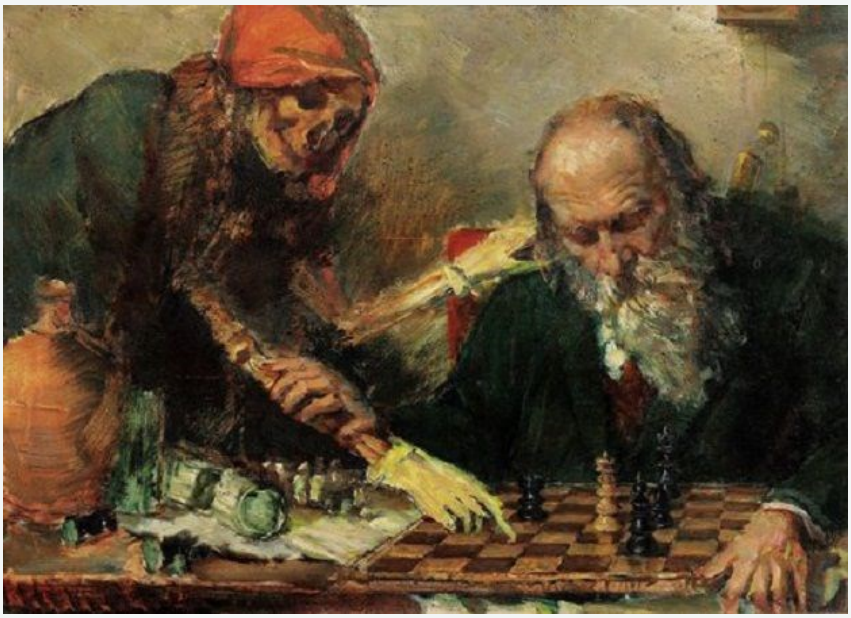
Theoretical Foundations of Adversarial Detection and Applications to Multimedia Forensics



Benedetta Tondi

Ph.D Thesis in Information Engineering

University of Siena

UNIVERSITÀ DEGLI STUDI DI SIENA

FACOLTÀ DI INGEGNERIA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Theoretical Foundations of Adversarial Detection and Applications to Multimedia Forensics

Benedetta Tondi

Ph.D Thesis in Information Engineering
XXVIII Cycle, 2012-2015

Supervisor
Prof. Mauro Barni

External reviewers
Prof. Rainer Böhme
Prof. Svyatoslav Voloshynovskiy

Examination Committee
Prof. Rainer Böhme
Prof. Giovanni Poggi
Prof. Alessandro Agnetis

SIENA
SEPTEMBER 20, 2016

Contents

List of Figures

List of Tables

List of Symbols

X	discrete random variable
$\mathcal{X}$	alphabet of symbols of X
$\lvert\mathcal{X}\rvert$	alphabet cardinality
x	realization of X, $x \in \mathcal{X}$
X^n	discrete random sequence of length n
x^n	sequence of length n (realization of X^n), $x^n \in \mathcal{X}^n$
X_i (res. x_i)	i-th element of X^n (res. x^n)
$\sigma(x^n)$	random permutation of x^n
$x^n \Vert y^N$	concatenation of x^n and y^N
$\mathcal{C}$	class of the memoryless information sources
$\mathcal{P}$	class of the probability density functions
P_X	probability mass function (pmf) of X $P_X(a)$, $a \in \mathcal{X}$, probability of symbol a
$P_X(x^n)$	probability that X emits the sequence x^n
P_{x^n}	empirical probability distribution induced by x^n or type of x^n
$P_{y^n\lvert x^n}$	empirical conditional probability distribution or conditional type of (y^n, x^n)
$\mathcal{P}_n$	set of types with denominator n

