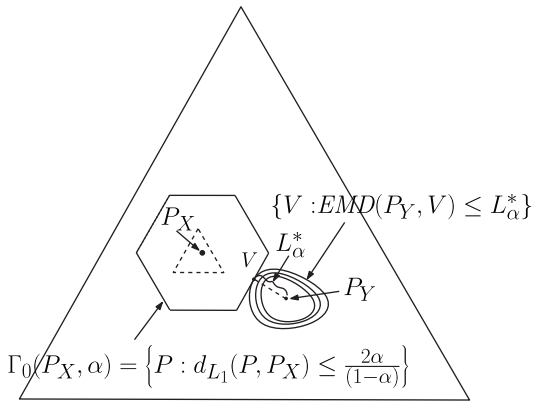


Figure 6.4: Geometrical representation of $\Gamma(P_X, \alpha, L)$ as stated in Theorem 6.5.

the larger the α the easier is for $\mathscr{A}$ to win the game. We can define the *blinding corruption level* α_b as the minimum value of α for which two sources X and Y can not be distinguished. Specifically, we have:

$$\alpha_b(P_X, P_Y) = \frac{d_{L_1}(P_Y, P_X)}{2 + d_{L_1}(P_Y, P_X)} = \frac{\sum_i [P_Y(i) - P_X(i)]^+}{1 + \sum_i [P_Y(i) - P_X(i)]^+}. \quad (6.58)$$

From (6.58), it is easy to see that α_b is always lower than $1/2$, with the limit case $\alpha_b = 1/2$ corresponding to a case in which P_X and P_Y have completely disjoint supports.[6] It is interesting to notice that α_b is symmetric with respect to the two sources. Since the Attacker is allowed only to add samples to the training sequence without removing existing samples, this might seem a counterintuitive result. Actually, the symmetry of α_b is a consequence of the worst case approach adopted by the Defender. In fact, $\mathscr{D}$ himself discards a subset of samples from the training sequence in such a way as to maximize the probability that the remaining part of the training sequence and the test sequence have been drawn from the same source.

Let us now consider the general case in which $L \neq 0$. For a given $\alpha < \alpha_b$, we look for the maximum distortion for which the two sources can be reliably distinguished. From Equation (6.52), we argue that the

[6] We remind that for any pair of pmf's (P, Q), $d_{L_1}(P, Q) \leq 2$.

attack does not succeed if the following condition holds:

$$\min_{V:\, EMD(P_Y,V)\leq L} d_{L_1}(V, P_X) > \frac{2\alpha}{(1-\alpha)}. \tag{6.59}$$

This leads to the extension of the concept of Security Margin, introduced in Chapter 4, to the more general setup considered in this chapter.

Definition 6.2 (Security Margin in the DG-CTRa Setup)**.** Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum distortion allowed to the Attacker for which the two sources can be reliably distinguished in the DG-CTRa setup with a fraction α of possibly corrupted samples is given by

$$\mathcal{SM}_\alpha(P_X, P_Y) = L^*_\alpha, \tag{6.60}$$

where $L^*_\alpha = 0$ if $P_Y \in \Gamma_0(P_X, \alpha)$, while, if $P_Y \notin \Gamma_0(P_X, \alpha)$, L^*_α is the quantity that satisfies

$$\min_{V:\, EMD(P_Y,V)\leq L^*_\alpha} d_{L_1}(V, P_X) = \frac{2\alpha}{(1-\alpha)}. \tag{6.61}$$

When $L > \mathcal{SM}_\alpha(P_X, P_Y)$, it is not possible for $\mathscr{D}$ to distinguish between the two sources with positive error exponents of the two kinds.

A geometric interpretation of L^*_α is given in Figure 6.5.

Remark 6.2. By focusing on the case $P_Y \notin \Gamma_0(P_X, \alpha)$, and by observing that

$$\min_{V:\, EMD(P_Y,V)\leq L} d_{L_1}(V, P_X) \tag{6.62}$$

is a monotonic non-increasing function of L, the Security Margin can be expressed in explicit form as

$$\mathcal{SM}_\alpha(P_X, P_Y) = \arg\min_{L'} \min_{V:\, EMD(P_Y,V)\leq L'} \left| d_{L_1}(V, P_X) - \frac{2\alpha}{(1-\alpha)} \right|. \tag{6.63}$$

By looking at the behavior of the Security Margin as a function of α, we see that $\mathcal{SM}_{\alpha_b}(P_X, P_Y) = 0$, meaning that, whenever the fraction of corrupted samples reaches the critical value α_b, the sources

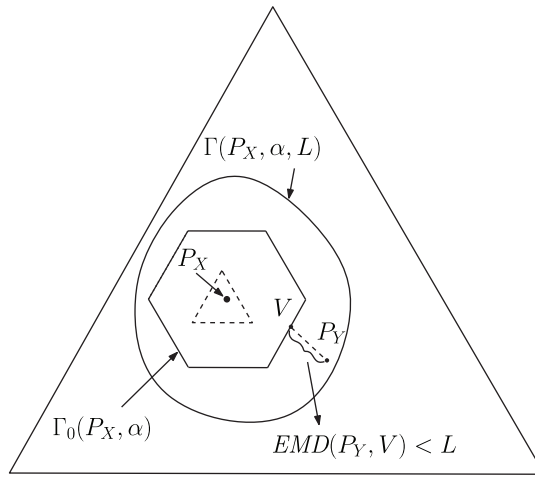


Figure 6.5: Geometrical interpretation of the Security Margin between two sources X ($\sim P_X$) and Y ($\sim P_Y$).

can not be distinguished even if the Attacker does not introduce any distortion. On the contrary, setting $\alpha = 0$ corresponds to studying the distinguishability of the sources with uncorrupted training; in this case we have $\mathcal{SM}_0(P_X, P_Y) = EMD(P_X, P_Y)$, in agreement with Definition 4.1. With reference to Figure 6.5, it is easy to see that when $\alpha = 0$ the hexagon representing $\Gamma_0(P_X, \alpha)$ collapses into the single point P_X and the Security Margin corresponds to the EMD between Y and X. Eventually, we notice that, for $\alpha > 0$, the value of the Security Margin in (6.63) is less than $EMD(P_X, P_Y)$. This is also an expected behavior since the general setting considered in this chapter is more favorable to the Attacker than the setting of Chapters 3 and 5.

We conclude our discussion by arguing that, as for the settings studied in the previous chapters, the Security Margin is symmetric with respect to the two sources X and Y, that is, $\mathcal{SM}_\alpha(P_Y, P_X) = \mathcal{SM}_\alpha(P_X, P_Y)$. To show that this is the case, we provide the following informal argument. By looking at (6.63), we observe that the pmf V' associated with the minimum L, for which we have $EMD(P_Y, V') = \mathcal{SM}_\alpha(P_X, P_Y)$, can be obtained through the application of a map $S_{P_Y V}$ that works as follows: it does not modify a portion $\alpha/(1-\alpha)$

of P_Y and moves the remaining mass into an equal amount of P_X in a convenient way (i.e., in such a way as to minimize the overall distance between the masses). The inverse map can be applied to bring the same quantity of mass from P_X to P_Y, while leaving as untouched the remaining mass, thus obtaining a V'' which satisfies $EMD(P_X, V'') = EMD(P_Y, V')$ (because of the symmetry of the per-symbol distortion d) and $d_{L_1}(V'', P_Y) = d_{L_1}(V', P_X) = 2\alpha/(1-\alpha)$. Arguably, V'' is the pmf for which $EMD(P_X, V'') = \mathcal{SM}_\alpha(P_Y, P_X)$; hence, $\mathcal{SM}_\alpha(P_Y, P_X) = \mathcal{SM}_\alpha(P_X, P_Y)$.

Bernoulli Sources

In order to get some insights on the practical meaning of α_b and $\mathcal{SM}_\alpha$, we consider the simple case of two Bernoulli sources with parameter $q = P_X(1)$ and $p = P_Y(1)$. Assuming that no distortion is allowed to the Attacker, the minimum fraction of samples that $\mathscr{A}$ must add to induce a decision error is, according to (6.58), $\alpha_b = \frac{|p-q|}{1+|p-q|}$. For instance, and rather obviously, when $|p-q| = 1$, to win the game $\mathscr{A}$ must introduce a number of fake samples equal to the number of samples of the correct training sequence, i.e., $\alpha = 0.5$. With regard to the Security Margin, we have:

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p| - \dfrac{\alpha}{1-\alpha} & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b. \end{cases} \tag{6.64}$$

Figure 6.6 illustrates the behavior of $\mathcal{SM}_\alpha(p,q)$ as a function of α when $p = 0.3$ and $q = 0.7$. The blinding corruption value is $\alpha_b = 0.286$. Obviously, when $\alpha = 0$, we obtain the same expression derived in Section 4.3 (Equation (4.10)).

6.6 The DG-CTRr Game

In this section, we study the second variant of the game with corrupted training, in which $\mathscr{A}$ observes the training sequence and can replace a selected fraction of samples.