$\mathcal{T}(P)$, $P \in \mathcal{P}_n$	type class of P (set of sequences in $\mathcal{X}^n$ having type P)
$\mathcal{T}(P_{x^n})$, $\mathcal{T}(x^n)$	type class of P_{x^n} (set of sequences having the same type of x^n)
$\mathcal{T}(P_{y^n\|x^n})$, $\mathcal{T}(y^n\|x^n)$	conditional type class or set of sequences having conditional type $P_{y^n\|x^n}$
$H(X)$	entropy of the source X
$H(P_{x^n})$, H_{x^n}	empirical entropy of a sequence x^n
$\mathcal{D}(P\|\|Q)$	K-L divergence between P and Q
$\mathcal{D}(P_{x^n}\|\|P_{y^n})$	empirical K-L divergence between P_{x^n} and P_{y^n}
$d(x^n, y^n)$	distance between sequence x^n and y^n d_H, Hamming distance; d_{L_1}, L_1 distance; $d_{L_2^2}$, squared Euclidean distance, d_{L_∞}, maximum or infinite distance
$[s]_+$	$[s]_+ \stackrel{\triangle}{=} \max\{s, 0\}$
S	metric space
$d(x, A)$	distance of x from set A, $d(x, A) = \inf_{a\in A} d(x, a)$ $x \in S$, $A \subseteq S$
$\delta_A(B)$	distance of set B from A, $\delta_A(B) = \inf_{b\in B} d(b, A)$ $A, B \subseteq S$
$\delta_H(A, B)$	Hausdorff distance between A and B, $\delta_H = \min\{\delta_A(B), \delta_B(A)\}$
S^n_{XY}	transportation map which moves P_{x^n} (or S^n_X) into P_{y^n} (or S^n_Y)
S^n_{PQ}	transportation map which moves P into Q, $P \in \mathcal{P}_n$, $Q \in \mathcal{P}_n$
S^n_{PQ}	transportation map which moves $P \in \mathcal{P}$ into $Q \in \mathcal{P}$
$\mathcal{A}^n(L, P)$	set of admissible transportation maps that can be applied to $P \in \mathcal{P}_n$, for a maximum distance L
$\mathcal{A}(L, P)$	set of admissible transportation maps that can be applied to $P \in \mathcal{P}$, for a maximum distance L

$\mathcal{S}_D$ (res. $\mathcal{S}_A$)	set of strategies for the Defender (res. Attacker)
u	payoff of zero-sum game
H_0 (res. H_1)	null (res. alternative) hypothesis of the test
P_{FP}	Type I error or false positive error probability (probability of deciding in favor of H_1 when H_0 holds)
P_{FN}	Type II error or false negative error probability (probability of deciding in favor of H_0 when H_1 holds)
η	false positive error exponent, $\eta = -\limsup_{n\to\infty} \frac{\log P_{FP}}{n}$
ε	false negative error exponent, $\varepsilon = -\limsup_{n\to\infty} \frac{\log P_{FN}}{n}$
Λ^n	acceptance region of the test (decision in favor of H_0)
Λ	asymptotic acceptance region of the test ($n \to \infty$)
Γ^n	indistinguishability region of the adversarial test
Γ	indistinguishability region of the game
A_0	attack channel under H_0
A_1	attack channel under H_1
Q_X (res. Q_Y)	pmf at the output of the attack channel under H_0 (res. H_1)
$\mathcal{D}(H_i\|x^n)$	probability of deciding in favor of H_i given x^n, $i = 1, 2$
$a_n \doteq b_n$	asymptotic equality of sequences $\{a_n\}$ and $\{b_n\}$, $\lim_{n\to\infty} 1/n \log(a_n/b_n) = 0$
$a_n \dot{\leq} b_n$	asymptotic relation between $\{a_n\}$ and $\{b_n\}$, $\limsup_{n\to\infty} 1/n \log(a_n/b_n) \leq 0$

"Niente è più pratico di una buona teoria."
Kurt Lewin

Chapter 1

Introduction

"The increasingly wired nature of the world means cyberspace will likely be the world's next large battlefield (if it isn't already..)"
Jeremy Bender, February 2014,
Business Insider, UK

We live in a hyper-connected digital word in which every digital system is daily exposed to several threats. While sometimes these threats are innocent, and accidentally pursued against the system, many time they have a malicious purpose. Cybercrime is a growing industry whose annual damage to the global economy is estimated in many hundreds of billions and is expected to increase in the next years [1]. Many of such crimes may threaten nations' security and financial health. Issues surrounding these types of crimes have become high-profile, particularly those surrounding hacking, copyright infringement, child pornography, and child grooming [2]. There are also problems of privacy when confidential information is intercepted or disclosed, lawfully or otherwise. In such a digital 'minefied', protecting information is becoming of primary importance [3].

Security-oriented disciplines of signal processing have received increasing attention in the last decades; multimedia forensics, digital watermarking, steganography and steganalysis, biometrics, network intrusion detection, spam filtering, traffic monitoring, videosurveillance, are just a few examples. Despite enormous differences, all these fields are characterized by a unifying feature: the presence of one or more adversaries aiming at making the system fail. So far, the problem of copying with an adversary has been addressed by different communities with very limited interaction among them. It is not surprising, then, that similar solutions are re-invented several times, and that the same problems are faced again and again by ignoring that satisfactory solutions have already been discovered in contiguous fields. As a result, similar errors are repeated, e.g., the security problem is misunderstood. In watermarking, for instance, robustness and security have been treated as a unique problem [4] and it took several years to recognize that they are instead contrasting requirements calling for the adoption of different countermeasures. In a similar way,

security issues in biometric research are often neglected, privileging a pattern recognition perspective more related to robustness than security [5]. Similar concerns apply to several other fields. The lack of a unifying view makes also difficult to grasp the essence of the addressed problems and work out effective solutions.

While each adversarial scenario has its own peculiarities, there are some common and fundamental problems whose solution under a unified framework would speed up the understanding of the associated security problems and the development of effective and general solutions. It is relevant, then, to go beyond limited views and lay the basis of a general theory that takes into account the impact that the presence of an adversary has on the design of effective signal processing tools, i.e. a theory of *Adversarial Signal Processing (Adv-SP)*, a.k.a. *Adversary-aware Signal Processing*.

Driven by the need of developing a general theoretical framework to analyze adversarial problems in signal processing, in this thesis we move a first step in this direction, studying one of the most recurrent problems, namely *binary decision*. Binary decision is also known in literature as binary detection, since, in many applications, the decision problem pertains to the detection of the presence or absence of a certain phenomenon or signal (e.g., in radar detection, or in fingerprint detection). Moreover, being the decision framed as an Hypothesis Test, such problem is also sometimes referred to as *Binary Hypothesis Testing*. Among binary decision problems, source identification is one of the most studied subjects, since it lies at the heart of several security-oriented disciplines: multimedia forensics, when an analyst wants to distinguish which between two sources (e.g. a photo camera and a scanner); biometric authentication, where the authentication system must decide whether a biometric trait belongs to a certain individual, anomaly detection, traffic monitoring, steganalysis and so on .

With the above ideas in mind, the first part of the thesis is devoted to the development of a general theoretical framework for the binary detection problem in the presence of an adversary. We start from the observation that in order to make a correct decision in an hostile environment, we need to model the interaction between the decision function designer and the adversary, and then define the decision test and evaluate its performance within this framework. To do so, we cast the adversarial binary detection problem into a game theoretical framework, which is studied by relying on methods typical of information theory. We consider different possible decision scenarios (detection based on a single-observation, or decision fusion based on multiple-observations) which provide frameworks for different application scenarios (e.g. fingerprinting or biometric authentication for the single observation case, sensor networks and cognitive radio networks for decision fusion of multiple observations). We consider different variants of the game depending on the behavior of the