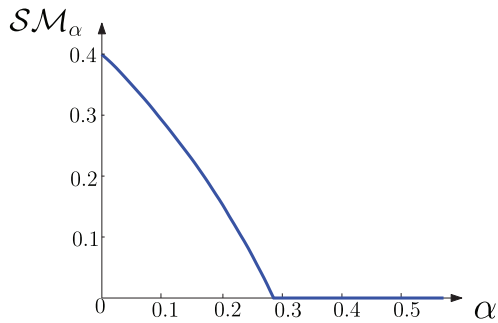


Figure 6.6: Security Margin as a function of α for Bernoulli sources with parameters $p = 0.3$ and $q = 0.7$ ($\alpha_b = 0.286$).

6.6.1 The Adversarial Setup

Let τ^m denote the original m-sample long training sequence drawn from X and let $\mathcal{M}$ be a subset of $m_2 = \alpha m$ indexes in $[1, 2, \ldots, m]$. Let $m_1 = m - m_2$. The Attacker can choose the index set $\mathcal{M}$ and replace the corresponding samples with m_2 fake samples. More formally, given the original training sequence τ^m, the training sequence observed by the Defender is $t^m = \sigma(\tau^{m_1}_{\bar{\mathcal{M}}} \| \tau^{m_2})$, where $\bar{\mathcal{M}}$ is the complement of $\mathcal{M}$ in $[1, 2 \ldots m]$, $\tau^{m_1}_{\bar{\mathcal{M}}}$ is the set of original (non-attacked) samples, and τ^{m_2} is the sequence with the fake samples introduced by the Attacker. The adversarial decision setup is illustrated in Figure 6.7 for the case of targeted attack, where the corruption of the training samples depends on the to-be-attacked sequence y^n. This is the version of the game we

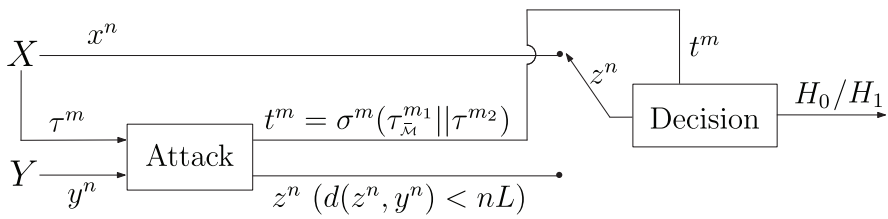


Figure 6.7: Schematic representation of the DG-CTRr game (with targeted corruption). Given the original training sequence τ^m, the adversary can replace a selected subset of m_2 training samples with fake ones.

focus on, the extension to the case of non-target attack being easily obtained by following the same approach used in Section 6.4.4.

Arguably, this scenario with replacement of the samples is more favorable to the Attacker with respect to the DG-CTRa setting.

6.6.2 Definition of the DG-CTRr Game

Below, we formally define the detection game with replacement of selected samples, namely the DG-CTRr $(\mathcal{S}_{\mathscr{D}}, \mathcal{S}_{\mathscr{A}}, u)$ game.

Defender's Strategies

As in the DG-CTRa game, in order to be sure that the false positive error probability is lower than $2^{-n\lambda}$, the Defender adopts a worst case strategy and considers the maximum of the false positive error probability over all the possible P_X and over all the possible attacks that the training sequence may have undergone, yielding:

$$\mathcal{S}_{\mathscr{D}} = \left\{ \Lambda^{n\times m} \subset \mathcal{P}_n \times \mathcal{P}_m \colon \max_{P_X \in \mathcal{P}} \max_{s \in \mathcal{S}_{\mathscr{A},T}} P_{\mathrm{FP}} \leq 2^{-\lambda n} \right\}. \tag{6.65}$$

While the above expression is formally equal to that of the DG-CTRa game (see (6.6)), the maximization over $\mathcal{S}_{\mathscr{A},T}$ is now more cumbersome, due to the additional degree of freedom available to the Attacker, who can selectively remove the samples of the original training sequence. In fact, even if $\mathscr{D}$ knew the position of the corrupted samples, simply throwing them away would not guarantee that the remaining part of the sequence would follow the same statistics of X, since the Attacker might have deliberately altered them by selectively choosing the samples to replace.

Attacker's Strategies

With regard to the Attacker, the part of the attack working on the test sequence y^n is the same as for the DG-CTRa case, while the part regarding the corruption of the training sequence must be redefined. To this purpose, we observe that the corrupted training sequence may be any sequence t^m for which $d_H(t^m, \tau^m) \leq \alpha m$, where d_H denotes the Hamming distance. Given that the Defender bases his decision

on the type of t^m, it is convenient to rewrite the constraint on the Hamming distance between sequences as a constraint on the L_1 distance between the corresponding types. In fact, by looking at the empirical distribution of the corrupted sequence, searching for a sequence t^m s.t. $d_H(t^m, \tau^m) \leq \alpha m$ is equivalent to searching for a pmf $P_{t^m} \in \mathcal{P}_m$ for which $d_{L_1}(P_{t^m}, P_{\tau^m}) \leq 2\alpha$ (see the proof of Lemma 3.5 in Chapter 3). Therefore, the set of strategies of the Attacker is defined by

$$\mathcal{S}_{\mathscr{A}} = \mathcal{S}_{\mathscr{A},T} \times \mathcal{S}_{\mathscr{A},O}, \tag{6.66}$$

where

$$\begin{aligned} \mathcal{S}_{\mathscr{A},T} = \{&Q(P_{\tau^m}, P_{y^n})\colon \mathcal{P}_m \times \mathcal{P}_n \to \mathcal{P}_m \\ &\text{such that } d_{L_1}(Q(P_{\tau^m}, P_{y^n}), P_{\tau^m}) \leq 2\alpha\}, \end{aligned} \tag{6.67}$$

$$\mathcal{S}_{\mathscr{A},O} = \{S^n_{YZ}(P_{y^n}, P_{t^m})\colon \mathcal{P}_n \times \mathcal{P}_m \to \mathcal{A}^n(L, P_{y^n})\}. \tag{6.68}$$

Note that, in this case, the function $Q(\cdot,\cdot)$ gives the type of the whole training sequence observed by $\mathscr{D}$ (not only the fake subpart, as it was in the DG-CTRa case), that is, $P_{t^m} = Q(P_{\tau^m}, P_{y^n})$.

In the following we find convenient to express the attack strategies in $\mathcal{S}_{\mathscr{A},T}$ in an alternative way. Since the Attacker *replaces* the samples of a subpart of the training sequence, the corruption strategy is equivalent to first *removing* a subpart of the training sequence and then *adding* a fake subsequence of the same length. Then, reordering is performed to hide the position of the fake samples. By focusing on the type of the observed training sequence, we can write:

$$P_{t^m} = P_{\tau^m} - \alpha Q_R(P_{\tau^m}, P_{y^n}) + \alpha Q_A(P_{\tau^m}, P_{y^n}). \tag{6.69}$$

where $Q_R(P_{\tau^m}, P_{y^n})$ and $Q_A(P_{\tau^m}, P_{y^n})$ (both belonging to $\mathcal{P}_{m_2}$) are the types of the removed and injected subsequences respectively. In order to simplify the notation, in the following we will avoid to indicate explicitly the dependence of $Q_R(P_{\tau^m}, P_{y^n})$ and $Q_A(P_{\tau^m}, P_{y^n})$ on P_{τ^m}, P_{y^n}, and will indicate them as $Q_R(\,)$ and $Q_A(\,)$. Furthermore, we will use the notation Q_R and Q_A whenever the dependence on the arguments is not relevant. By varying Q_R and Q_A, we obtain all the pmf's that can be produced from P_{τ^m} by first removing and later adding m_2 samples. Of course not all pairs $(Q_R,\ Q_A)$ are admissible since the P_{t^m} resulting

from (6.69) must be a valid pmf, i.e., it must be nonnegative for all the symbols of the alphabet $\mathcal{X}$.

Payoff

The payoff function is defined as before, that is

$$u(\Lambda^{n\times m}, (Q(\cdot,\cdot), S^n_{YZ}(\cdot,\cdot))) = -P_{\text{FN}}. \tag{6.70}$$

6.7 Solution of the DG-CTRr Game

Let us first rewrite the set of strategies available to $\mathscr{D}$ by using the attack formulation given in (6.69). For given P_X, Q_R and Q_A, P_{FP} is the probability that X generates two sequences x^n and τ^m, such that the pair of type classes $(P_{x^n}, P_{\tau^m} - \alpha(Q_R(\,) - Q_A(\,)))$ falls outside $\Lambda^{n\times m}$. Therefore:

$$\begin{aligned}\mathcal{S}_{\mathscr{D}} = \Bigg\{ \Lambda^{n\times m}\colon & \max_{P_X\in\mathcal{P}} \max_{Q_R(\,),Q_A(\,)} \sum_{P_{y^n}\in\mathcal{P}_n} P_Y(T(P_{y^n})) \\ & \cdot \sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} P_X(T(P_{x^n})) \\ & \cdot \sum_{\substack{P_{\tau^m}\in\mathcal{P}_m:\\ P_{\tau^m}-\alpha(Q_R(\,)-Q_A(\,))=P_{t^m}}} P_X(T(P_{\tau^m})) \le 2^{-\lambda n} \Bigg\}. \end{aligned} \tag{6.71}$$

By proceeding as in the proof of Lemma 6.1, it is easy to prove that the asymptotically optimal strategy for the Defender corresponds to the following:

$$\begin{aligned}\Lambda^{n\times m,*} = \Bigg\{ & (P_{x^n}, P_{t^m})\colon \\ & \min_{Q_R,Q_A\in\mathcal{P}_{m_2}} h\left(P_{x^n}, P_{t^m} + \alpha(Q_R - Q_A)\right) \le \lambda - \delta_n \Bigg\}, \end{aligned} \tag{6.72}$$

where δ_n tends to zero as $n\to\infty$ and the minimization is limited to the Q_R and Q_A in $\mathcal{P}_{m_2}$ such that $P_{t^m} + \alpha(Q_R - Q_A)$ is a valid pmf.