adversary (who may act under only one or both hypotheses under test), the decision setup (Neyman-Pearson, Bayesian), the a priori knowledge that the designer and the adversary have about the phenomenon under investigation (perfect knowledge of the statistical characterization of the system or partial knowledge based on observables) and finally, the possibility for the attacker to interfere with the learning phase by corrupting the training set (observables) used by the designer to make the decision. Drawing a parallelism with optimal transport theory, this thesis also contributes to the definition of a measure of statistical distinguishability of information sources under adversarial conditions, namely the Security Margin, which summarizes in a single quantity the expected result of the game. In order to make the analysis tractable, we focus on an asymptotic versions of the studied problems (with respect to the length of the observed samples) and confine our analysis to the case in which the decision is based of first order statistics.

The use of game theory to model the impact that the presence of an adversary has on Hypothesis Testing is not an absolute novelty. In many security oriented fields in which Hypothesis Testing plays a central role, game theory has been advocated to avoid entering the so called 'cat & mouse' loop in which researchers alternatively play the role of the designer and the adversary, and continuously develop new countermeasures, each time by attacking a specific algorithm or strategy. Despite isolated works in the various fields, a game theoretical formulation that permits to cast under a unique umbrella all the similar versions of the binary decision problem encountered in different applications is still missing.

Although the results of the first part of the thesis are general ones, we often use Multimedia Forensics as leading example to motivate the analysis. The unknowledgeable reader may refer to the introduction to this topic provided in the second part of the thesis.

In the second part of the thesis, we focus on the application of the theoretical findings to some selected problems in multimedia forensics. Multimedia Forensics is a rather mature discipline which pertains to the study of the techniques aimed at gathering information on the 'history' of a multimedia document (e.g. an image, a video, an audio track), i.e., on its origin, the processing it has undergone and its authenticity. The need for such tools is the natural consequence of the widespread diffusion of digital content, which anyone can modify, manipulate and distribute almost effortlessly. It is not surprising, then, that restoring the credibility of digital content has become a task of paramount importance. By playing the role of the counterfeiter, we develop a *universal* counter-forensic attack: as long as first order statistics are concerned, our proposed technique can be adopted to counter any forensic detector, that is, to conceal the traces left by *any* processing tool. We test our

algorithm in the pixel domain, to counter the detection of the contrast enhancement, i.e., a common operation which increases the image contrast to improve image quality. We then adapt our attack to make it work in the frequency domain for countering the detection of multiple JPEG compressions. Within this framework, we test the effectiveness of our method against state-of-the-art detectors. Besides, we evaluate experimentally the theoretical quantities which, according to the theory, summarize the ultimate achievable performance of the analysis.

1.1 Overview and contribution

This thesis is organized in two parts. Before starting with the first part, in Chapter 2, we introduce the reader to Adversarial Signal Processing and provide the basic tools for studying it. We also give an overview of the general structure of the adversary-aware binary detection problems addressed in the thesis.

In the first part of the thesis we develop the theoretical framework for the study of several binary detection problems in the presence of an adversary. In Chapter 3, we define and study the first simple case of binary detection when the statistical characterization of the observed system is known to both the decision function designer, referred to as the Defender, and the adversary, namely the Attacker. In the considered scenario, the Attacker is active only under one of the two hypotheses (one-side attack scenario). Such analysis is extended in Chapter 4 to the case in which the statistics of the system are known through training data. Chapter 5 is devoted to the analysis of the final achievable performance of the games studied in the first two chapters and to the definition of the concept of Security Margin. Then, in Chapter 6, we generalize the analysis of the adversarial setup studied in Chapter 4 by considering a version of the game in which the adversary can corrupt part of the training data available to the Defender. A different scenario in which the decision is based on multiple observations is addressed in Chapter 7. Finally, the case in which the Attacker is active under both hypotheses (two-side attack) is considered in Chapter 8, where two versions of the game are analyzed.

The second part of the thesis is devoted to the application of the theoretical findings to some practical problems in Multimedia Forensics. After a brief introduction to Multimedia Forensics and Counter-forensics in Chapter 9, in Chapter 10 we take the role of the counterfeiter and develop a universal counter-forensic attack against first order based detectors, i.e., detectors based on the analysis of the image histogram. Our universal scheme is extended to the frequency domain in Chapter 11 and used to counter the detection of multiple JPEG compressions. With reference to such forensic applications, in Chapter 12 we evaluate the Security Margin in practice. Chapter 13 concludes the thesis, summarizing the lessons learned and outlining

a possible path for future research.

1.2 Activity within research projects

The activity of this thesis has been partly carried out within the REWIND European projects, funded by the European Commission under FP7-FET programme and the AMULET project, funded by the European Office of Aerospace Research and Development (EOARD).

The REWIND (*REVerse engineering of audio-Visual content Data*) project[1], ended in June 2014, supported the activity presented in this thesis. The aim of the project was to develop new theories and tools for investigating the digital history of multimedia contents by synergistically combining principles of signal processing, machine learning and information theory. The REWIND project successfully reached its objectives, and we are proud of having contributed to its success. This research project was essential especially for the development of the second part of this thesis. Moreover, the project gave me the opportunity to establish contacts and fruitful collaborations, with the Signal Processing in Communications Group, University of Vigo (Spain), where the author of this thesis worked, as a visiting student, from October 2014 to February 2015.

We worked also on the AMULET (*A MUlti-cLuE approach To image forensics*) project, ended in December 2014. The project focused on the development of new techniques for multi-clue forensic analysis that, starting from the indications provided by a pool of tools thought to detect the presence of specific artifacts, make a global decision about the authenticity of a given image. This thesis contributed to this goal with the results presented in Chapter 7.

1.3 Publications

The activity of this thesis resulted into the following publications.

Chapter 3:

M.Barni, **B. Tondi**. "The Source Identification Game: an Information-Theoretic Perspective", *IEEE Transactions on Information Forensics and Security, Vol. 8, no. 3, pp 450-463, March 2013.*

[1] http://www.rewindproject.eu/

Chapter 4:

M. Barni, **B. Tondi**. "Optimum Forensic and Counter-forensic Strategies for Source Identification with Training Data". *In Proc. of IEEE International Workshop on Information Forensics and Security*, WIFS 2012.

M. Barni, **B. Tondi**. "Binary Hypothesis Testing Game with Training Data", *IEEE Transactions on Information Theory, Vol. 60, no. 8, pp 4848 - 4866, Aug. 2014.*

Chapter 5 (and 12):

M. Barni, **B. Tondi**. "The Security Margin: a Measure of Source Distinguishability under Adversarial Conditions". *Proc. of GlobalSip'13, IEEE Global Conference on Signal and Information Processing, 3-5 December 2013, Austin, Texas.*

M. Barni, **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, Oct. 2016.*

Chapter 6:

M. Barni, **B. Tondi**. "Source Distinguishability under Corrupted Training". *Proc. of WIFS'14, IEEE International Workshop on Information Forensics and Security, 3-5 December 2014, Atlanta, Georgia.*

Chapter 7:

M. Barni, **B. Tondi**. "Multiple-Observation Hypothesis Testing under Adversarial Conditions". *Proc. of WIFS'13, IEEE International Workshop on Information Forensics and Security, 18-21 November 2013, Guangzhou, China.*

Chapter 8:

B. Tondi, M. Barni, N. Merhav. "Detection Games with a Fully Active Attacker" *WIFS'15, IEEE International Workshop on Information Forensics and Security (WIFS), 16-19 Nov. 2015, Rome, Italy.*

Chapter 10:

M. Barni, M. Fontani, **B. Tondi**. "A Universal Technique to Hide Traces of Histogram-Based Image Manipulations". *In Proc. of the 14th ACM workshop on Multimedia and Security*, MMSEC 2012.