Consequently, the optimal attack strategy is given by:

$$\begin{aligned}&(Q^*(P_{\tau^m}, P_{y^n}), S^{n,*}_{YZ}(P_{y^n}, P_{t^m}))\\ &\quad= \underset{\substack{P_{t^m} \text{ s.t. } d_{L_1}(P_{t^m},P_{\tau^m})\leq 2\alpha\\ S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})}}{\operatorname{argmin}} \Big[\min_{Q_R,Q_A} h\left(P_{z^n}, P_{t^m}+\alpha(Q_R-Q_A)\right)\Big],\end{aligned} \tag{6.73}$$

hence resulting in the following theorem.

Theorem 6.6 (Equilibrium Point of the DG-CTRr Game)**.** The DG-CTRr game with targeted corruption is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile $(\Lambda^{n\times m,*}, (Q^*,\ S^{n,*}_{YZ}))$ given by Equations (6.72) and (6.73).

In order to study the asymptotic payoff of the DG-CTRr game at the equilibrium, we parallel the analysis carried out in Section 6.4.3. By considering the case $L = 0$, as a consequence of the attack to the training sequence, the set of pairs of types for which $\mathscr{D}$ will accept H_0 is given by

$$\begin{aligned}\Gamma^n_0(\lambda,\alpha) = \{(P_{y^n}, P_{\tau^m})\colon \exists\ P_{t^m} \text{ s.t. } d_{L_1}(P_{t^m}, P_{\tau^m}) \leq 2\alpha\\ \text{and } (P_{y^n}, P_{t^m}) \in \Lambda^{n\times m,*}\}.\end{aligned} \tag{6.74}$$

If we fix the type of the original training sequence, we get:

$$\begin{aligned}\Gamma^n_0(P_{\tau^m},\lambda,\alpha) &= \{P_{y^n}\colon \exists\ P_{t^m} \text{ s.t. } d_{L_1}(P_{t^m}, P_{\tau^m}) \leq 2\alpha\\ &\qquad \text{and } P_{y^n} \in \Lambda^{n,*}(P_{t^m})\}\\ &= \{P_{y^n}\colon \exists\ P_{t^m},\ \exists\ Q, Q' \in \mathcal{P}_{m_2},\ \text{s.t.}\\ &\qquad d_{L_1}(P_{t^m}, P_{\tau^m}) \leq 2\alpha\\ &\qquad \text{and } h(P_{x^n}, P_{t^m} - \alpha Q' + \alpha Q) \leq \lambda - \delta_n\}.\end{aligned} \tag{6.75}$$

By letting n go to infinity, we obtain the asymptotic counterpart of the above set, which, for a generic $R \in \mathcal{P}$, takes the following expression:

$$\begin{aligned}\Gamma_0(R,\lambda,\alpha) = \{P\colon \exists\ P', Q, Q',\ \text{s.t. } d_{L_1}(P', R) \leq 2\alpha\\ \text{and } h_c(P, P' - \alpha Q' + \alpha Q) \leq \lambda\}.\end{aligned} \tag{6.76}$$

When $L \neq 0$, we obtain:

$$\Gamma(R,\lambda,\alpha,L) = \{P\colon \exists V \in \Gamma_0(R,\lambda,\alpha) \text{ s.t. } EMD(P,V) \leq L\}. \tag{6.77}$$

Given the above definitions, it is straightforward to extend Theorem 6.3 to the DG-CTRr case, thus proving that the set in (6.77) evaluated in $R = P_X$ represents the indistinguishability region of the DG-CTRr game.

6.8 Source Distinguishability in the DG-CTRr Setup

We are now interested in studying the distinguishability of two sources X and Y in the DG-CTRr setup and compare it with the result we have obtained for the DG-CTRa case. To do so, we consider the behavior of the indistinguishability region when λ tends to zero. We have:

$$\Gamma(P_X, \alpha, L) = \{P\colon \exists\, V \in \Gamma_0(P_X, \alpha) \text{ s.t. } EMD(P, V) \leq L\}, \tag{6.78}$$

where

$$\begin{aligned}\Gamma_0(P_X, \alpha) &= \{P\colon \exists\, P', Q, Q' \text{ s.t. } d_{L_1}(P', P_X) \leq 2\alpha \\ &\qquad \text{and } P = P' + \alpha(Q - Q')\} \\ &= \{P\colon \exists\, P' \text{ s.t. } d_{L_1}(P', P_X) \leq 2\alpha \\ &\qquad \text{and } d_{L_1}(P, P') \leq 2\alpha\}. \end{aligned} \tag{6.79}$$

We observe that, the set in (6.79) can be equivalently rewritten as

$$\Gamma_0(P_X, \alpha) = \{P\colon d_{L_1}(P, P_X) \leq 4\alpha\}. \tag{6.80}$$

To see why, we first notice that the set in (6.79) is contained in (6.80). Indeed, from the triangular inequality we have that, for any P', $d(P, P_X) \leq d_{L_1}(P, P') + d_{L_1}(P', P_X)$. Then, if P belongs to $\Gamma_0(P_X, \alpha)$ in (6.79), it also belongs to the set in (6.80). To see that the two sets are indeed equivalent, it is sufficient to show that the reverse implication also holds. To this purpose, we observe that, whenever $d_{L_1}(P, P_X) \leq 4\alpha$, a type P^* can be found such that its distance from both P and P_X is less or at most equal to 2α. In fact, by letting $P^* = \frac{P+P_X}{2}$, we have

$$\begin{aligned} d_{L_1}(P, P^*) = d_{L_1}(P^*, P_X) &= \sum_i \left|\frac{P(i) - P_X(i)}{2}\right|, \\ d_{L_1}(P, P_X) &= \sum_i |P_X(i) - P(i)| = 2d_{L_1}(P, P^*). \end{aligned} \tag{6.81}$$

If $d_{L_1}(P, P_X) \leq 4\alpha$, then

$$d_{L_1}(P, P^*) = d_{L_1}(P^*, P_X) = d_{L_1}(P, P_X)/2 \leq 2\alpha, \tag{6.82}$$

permitting us to conclude that the sets in (6.79) and (6.80) are equivalent.

Therefore, the set in (6.78) can be rewritten as

$$\Gamma(P_X, \alpha, L) = \Big\{P\colon \min_{V\colon\, EMD(P,V)\leq L} d_{L_1}(V, P_X) \leq 4\alpha\Big\}. \tag{6.83}$$

Arguably, $\Gamma(P_X, \alpha, L)$ corresponds to the *ultimate indistinguishability region* of the DG-CTRr game (the extension of Theorem 6.5 to this case going along the same steps).

Upon inspection of (6.80), we see that, as expected, the ultimate indistinguishability region for $L = 0$ (and hence, also for the case $L \neq 0$ in (6.83)) is larger than that of the DG-CTRa game (see (6.54)), thus confirming that the game with sample replacement is more favorable to the Attacker (a graphical comparison between the indistinguishability regions for the two setups is shown in Figure 6.8). As a matter of fact, for the Attacker, the advantage of the DG-CTRr setup with respect to the DG-CTRa setup depends on α. For small values of α and for α close to 1/2, the indistinguishability regions of the two games are

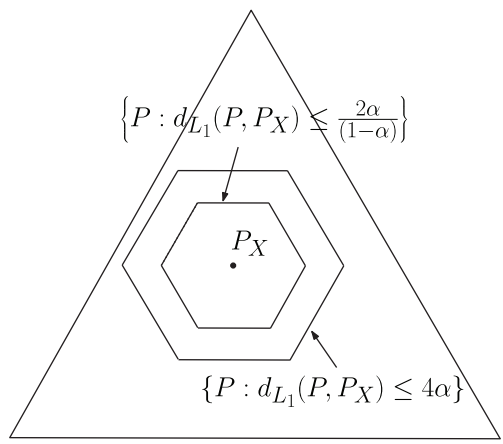


Figure 6.8: Comparison of the ultimate indistinguishability regions for the DG-CTRa and DG-CTRr games with $L = 0$ (no corruption of the test).

very similar, while for intermediate values of α the indistinguishability region of the DG-CTRr game is considerably larger than that of the DG-CTRa game (the maximum difference between the two regions is obtained for $\alpha \approx 0.3$). When $\alpha = 1/2$ the Attacker always wins, since he is able to bring any pmf inside the acceptance region regardless of the corruption setup, while for $\alpha = 0$, we fall back into the detection game without corruption of the training sequence, thus making the two versions of the game equivalent.

Given two sources X and Y, the blinding corruption level takes the expression:

$$\alpha_b = \frac{d_{L_1}(P_Y, P_X)}{4}. \tag{6.84}$$

Since $d_{L_1}(P_Y, P_X) \leq 2$ for any pair (P_Y, P_X),[7] the blinding value for the DG-CTRr game is lower than the blinding value for the DG-CTRa game. The two expressions are identical when the two sources have disjoint support, in which case $\alpha_b = 1/2$.

Below, we give the definition of the Security Margin.

Definition 6.3 (Security Margin in the DG-CTRr Setup)**.** Let $X \sim P_X$ and $Y \sim P_Y$ be two discrete memoryless sources. The maximum distortion for which the two sources can be reliably distinguished in the DG-CTRr setup, in the presence of a fraction α of corrupted samples, is given by

$$\mathcal{SM}_\alpha(P_X, P_Y) = L^*_\alpha, \tag{6.85}$$

where L^*_α is the quantity which satisfies the following relation

$$\min_{V:\ EMD(P_Y,V) \leq L^*_\alpha} d_{L_1}(V, P_X) = 4\alpha, \tag{6.86}$$

if $P_Y \notin \Gamma_0(P_X, \alpha)$, and $L^*_\alpha = 0$ otherwise.

Considering again the case of two Bernoulli sources, and by adopting the same notation of Section 6.5.2, we have that $\alpha_b = |p - q|/4$, while the Security Margin is

$$\mathcal{SM}_\alpha(p, q) = \begin{cases} |q - p| - 2\alpha & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b. \end{cases} \tag{6.87}$$

[7] The maximum value 2 is taken when the two distribution have disjoint support.

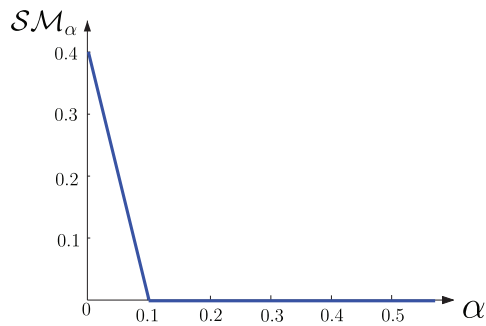


Figure 6.9: Security Margin as a function of α for Bernoulli sources with parameters $p = 0.3$ and $q = 0.7$ ($\alpha_b = 0.1$).

Figure 6.9 plots $\mathcal{SM}_\alpha$ as a function of α when $p = 0.3$ and $q = 0.7$. The blinding value is $\alpha_b = 0.1$ which, as expected, is lower than the value we found for the DG-CTRa setup.

7

Summary and Outlook

Aiming at giving adversarial signal processing a sound theoretical basis, in this monograph we lay the basis of a general theory that takes into account the impact that an adversary has on the design of effective signal processing tools, by focusing on the most common problem in adversarial signal processing, namely binary detection or hypothesis testing.

The main idea behind the theory consists in casting the adversarial binary decision problem into a game-theoretic framework, which permits to rigorously define the goals and the actions available to the contenders, namely, the Defender and the Attacker, and study the interplay between them by resorting to methods of information theory and large deviation theory. The outcome of the game opens the way to the analysis of the distinguishability of information sources in the presence of attacks by resorting to concepts of optimal transport, which permits to summarize the source distinguishability in a concise and elegant way. Specifically, the theory permits to state a necessary and sufficient condition under which it is possible to devise a detector for which both the false negative and false positive error probabilities tend to zero exponentially fast. Such a condition requires that the distortion introduced into the test sequence

during the attack is lower than a quantity, which we called Security Margin, strictly related to the Earth Mover Distance between the pdf's governing the generation of the test sequence under the two hypotheses being tested. The theory also permits to derive the best achievable false negative error exponent for a fixed false positive exponent, even though a closed form expression can be obtained only in some specific cases.

Several versions of the binary detection game have been addressed, depending on the knowledge available to the Defender and the Attacker about the statistical characterization of the system under analysis, and the behaviour of the Attacker. In all these cases, the game is solved under some limiting, yet reasonable, assumptions on the statistics used by the Defender to make a decision.

Overall, the bulk of theory summarized in this monograph contributes to show the potentiality of game-theoretic concepts coupled with tools of information theory and statistics, and points out interesting synergies with optimization theory.

The study of signal processing in adversarial setup is an open research field and several directions for future research can be pointed out starting from the work of this monograph. To start with, efforts can be made to relax some of the assumptions behind the theory. In particular, the memoryless assumption for the sources could be removed by considering more realistic models, e.g., Markov sources or renewal processes, which are commonly used to describe a wide variety of sources with memory and, at the same time, still amenable to be studied with the method of types. Relaxing the other main assumption behind the analysis, i.e., the assumption that the detection is based on a first-order statistical analysis, would allow to extend the applicability of the theory, especially for sources with memory, to cases wherein looking at higher-order statistics may help in making a correct decision. This extension – however comes with a number of additional difficulties, since it complicates the derivation of the optimal attack and then the analysis of the source distinguishability. We also mention the possibility of extending the analysis to the case of continuous sources. While the general ideas would remain the same, passing from discrete to continuous sources is non trivial since it requires that the method of types be extended to continuous sources. In a more elegant way, the

case of continuous sources can be studied by resorting to the Laplace integration method and, more generally, to the saddle point method [89, Chapter 4, p. 101], representing an analog of the method of types for the case of continuous sources.

More generally, adversarial classification or multiple hypothesis testing is another interesting problem which is worth studying under a unified framework, permitting to extend the theory to a large number of practical applications where the detector must distinguish among different classes of sources. Finally, it would be interesting to strengthen the connection with the field of adversarial machine learning, by extending the theoretical framework and analysis to other problems of adversarial binary detection with corruption of the training data, e.g., by considering a scenario where the Defender relies on multiple training sequences, and the Adversary interferes with the learning process by adopting a corruption strategy explicitly thought for such a case.

Acknowledgements

The authors are grateful to Alessandro Agnetis, of the University of Siena, for useful discussions on optimization concepts underlying the derivation of the quantities related to optimal transport theory.

The version of record is available at: http://dx.doi.org/10.1561/0100000102

Appendices

A

Regularity Properties of the Admissibility Set

To derive the theorems on the asymptotic behavior of the payoff in the various versions of the detection game, we need to prove some regularity properties of the set of admissible transportation maps $\mathcal{A}$ defined in (3.14), characterizing the set of strategies available to the Attacker. Such properties hold since the admissible set $\mathcal{A}$ is a *convex polytope*, i.e., the set of constraints defining $\mathcal{A}$ is linear.

To derive our results, we first need to define a distance measure between transportation maps, that is a function d_s: $\mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \to \mathbb{R}^+$, where we remind that $|\mathcal{X}|$ corresponds to the cardinality of the space the simplex $\mathcal{P}$ lives in Section 2.1. Let us (arbitrarily) consider the L_1 distance; then, given two maps (S_{PV}, S_{QR}), $d_s(S_{PV}, S_{QR}) = \sum_{i,j} |S_{PV}(i,j) - S_{QR}(i,j)|$.

Lemma A.1. Let $P \in \mathcal{P}$ and let P' be any pmf in the neighborhood of P of radius τ, for some $\tau > 0$, i.e., $P' \in \mathcal{B}(P,\tau)$. Then, $\delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)) \leq |\mathcal{X}|^2 \cdot \tau$, implying that $\delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)) \to 0$ as $\tau \to \infty$, uniformly in $\mathcal{P}$.

Furthermore, if we take $P' \in \mathcal{P}_n$, the following result holds: for any $\varepsilon > 0$, there exists τ^* and n^* such that $\forall\ \tau < \tau^*$ and $n > n^*$, $\delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P)) \leq \varepsilon$, $\forall P' \in \mathcal{B}(P,\tau) \cap \mathcal{P}_n$, and $\forall P \in \mathcal{P}$.

Proof. The lemma follows from the fact that $\mathcal{A}(L,P)$ is built by intersecting a finite number of half-spaces and is also bounded, i.e., is a convex polytope [90, Chapter 2], [91, p. 31]. By considering a P' close to P, we are perturbing the vector of the known terms of the linear constraints of the system which defines the admissibility set.

Given $P \in \mathcal{P}$ and $P' \in \mathcal{B}(P,\tau)$, for any map in $\mathcal{A}(L,P)$ we can choose a map $S_{P'V'}$ that works as follows: for the bins i such that $P'(i) \geq P(i)$, the same mass $S_{PV}(i,j)$ is moved from bin i to j, $\forall j \neq i$, while for $j = i$, $S_{P'V'}(i,j) = S_{PV}(i,j) + (P'(i) - P(i))$. For the bins i such that $P'(i) < P(i)$, first the index set $\{j\colon S_{PV}(i,j) \neq 0\}$ is sorted in decreasing order with respect to the amount of distortion introduced per unit of mass delivered $d(i,j)$; then, the mass is moved from bin i to the first j in the ordered list, until the amount $S_{PV}(i,j)$ is reached. Then, we pass to the second bin j in the list and go on until all the mass is moved from bin i. It is easy to argue that the map built in this way satisfies the distortion constraint (by construction, the distortion associated to $S_{P'V'}$ is less than that introduced by the admissible map S_{PV})[1] both in the case of additive distortion constraint (see (3.14)) and L_∞ distortion constraint (see (3.43)), which are the cases we focus on in this monograph. Then, $S_{P'V'} \in \mathcal{A}(L,P')$. Besides, by construction $|S_{P'V'}(i,j) - S_{PV}(i,j)| \leq \tau$, $\forall i,j$. Accordingly, $\max_{S_{PV} \in \mathcal{A}(L,P)} d(S_{PV}, \mathcal{A}(L,P')) \leq d_s(S_{PV}, S_{P'V'}) \leq |\mathcal{X}|^2 \cdot \tau$ and then $\delta_H(\mathcal{A}(L,P'), \mathcal{A}(L,P)) \leq |\mathcal{X}|^2 \cdot \tau$, thus concluding the proof of the first part.[2]

Let us now take $P' \in \mathcal{P}_n$. By exploiting the density of the rational numbers within the real ones, for any given map $S_{P'V} \in \mathcal{A}(L,P')$, we can find a map $S^n_{P'V'} \in \mathcal{A}^n(L,P')$ (i.e., having the same input marginal P' and satisfying the distortion constraint), such that $|S^n_{P'V'}(i,j)$

[1] Remember that any move from a bin to itself does not increase the distortion.