M. Barni, M. Fontani, **B. Tondi**. "A Universal Attack Against Histogram-Based Image Forensics", *International Journal of Digital Crime and Forensics (IJDCF), IGI Global, USA, Vol. 5, no. 3, 2013.*

Chapter 11:

M. Barni, M. Fontani, **B. Tondi**. "Universal Counterforensics of Multiple Compressed JPEG Images". *IWDW 2014, The 13th International Workshop on Digital-forensics and Watermarking, October 01-04, 2014, Taipei, Taiwan.*

The author of this thesis also contributed to the publications listed below, which are not included in this thesis.

F. Pérez-González, P. Comesaña-Alfaro, M. Barni, **B. Tondi**. "Are you threatening me?: Towards smart detectors in watermarking". *IS&T/SPIE Electronic Imaging 2014, 2-6 February 2014, San Francisco, California, United States.*

A. Abrardo, M. Barni, K. Kallas, **B. Tondi**. "Decision fusion with corrupted reports in multi-sensor networks: A game-theoretic approach" *Proc.of CDC 2014, IEEE 53rd Annual Conference on Decision and Control (CDC), 15-17 Dec. 2014, Los Angeles, CA.*

B. Tondi, P. Comesaña-Alfaro, F. Pérez-González, M. Barni. "The Effectiveness of the Meta-Detection for Countering Oracle Attacks in Watermarking" *WIFS'15, IEEE International Workshop on Information Forensics and Security (WIFS), 16-19 Nov. 2015, Rome, Italy.*

B. Tondi, P. Comesaña-Alfaro, F. Pérez-González, M. Barni "Smart Detection of Line Search Oracle Attacks", *IEEE Transactions on Information Forensics and Security on March, 2016 (under major revision).*

A. Abrardo, M. Barni, K. Kallas, **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, June 2016.*

K. Kallas, **B. Tondi**, M. Barni. "Consensus Algorithm with Censored Data for Distributed Detection with Corrupted Measurements: *A Game-Theoretic Approach", 2016 Conference on Decision and Game Theory for Security (GameSec), 2-4 November, 2016, New York, USA.*

M. Barni, Z. Chen, **B. Tondi**. "Adversary-aware, data-driven detection of double JPEG compression: how to make counter-forensics harder", *submitted to 2016 IEEE International Workshop on Information Forensics and Security (WIFS), 4-7 December, 2016, Abu Dhabi, UAE.*

1.4 Awards

The work behind this thesis led to the following awards:

- Best Student Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS), December 3-5, 2014, Atlanta, Georgia, USA
- Best Paper Award at the IEEE International Workshop on Information Forensics and Security (WIFS), November 16-19, 2015, Rome, Italy

1.5 Acknowledgements

I would like to express my gratitude to my advisor Prof. Mauro Barni for his guidance since the very beginning of my academic studies, for his motivation, enthusiasm and knowledge. I also thank him for the continuous support to my research and for his efforts to make me a good researcher. Besides my advisor, I wish to thank my thesis reviewers and members of the committee, Prof. Rainer Böhme (University of Münster) and Prof. Svyatoslav Voloshynovskiy (University of Geneva), for their insightful comments and suggestions. A sincere thank goes to Prof. Pedro Comesaña Alfaro and Fernando Pérez González (University of Vigo); working with them during my stay in Vigo has been a great and enriching experience. Moreover, I want to acknowledge the REWIND European projects for its financial support during the Ph.D course, and for making possible the fruitful collaboration with very professional and friendly researchers. Last but not least, a special thank goes to Prof. Neri Merhav (Technion, Israel Institute of Technology) for his contribution to my academic growth.