[2] We are implicitly exploiting the symmetry of the problem w.r.t. P and P', according to which $\max_{S_{PV} \in \mathcal{A}(L,P)} d(S_{PV}, \mathcal{A}(L,P')) = \max_{S_{P'V'} \in \mathcal{A}(L,P')} d(S_{P'V'}, \mathcal{A}(L,P))$ (see the definition of the Hausdorff distance in Section 2.1).

$- S_{P'V}(i,j)| \leq 1/n$. In fact, for any fixed i, we can define $S^n_{P'V'}$ as:

$$S^n_{P'V'}(i,j) = \max\{k\colon k/n \leq S_{P'V}(i,j)\}/n, \quad \forall j \neq i, \tag{A.1}$$

$$S^n_{P'V'}(i,i) = 1 - \sum_{j\neq i} S_{P'V}(i,j), \tag{A.2}$$

where $S^n_{P'V'}(i,i) \in \mathbb{Q}_n$ by construction (since the input distribution belongs to $\mathcal{P}_n$). It is easy to argue that the map defined in (A.2) belongs to $\mathcal{A}^n(L,P')$. By observing that $S_{P'V}(i,j) - 1/n \leq S^n_{P'V'}(i,j) \leq S_{P'V}(i,j)$, $\forall i,j, j \neq i$, and $S_{P'V}(i,i) \leq S^n_{P'V'}(i,i) \leq S_{P'V}(i,i) + (|\mathcal{X}| - 1)/n$, $\forall i$, we argue that $d_s(S^n_{P'V'}, S_{P'V}) \leq 2|\mathcal{X}|^2/n$. Therefore, by considering the discrete set $\mathcal{A}^n$, we can write

$$\begin{aligned}\delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P)) &\leq \delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P')) \\ &\quad + \delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)) \\ &\leq \delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P')) + |\mathcal{X}|^2 \cdot \tau \\ &\leq 2|\mathcal{X}|^2/n + |\mathcal{X}|^2 \cdot \tau. \end{aligned} \tag{A.3}$$

Then, for a fixed ε, by choosing τ^* and n^* such that $|\mathcal{X}|^2 \cdot (2/n^* + \tau^*) = \varepsilon$, we have that for any τ smaller than τ^* and n larger than n^*, $\delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P)) \leq \varepsilon$, thus concluding the second part of the proof. □

From the above lemma, it is easy to prove the following theorem.

Theorem A.2. Let $S_{PV} \in \mathcal{A}(L,P)$ for some $P \in \mathcal{P}$. For any point $P' \in \mathcal{B}(P,\tau)$, for some $\tau > 0$, we can find a map $S_{P'V'} \in \mathcal{A}(L,P')$ such that $V' \in \mathcal{B}(V,\varepsilon)$, with $\varepsilon \leq |\mathcal{X}|^2 \cdot \tau$.

Similarly, for any $\varepsilon' > 0$, there exists τ^* and n^* such that for all $\tau < \tau^*$ and $n > n^*$ we have the following: for any map $S_{PV} \in \mathcal{A}(L,P)$ a map $S^n_{P'V'}$ in $\mathcal{A}^n(L,P')$ can be found such that $V'_n \in \mathcal{B}(V,\varepsilon')$, $\forall P' \in \mathcal{B}(P,\tau) \cap \mathcal{P}_n$, and $\forall P \in \mathcal{P}$.

Proof. It is easy to see that for any map $S_{PV} \in \mathcal{A}(L,P)$ we can choose a map $S_{P'V'} \in \mathcal{A}(L,P')$ such that, for all j,

$$\begin{aligned} V'(j) = \sum_i S_{P'V'}(i,j) &< \sum_i (S_{PV}(i,j) + |S_{P'V'}(i,j) - S_{PV}(i,j)|) \\ &\leq V(j) + d_s(S_{P'V'}, S_{PV}) \\ &\leq V(j) + \delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)), \end{aligned} \tag{A.4}$$

and, similarly, $V'(j) \geq V(j) - \delta_H(\mathcal{A}(L, P'), \mathcal{A}(L, P))$. Accordingly, if $P' \in \mathcal{B}(P, \tau)$, by exploiting Lemma A.1, we get

$$|V'(j) - V(j)| < \delta_H(\mathcal{A}(L, P'), \mathcal{A}(L, P)) < |\mathcal{X}|^2 \cdot \tau, \tag{A.5}$$

and hence $V' \in \mathcal{B}(V, |\mathcal{X}|^2 \cdot \tau)$. Similarly, for the second part, we observe that, from Lemma A.1, for a proper choice of the admissible map $S^n_{P'V'}$ we have

$$|V'_n(j) - V(j)| < \delta_H(\mathcal{A}^n(L, P'), \mathcal{A}(L, P)) \leq 2|\mathcal{X}|^2/n + |\mathcal{X}|^2 \cdot \tau. \tag{A.6}$$

Then, for a fixed ε, we can choose τ^* and n^* such that $2|\mathcal{X}|^2/n^* + |\mathcal{X}|^2 \cdot \tau^* = \varepsilon$. □

B

Asymptotic Behavior of the Indistinguishability Regions

B.1 Behavior of Set Γ (and Γ_{tr}) for $\lambda \to 0$

We start by studying the behavior of $\Gamma(P_X, \lambda, L)$ when $\lambda \to 0$. More specifically, we show that for small values of λ the set $\Gamma(P_X, \lambda, L)$ approaches $\Gamma(P_X, L)$ smoothly.

As a first step, we highlight the following proposition.

Proposition B.1. *EMD*(P, Q) is a continuous and convex function of P and Q.

Proof. The proposition follows immediately if we look at the *EMD* as the solution of a LP problem, wherein P and Q are the known terms of the linear constraints. In fact, the minimum of the objective function of an LP problem is a continuous and convex function of the known terms of the linear constraints (known result in operations research [90, Chapter 2]). □

By exploiting the continuity of the divergence and the continuity and convexity of the *EMD*, we now show that when λ tends to zero, the set $\Gamma(P_X, \lambda, L)$ tends to $\Gamma(P_X, L)$ regularly. More precisely, the following lemma holds.

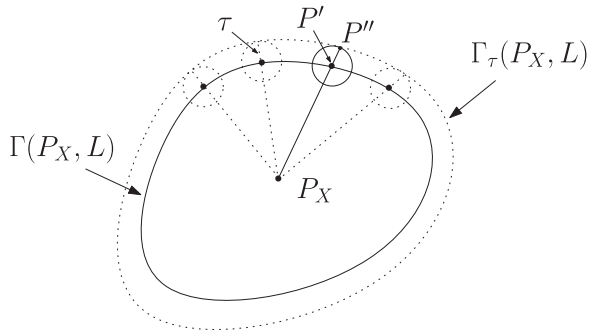


Figure B.1: Graphical representation of the set $\Gamma_\tau(P_X, L)$.

Lemma B.1. Let $X \sim P_X$ be an information source and L the maximum allowable average per-letter distortion in the DG-KS setup. The set $\Gamma(P_X, \lambda, L)$, defined in (4.1), satisfies the following property:

$$\forall \tau > 0,\ \exists \lambda > 0 \text{ s.t. } \forall P \in \Gamma(P_X, \lambda, L)\ \exists P' \in \Gamma(P_X, L) \text{ s.t. } P \in \mathcal{B}(P', \tau),$$

where $\Gamma(P_X, L)$ is defined as in (4.2) and $\mathcal{B}(P', \tau)$ is a ball centered in P' with radius τ.

Proof. Throughout the proof we will refer to Figure B.1 where all the sets and quantities involved in the proof are sketched. For any $\tau > 0$, we consider the set:

$$\Gamma_\tau(P_X, L) = \{P\colon \exists P' \in \Gamma(P_X, L) \text{ s.t. } P \in \mathcal{B}(P', \tau)\}. \tag{B.1}$$

With such a definition, we can rephrase (B.1) as follows:

$$\forall \tau > 0,\ \exists\ \lambda > 0 \text{ s.t. } \Gamma_{ks}(P_X, \lambda, L) \subseteq \Gamma_\tau(P_X, L). \tag{B.2}$$

For the sake of simplicity, we will prove a slightly stronger version of the lemma by means of the following 2-step proof. First, we will show that a subset of $\Gamma_\tau(P_X, L)$ exists having the following form:

$$\Gamma_\tau^{sub}(P_X, L) = \{P\colon EMD(P, P_X) \leq L + \delta(\tau)\}, \tag{B.3}$$

for some $\delta(\tau) > 0$. Then, we will prove that for small enough λ, any $P \in \Gamma(P_X, \lambda, L)$ belongs to $\Gamma_\tau^{sub}(P_X, L)$.

To start with, let P' be any point on $\mathcal{C}(\Gamma(P_X, L))$, the boundary of $\Gamma(P_X, L)$. Among all the points on the boundary of the ball of radius τ and centered in P', consider the one, name it P'', lying along the direction given by the line joining P_X and P' and falling outside $\Gamma(P_X, L)$ (see Figure B.1). By the convexity of the *EMD* (Property B.1) and since $EMD = 0$ if and only if $P = P_X$, we conclude that $EMD(P'', P_X) > EMD(P', P_X)$. Since P' lies on the boundary of $\Gamma(P_X, L)$ we know that $EMD(P'', P_X) = L + \mu$, where $\mu = \mu(P', \tau)$ is a strictly positive quantity. We now show that the first part the proof holds by letting $\delta(\tau) = \min_{P' \in \mathcal{C}(\Gamma(P_X, L))} \mu(P', \tau)$. To this purpose, let P be any point in set $\Gamma_\tau^{sub}(P_X, L)$ for the above choice of $\delta(\tau)$. If $P \in \Gamma(P_X, L)$, then, by definition, P also belongs to $\Gamma_\tau(P_X, L)$. On the other side, if P lies outside $\Gamma(P_X, L)$, let us denote by P^* the point lying on the boundary of the set $\Gamma(P_X, L)$ along the line joining P and P_X, and let P^{**} be the point where the same line crosses the ball $\mathcal{B}(P^*, \tau)$ outside $\Gamma(P_X, L)$. Now, $EMD(P, P_X) \leq L + \delta(\tau) \leq EMD(P^{**}, P_X)$ by construction. Because of the convexity of *EMD*, then $P \in \mathcal{B}(P^*, \tau)$ as required.

Let us now pass to the second part of the proof. First, we notice that set $\Gamma(P_X, \lambda, L)$ depends on λ only through the acceptance region $\Lambda^*(P_X, \lambda)$. If λ is small, due to the continuity of the divergence, for any $Q \in \Lambda^*(P_X, \lambda)$ we will have $Q \in \mathcal{B}(P_X, \kappa(\lambda))$ for some $\kappa(\lambda)$ such that $\kappa(\lambda) \to 0$ when $\lambda \to 0$. Let, then, P be a pmf in $\Gamma(P_X, \lambda, L)$. By definition, a $Q \in \Lambda^*(P_X, \lambda)$ exists s.t. $EMD(P, Q) \leq L$. If λ is small, due to the proximity of Q to P_X and the continuity of the *EMD* we have that $EMD(P, P_X) < EMD(P, Q) + \eta(\lambda) \leq L + \eta(\lambda)$ with $\eta(\lambda)$ approaching 0 when $\lambda \to 0$. In particular, if λ is small enough $\eta(\lambda) < \delta(\tau)$ and hence $P \in \Gamma_\tau^{sub}(P_X, L)$ which in turn is entirely contained in $\Gamma_\tau(P_X, L)$ thus completing the proof. □

In the same way, we can prove that Lemma B.1 holds also when $\Gamma(P_X, \lambda, L)$ is replaced by $\Gamma_{tr}(Q, \lambda, L)$ and $\Gamma(P_X, L)$ by $\Gamma_{tr}(Q, L)$ with a generic Q instead of P_X. To be convinced about that, it is sufficient to note that the only difference between Γ and Γ_{tr} relies on the test function which defines the acceptance region, respectively the divergence and the h_c function. Since the h_c function is still a continuous and convex

function and, likewise $\mathcal{D}$, is equal to zero if and only if its arguments are identical, the proof that we used for Lemma B.1 still holds.

B.2 Behavior of Γ_{L_∞} for $\lambda \to 0$

We prove that when $\lambda \to 0$ $\Gamma_{L_\infty}(P_X, \lambda, L)$ approaches $\Gamma_{L_\infty}(P_X, L)$ regularly, in the sense stated by the following lemma.

Lemma B.2 (Extension of Lemma B.1 to the L_∞ Case)**.** Let $X \sim P_X$ be an information source and L the maximum per-sample distortion allowed to the Attacker. The set $\Gamma_{L_\infty}(P_X, \lambda, L)$, defined in Section 3.5, satisfies the following property:

$$\begin{aligned} &\forall \tau > 0,\ \exists\, \lambda > 0 \text{ s.t., } \forall P \in \Gamma_{L_\infty}(P_X, \lambda, L) \\ &\qquad \exists P' \in \Gamma_{L_\infty}(P_X, L) \text{ s.t. } P \in \mathcal{B}(P', \tau). \end{aligned} \tag{B.4}$$

Proof. We will prove the lemma by assuming that the distance defining the ball $\mathcal{B}(P', \tau)$ is the L_1 distance, extending the proof to other distances being straightforward.

For a fixed $\tau > 0$, let P be a pmf in $\Gamma_{L_\infty}(P_X, \lambda, L)$ for some λ. This means that at least one pmf $Q \in \Lambda^*(P_X, \lambda)$ exists, such that P can be mapped into Q with maximum shipment distance lower than or equal to L. From Equation (3.22) and by exploiting the continuity of the divergence function, we argue that $Q \in \mathcal{B}(P_X, \gamma(\lambda))$ for some positive $\gamma(\lambda)$, and where $\gamma(\lambda) \to 0$ as $\lambda \to 0$. Accordingly, P_X can be written as $P_X(j) = Q(j) + \gamma(j)$, $\forall j$, where $\sum_{j \in \mathcal{X}} |\gamma(j)| < \gamma(\lambda)$. Note that, by construction, $\sum_j \gamma(j) = 0$ and $\gamma(j) \to 0$ when $\lambda \to 0$. Let S_{PQ} be an admissible map bringing P into Q (such a map surely exists by construction). We prove the lemma by explicitly building a pmf P' and a new admissible transportation map S', such that, P' is arbitrarily close to P (for a small enough λ) and S' maps P' into P_X. We start by introducing two new quantities, namely $\gamma^+(j)$, defined as follows:

$$\begin{aligned} \gamma^+(j) &= \gamma(j) \quad \text{if } P_X(j) - Q(j) \geq 0 \\ \gamma^+(j) &= 0 \qquad \text{if } P_X(j) - Q(j) < 0, \end{aligned} \tag{B.5}$$

and $\gamma^-(j)$ defined as

$$\begin{aligned} \gamma^-(j) &= -\gamma(j) \quad \text{if } P_X(j) - Q(j) < 0 \\ \gamma^-(j) &= 0 \qquad \text{if } P_X(j) - Q(j) \geq 0. \end{aligned} \tag{B.6}$$

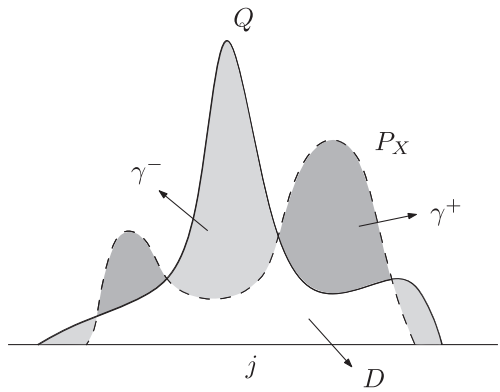


Figure B.2: Geometric interpretation of γ^+, γ^- and $D(j)$.

A graphical interpretation of γ^+ and γ^- is given in Figure B.2. Clearly, $\sum_j \gamma^-(j) = \sum_j \gamma^+(j)$. With the above definitions, we can look at the demand distribution Q as consisting of two amounts: the mass distribution D, with $D(j) = \min\{P_X(j), Q(j)\}$, and γ^-. According to the superposition principle, the map S_{PQ} can then be split into two sub-maps: one that satisfies the demand of D (let us call it S^D_{PQ}), and one that satisfies the demand of γ^- (let us call it S^γ_{PQ}). The same distinction can be made in the source distribution:

$$P(i) = \sum_j S^D_{PQ}(i,j) + \sum_j S^\gamma_{PQ}(i,j) = P_D(i) + P_\gamma(i), \tag{B.7}$$

where P_D and P_γ are the masses in the source distribution which are used to satisfy the mass demand pertaining to D and γ^- according to mapping S_{PQ}. Then, $\sum_i P_D(i) = D$ and $\sum_i P_\gamma(i) = \gamma^-$. In order to construct the pmf P' we are looking for, we simply remove from P the amount of mass P_γ used to fill γ^- and redistribute it according to γ^+. Specifically, we have

$$P'(i) = P_D(i) + \gamma^+(i) \tag{B.8}$$

$$S'(i,j) = S^D_{PQ}(i,j) + \gamma^+(j)\delta(i,j), \tag{B.9}$$

where $\delta(i,j)$ is equal to 1 if $i = j$ and 0 otherwise. It is easy to see that applying the transportation map $S'(i,j)$ to P' yields P_X. Besides, from the procedure adopted to build S', it is evident that

$$\max_{(i,j):\, S'(i,j)\neq 0} |i-j| \leq \max_{(i,j):\, S_{PQ}(i,j)\neq 0} |i-j| \leq L, \tag{B.10}$$

(the only new shipments introduced are from a bin to itself). In addition, the distance between P' and P is, by construction, lower than $\gamma(\lambda)$, which can be made arbitrarily small by decreasing λ, thus completing the proof of the lemma. □

C

Security Margin Computation as a Minimum Cost Flow Problem

This appendix includes the proof of the closed form expression in (4.13) for the Security Margin in the case of L_1 distance (Section 4.4.3).