Chapter 2

Introduction to Adversarial Signal Processing

"If you know the enemy and know yourself,
you need not fear the result of a hundred battles.
If you know yourself but not the enemy,
for every victory gained you will also suffer a defeat.
If you know neither the enemy nor yourself,
you will succumb in every battle."
Sun Tzu, The Art of War

Adversarial Signal Processing (Adv-SP) is an emerging discipline targeting the study of signal processing techniques explicitly thought to withstand the intentional attacks of one or more adversaries aiming at system failure. The aim of AdvSP is then modeling the interplay between a Defender, wishing to carry out a certain processing task, and an Attacker, aiming at impeding it. Adv-SP methods can be applied to a wide variety of security-oriented applications including multimedia forensics, biometrics, digital watermarking, steganography and steganalysis, network intrusion detection, traffic monitoring, video-surveillance, just to mention a few [6].

While in classical signal processing, a designer carries out a certain processing task without any interference, in Adversarial Signal Processing, the Defender carries out its processing task in the presence of an Attacker, aiming at impeding it. Within such a framework, the classical signal processing theory can no longer be applied to model the problem: being the processing carried out in an adversarial environment, in fact, we must account for the presence of two players who interact each other. A quite natural framework for studying Adversarial Signal Processing is then provided by Game Theory. A brief introduction to Game Theory and to the main concepts necessary for deriving the results of this thesis is provided in section 2.2. After that, in Section 2.3 we give the basics of Adversarial Hypothesis Testing and introduce the various setups studied in this thesis.

Before doing that, in the following section, we provide a review of the most recent advances in the security-oriented fields where signal processing designers must cope with the presence of an adversary, highlighting the similarities between some existing approaches in the various fields, thus motivating the interest in building a unitary view of the most recurrent and basic problems.

2.1 Copying with an adversary: prior art

Copying with an enemy is a common problem to many signal processing fields and in particular to all security oriented disciplines.

In the various communities, researchers have already started to face this problem: adversarial machine learning [7], watermarking [8, 4], steganography and steganalysis [9, 10], biometric spoofing [11], traffic analysis [12] and intrusion detection [13] are among the most popular ones. Examples from other fields include security of reputation systems [14], cognitive radio [15], content based information retrieval [16], and many others.

In some cases, researchers are aware of the challenges set by the presence of *intentional* or malicius attacks (with respect to the *unintentional* ones, e.g. noise addition) and have started addressing them. In other cases, such awareness has still to be fully developed and the presence of an adversary is properly treated only in some scattered works.

For instance, in adversarial machine learning, the problem of classifier security in adversarial environment have attracted a growing interest [17, 18]. However, as claimed in [19], the related issues have only been sparsely addressed, under different perspectives and to a limited extent. Most of the works has focused on application-specific issues related to spam filtering [20] and network intrusion detection (e.g., evasion attacks [21, 22]) while the problem of adversarial classification has started to be addressed in a more systematic way only in a few works [23, 24, 7]. In particular, in [23], the classification problem is viewed as a game between the classifier and the adversary (intruder), although the unrealistic assumption of perfect knowledge of the classifier is made. Starting from this work, game theoretical approaches have been proposed which go beyond this assumption and study the strategic interaction: for instance, the intruder classification games in [25, 26, 27], or the game between a spam filter and a spammer studied in [28].

Steganography and steganalysis is another field where the resilience of the embedding scheme to attacks have been addressed in a large number of publications. However, a clear distinction between the concept of security and robustness is still missing in many works [29]. Several definitions of security have been given: an information-theoretic definition is given by Cachin [9], based on the K-L divergence

between cover and stego. Another related and recurrent notion is that of provably secure steganography [10]. However, provable secure models are difficult to put into practice and how to prove the security of practical steganographic schemes is still unclear [30]. Model-based approaches take into account these theoretical concepts in the design of practical steganographic algorithms; however, it has been shown that such approaches are flawed against strategic adversaries [31]. Then, it is becoming common sense that the security should be evaluated against an adversary (steganalyzer) who anticipates the behavior of the embedder.