When d corresponds to the L_1 distance, the per-letter distortion (or cost) is $d(i,j) = |i-j|$, for $i,j \in \mathcal{X}$, which obviously satisfies the Monge property (see Section 2.3); hence, the *EMD* can be computed by applying the NWC rule (see Section 2.3.2). By formulating the TP problem as a minimum cost flow problem [50, Section 1.2], it is possible to find a closed form expression for the minimum transportation cost, which corresponds to (4.13).

To present the arguments, we refer to the example of TP problem illustrated in Figure C.1(a) (w.l.o.g.). Specifically, Figure C.1(a) illustrates two pmf's P_X and P_Y defined on an alphabet $\mathcal{X}$; the graphical representation of the optimum transportation map between P_X (source) and P_Y (sink) based on the NWC rule, namely S^*_{XY}, is reported in Figure C.1(b).[1] Let us consider the flow graph representation associated to the TP problem. In a flow graph representation, each bin of the alphabet is represented as a node; if, at bin i, the value of $P_X(i)$ is

[1] Although we limit our discussion to probability distributions, the TP problem, and hence the derivation in this appendix, holds for arbitrary mass functions (see the general formulation of the Hitchcock TP in (2.17)).

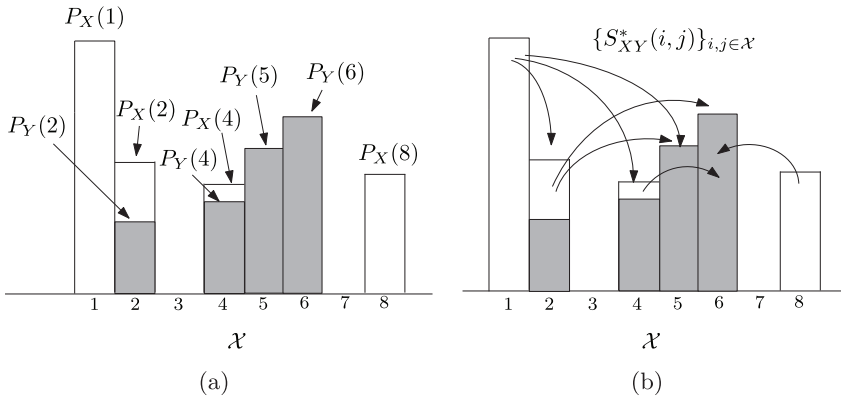


Figure C.1: Example of TP problem: (a) Two pmf's, P_X and P_Y, defined over $\mathcal{X}$; (b) optimum transportation map from P_X (source) to P_Y (sink). The arrows represent (non-zero) mass shipments according to the optimum map S^*_{XY} obtained through the application of the NWC rule.

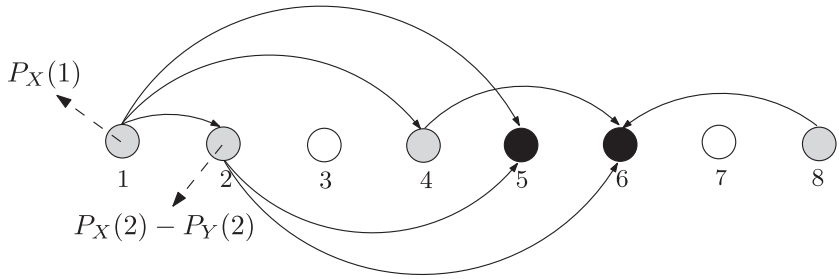
larger than $P_Y(i)$, then the corresponding node i is a (so-called) *surplus* node; conversely, if it is smaller, then i is a *demand* node. The quantity $S_{XY}(i,j)$ denotes the amount of flow on the arc between node i and j. For a *surplus* (*demand*) node, the sum of the output flows, i.e., the flows leaving the node, is larger (smaller) than the sum of the incoming flows, i.e., the flows entering the node.[2] An admissible solution for the TP problem corresponds to a feasible flow in the graph, i.e., a flow which satisfies: (i) the constraint on the flow capacity, and non-negativity, that is, $0 \leq S_{XY}(i,j) \leq 1$, $\forall i,j$; (ii) the flow conservations constraint at the nodes, that is, $\forall i$:

$$\sum_{j\neq i} S_{XY}(i,j) - \sum_{k\neq i} S_{XY}(k,i) = P_X(i) - P_Y(i), \tag{C.1}$$

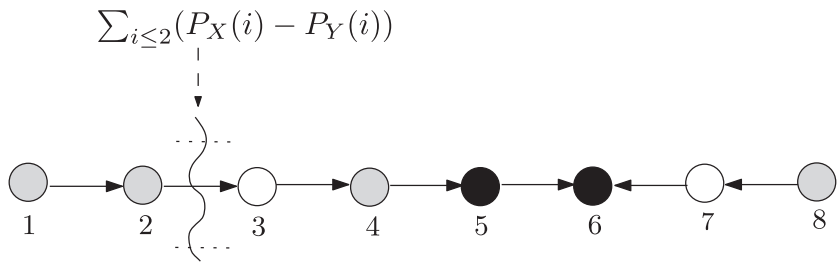
where the first term in the left-hand side denotes the sum of the output flows at node i and the second term the sum of the incoming ones. The overall cost can be computed by summing the costs for all the arc flows between pairs of nodes.

Solving the TP problem corresponds to find the feasible flow S_{XY} that minimizes the cost. Figure C.2(a) illustrates the minimum cost

[2] If $P_X(i) = P_Y(i)$, the node i is called *transhipment* node.



(a) Minimum cost flow graph associated to the TP problem in Figure C.1. The surplus nodes are those for which $P_X(i) > P_Y(i)$ (in light gray), the demand nodes are those for which $P_X(i) < P_Y(i)$ (in black). Node 3 and 7 are transhipment nodes, where P_X and P_Y are both null.



(b) Equivalent cost flow graph with single hops.

Figure C.2: Optimal transport problem in Figure C.1 represented as a minimum cost flow problem.

flow graph associated to the TP problem in Figure C.1 (the flows over the arcs are not reported for ease of representation); the flow map corresponds then to S^*_{XY}, obtained with the NWC rule. According to the flow decomposition principle [79], given any feasible flow map for a graph G, adding (or removing) an amount of flow on a cycle (i.e., a path from a node to the same node), as well as adding and removing a similar amount at any intermediate node on a path, leads to a feasible flow map, since the overall flows at the nodes are preserved (flow conservation property) [50, Section 1.2, p. 5]. By following this principle, we can decompose a flow over multiple hops into a series of flows over single hops (i.e., hops between consecutive nodes), getting a new graph G'. Given the particular cost function adopted, the cost of moving a unit of flow from i to j, $j > i+1$ (i.e., a multiple hop move) is equivalent to

the sum of the costs of the moves (of the same unit of flow) from i to h and from h to j, for $i \leq h \leq j$. Then, G' achieves the same cost of the initial graph G.

Figure C.2(b) shows the graphs with single hops obtained from the graph in Figure C.2(a) by decomposing the flows as detailed above. Because of the flow conservation, it is easy to argue that, if G is the minimum cost flow graph associated to a TP problem, then, considering G', the flow on the arc from node 1 to 2 equals $(P_X(1) - P_Y(1))$, the flow from node 2 to 3 equals $(P_X(1) - P_Y(1)) + (P_X(2) - P_Y(2))$, and so on. When, for some node s, $\sum_{i \leq s}(P_X(i) - P_Y(i))$ is negative, this corresponds to a (positive) flow in the opposite direction i.e., from $s+1$ to s. With reference to the graph in Figure C.2, this happens when $s = 6, 7$.

Therefore, the overall cost associated to such a graph takes the expression:

$$\sum_{s \in \mathcal{X}} \left| \sum_{i=1}^{s} (P_X(i) - P_Y(i)) \right|. \tag{C.2}$$

Then, (4.13) follows immediately.

References

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8682–8686, Vancouver, Canada, 2013.

[2] M. Barni and B. Tondi, "The source identification game: An information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, 2013b.

[3] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, 2014. DOI: 10.1109/TIT.2014.2325571.

[4] M. Barni and B. Tondi, "Source distinguishability under distortion-limited attack: An optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, 2016. DOI: 10.1109/TIFS.2016.2570739.

[5] M. Barni and B. Tondi, "Adversarial source identification game with corrupted training," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3894–3915, 2018. DOI: 10.1109/TIT.2018.2806742.

[6] G. Carl, G. Kesidis, R. R. Brooks, and Suresh Rai, "Denial-of-service attack-detection techniques," *IEEE Internet Computing*, vol. 10, no. 1, pp. 82–89, 2006.

[7] G. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, pp. 335–455, Jan. 2006. DOI: 10.1561/1500000006.

[8] T. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, pp. 10 206–10 222, Sep. 2009. DOI: 10.1016/j.eswa.2009.02.037.

[9] F. Marturana, S. Tacconi, and G. Italiano, "Handbook of digital forensics of multimedia data and devices," *A Machine Learning-Based Approach to Digital Triage,* John Wiley & Sons, Ltd., pp. 94–132, 2015.

[10] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds., Springer, Berlin/Heidelberg, 2012.

[11] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.

[12] C. Cachin, "An information-theoretic model for steganography," in *Proceedings of IH98, Second International Workshop on Information Hiding,* Lecture Notes in Computer Science Series, vol. 6958, Springer, Berlin/Heidelberg, 1998, pp. 306–318.

[13] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *Proceedings of EUSIPCO'05, European Signal Processing Conference*, pp. 469–472, 2005.

[14] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability," *Computer Vision and Image Understanding*, vol. 197–198, 102988, 2020. DOI: 10.1016/j.cviu.2020.102988.

[15] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ACM, pp. 43–58, 2011.

[16] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018. DOI: 10.1016/j.patcog.2018.07.023.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory*. Prentice Hall, 1998.

[18] B. Tondi, N. Merhav, and M. Barni, "Detection games under fully active adversaries," *Entropy*, vol. 21, no. 1, 2019. DOI: 10.3390/e21010023.

[19] S. Yasodharan and P. Loiseau, "Nonzero-sum adversarial hypothesis testing games," in *Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 7312–7322, 2019.

[20] M. Barni and B. Tondi, "Multiple-observation hypothesis testing under adversarial conditions," in *2013 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 91–96, 2013a. DOI: 10.1109/WIFS.2013.6707800.

[21] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," in *Concurrency: The Works of Leslie Lamport*, 2019, pp. 203–226.

[22] A. Abrardo, M. Barni, K. Kallas, and B. Tondi, "A game-theoretic framework for optimum decision fusion in the presence of Byzantines," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1333–1345, 2016.

[23] A. Abrardo, M. Barni, K. Kallas, and B.Tondi, *Information Fusion in Distributed Sensor Networks with Byzantines*. Signals and Communication Technology, Springer, Singapore, 2021.

[24] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, 2009.

[25] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed Bayesian detection in the presence of Byzantine data," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5250–5263, 2015.

[26] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[27] J. Zhang, R. S. Blum, X. Lu, and D. D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1086–1101, 2014.

[28] A. Vempaty, T. Lang, and P. Varshney, "Distributed inference with Byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 65–75, 2013.

[29] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[30] X. Wang, J. Li, X. Kuang, Y. Tan, and J. Li, "The security of machine learning in an adversarial setting: A survey," *Journal of Parallel and Distributed Computing*, vol. 130, pp. 12–23, 2019.

[31] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proceedings of the ACM Multimedia and Security Workshop*, pp. 97–104, Coventry, UK, Jun. 2012.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[33] J. R. Munkres, *Topology,* Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. URL: https://books.google.it/books?id=XjoZAQAAIAAJ.

[34] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 2007.

[35] M. J. Osborne, *An Introduction to Game Theory*, vol. 3. Oxford University Press, New York, 2004.

[36] J. Nash, "Equilibrium points in *n*-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[37] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[38] J. Von Neumann, "Zur theorie der gesellschaftsspiele," ger, *Mathematische Annalen*, vol. 100, pp. 295–320, 1928. URL: http://eudml.org/doc/159291.

[39] V. Chvatal, "Linear programming," *A Series of Books in the Mathematical Sciences,* New York: W. H. Freeman and Company 1983, vol. 1, 1983.

[40] A. Charnes and W. W. Cooper, "Management models and industrial applications of linear programming," *Management Science*, vol. 4, no. 1, pp. 38–91, 1957.

[41] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium," *SIAM Journal on Computing*, vol. 39, no. 1, pp. 195–259, 2009.

[42] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.

[43] Y. C. Chen, N. Van Long, and X. Luo, "Iterated strict dominance in general games," *Games and Economic Behavior*, vol. 61, no. 2, pp. 299–315, 2007.

[44] P. Weirich, *Equilibrium and Rationality: Game Theory Revised by Decision Rules*. Cambridge University Press, 2007.

[45] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1991. URL: http://www.jstor.org/stable/j.ctvjsf522.

[46] I. L. Glicksberg, "A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points," *Proceedings of the American Mathematical Society*, vol. 3, no. 1, pp. 170–174, 1952. URL: http://www.jstor.org/stable/2032478.

[47] G. Monge, *Mémoire sur la Théorie des Déblais et des Remblais*. De l'Imprimerie Royale, 1781.

[48] C. Villani, *Topics in Optimal Transportation*, vol. 58. Graduate Studies in Mathematics Series: American Mathematical Society, 2003.

[49] S. T. Rachev, *Mass Transportation Problems: Volume I: Theory*, vol. 1. Springer, 1998.

[50] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ: Prentice-Hall, Inc., 1993.

[51] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal on Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[52] C. Villani, *Optimal Transport: Old and New*. Berlin: Springer-Verlag, 2009.

[53] T. S. Rachev, "The Monge–Kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[54] E. Levina and P. Bickel, "The Earth Mover's Distance is the Mallows distance: Some insights from statistics," in *Proceedings of the 8th IEEE International Conference on Computer Vision*, vol. 2, pp. 251–256, 2001. DOI: 10.1109/ICCV.2001.937632.

[55] A. J. Hoffman, "On simple linear programming problems," in *Proceedings of Symposia in Pure Mathematics*, World Scientific, vol. 7, pp. 317–327, 1963.

[56] R. E. Burkard, B. Klinz, and R. Rudolf, "Perspectives of Monge properties in optimization," *Discrete Applied Mathematics*, vol. 70, no. 2, pp. 95–161, 1996.

[57] J. B. Orlin, "A faster strongly polynomial minimum cost flow algorithm," *Operations Research*, vol. 41, no. 2, pp. 338–350, 1993.

[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

[59] I. N. Sanov, "On the probability of large deviations of random variables," *Matematicheskii Sbornik*, vol. 42, pp. 11–44, 1957.

[60] I. Csiszar, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[61] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38. Springer Science & Business Media, 2009.

[62] K. Kuratowski, *Topology*, vol. 2. Academic Press, 1968.

[63] G. Salinetti and J. B. Wets, "On the convergence of sequences of convex sets in finite dimensions," *Siam Review*, vol. 21, no. 1, pp. 18–33, 1979.

[64] N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 255–274, 2008.

[65] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 515–525, 2014.

[66] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *2012 IEEE International Conference on Computational Photography (ICCP)*, IEEE, pp. 1–10, 2012.

[67] A. C. Popescu and H. Farid, "Statistical tools for digital forensics," in *Proceedings of the 6th International Conference on Information Hiding,* IH'04, pp. 128–147, Toronto, Canada: Springer-Verlag, 2004. DOI: 10.1007/978-3-540-30114-1_10.

[68] B. Tondi, *Theoretical Foundations of Adversarial Detection and Applications to Multimedia Forensics*. University of Siena: PhD Thesis, Sep. 2016.

[69] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, 1965.

[70] M. R. Bussieck and A. Pruessner, "Mixed-integer nonlinear programming," *SIAG/OPT Newsletter: Views & News*, vol. 14, no. 1, pp. 19–22, 2003.

[71] M. Bussieck and S. Vigerske, "MINLP solver software," 2011. DOI: 10.1002/9780470400531.eorms0527.

[72] P. Bonami, M. Kilinc, and J. Linderoth, "Algorithms and software for convex mixed integer nonlinear programs," *Tech. Rep.* Computer Sciences Department, University of Wisconsin-Madison, 2009.

[73] S. Mallat, *A Wavelet Tour of Signal Processing*. Elsevier, 1999.

[74] F. F. Leimkuhler, *Introduction to Operations Research*. Taylor & Francis Group, 1968.

[75] R. Ansari, N. Memon, and E. Ceran, "Near-lossless image compression techniques," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 486–495, 1998.

[76] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 337–346, 2009.

[77] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *Vision Models for Target Detection and Recognition*, vol. 2, pp. 245–357, 1995.

[78] O. Pele and M. Werman, "Fast and robust Earth Mover's Distances," in *Proceedings ICCV'09, 12th IEEE International Conference on Computer Vision*, pp. 460–467, 2009.

[79] A. C. Williams, "A treatment of transportation problems by decomposition," English, *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 1, pp. 35–48, 1962.

[80] M. Kendall and S. Stuart, *The Advanced Theory of Statistics,* vol. 2, 4th edition. New York: MacMillan, 1979.

[81] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, 1989.

[82] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[83] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial*. Now Publishers Inc., 2004.

[84] W. A. Sutherland, *Introduction to Metric and Topological Spaces*. Oxford University Press, 1975.

[85] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in CNNs by training set corruption without label poisoning," *Proceedings of 2019 IEEE International Conference on Image Processing, ICIP 2019, arXiv:1902.11237*, 2019.

[86] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, 2015.

[87] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[88] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018,* San Diego, CA, USA, February 18–22, 2018, The Internet Society, 2018.

[89] N. Merhav, "Statistical physics and information theory," *Foundations and Trends in Communications and Information Theory*, vol. 6, pp. 1–212, 2010. DOI: 10.1561/0100000052.

[90] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, vol. 6. Athena Scientific Belmont, MA, 1997.

[91] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.