There is a branch of research in the field, where game theoretical approaches to steganography have been explored and the problem of designing secure steganographic algorithms addressed more systematically. The first work in which steganography and game theory are combined traces back to 1998 [32]. In such a work the author proposes to use zero-sum games to model the contest between a data-hider and an attacker. There, the purpose of the active attacker is not only to detect, but to suppress hidden communication and then is subject to a distortion constraint. In [33], game theory is used to find best strategies for a steganographer who can spread the secret message over several homogeneous cover media (batch steganography), and a steganalyst who anticipates this and tries to detect the existence of at least one secret message (pooled steganalysis). Some interesting game theoretical approaches have been proposed recently in the field of content-adaptive steganography. Content-adaptive steganographic schemes embed the stego-message in the locations of the cover medium which are most suitable for embedding [34], i.e., where the changes are (supposed to be, according to conventional wisdom) harder to detect. Schöttle et al. [35] have recently drawn the attention to the fact that, if the steganalyzer behaves in a strategic manner (and then can recalculate the adaptivity criterion) adaptive embedding schemes are less secure than random embedding. In [35] the authors provide a rigorous approach to secure content-adaptive steganography by means of a game theoretic model: defender and attacker must decide in which position to hide and look for evidence of embedding, respectively, by taking into account the opponent's action. Using the notion of Nash equilibrium, an optimal adaptive embedding strategy which maximizes the security against a strategic detector is identified in a simple case. The simple model is later extended in [36]. Furthermore, in [37], the approach in [35] is applied to the case of Gaussian cover and embedding changes using LSB matching.

Game theory and information theory have also been used in watermarking to model the interplay between the watermarker and the attacker. In [38, 39, 40], the game is played between the watermark embedder/decoder and an attacker who attempts to degrade the embedded message by modifying the watermarked signal, e.g. by adding some noise. The payoff of the game is usually the capacity of the

watermark channel.

Multimedia Forensics is a relatively novel research field and then less mature in thought. Prior to our contributions (see [41, 42, 43], which are included as part of this thesis), attempts to define a general framework to account for the presence of the adversary have been made only recently in [44], where the validity of an attack is assessed regardless of the adopted countermeasures. We refer to Chapter 9 for a more extensive discussion.

It is worth mentioning that the need for adversarial modeling is becoming evident in many other security-related applications and game theory is often advocated as a possible useful tool [45]. For this reason, game theory is gaining popularity in all these areas: among these we mention network security [46, 47, 48], risk control [49] and cybersecurity [50].

2.1.1 Binary decision in the presence of adversary

In the above security-oriented fields, there are some common problems whose solution under a unified umbrella would speed up the understanding of the associated security problems and the development of effective solutions.

The most prominent of these problems is Binary Detection or Hypothesis Testing. Below, we list some examples of binary detection problems in adversarial environments from the various fields.

In Multimedia Forensics [51], a forensic analyst may be asked to decide whether an image has been acquired by a given camera, notwithstanding the presence of an adversary aiming at deleting the acquisition traces left by the camera. Similarly, the analyst may be asked to decide whether a signal has undergone a certain processing or not, by taking into account the possibility that someone deliberately tried to delete the traces left by the processing. Another popular example comes from spam filtering [23], wherein an anti-spam filter is presented with a test e-mail and must decide whether the e-mail contains a genuine or a spam message. It is obvious that such a test can not neglect the presence of an adversary trying to shape the message in such a way to fool the filter. Biometric authentication provides a further example. In this case, the authentication system must decide whether a biometric template belongs to a certain individual, despite the opposite efforts of an attacker aiming at building a fake template that passes the authentication test [11, 52]. Yet another example is watermarking, where the detector is asked to decide whether a document contains a given watermark or not, possibly in presence of an attacker aimed at injecting or removing the watermark from the content [53], or cognitive radio [15, 54], where the system has to decide if the spectrum is free or busy for transmission based on the information collected from many users which may be interested in gaining usage of

the spectrum resources. Other possible examples include: steganalysis, in which the steganalyzer has to distinguish between cover and stego images [32, 55], network intrusion detection, wherein anomalous traffic conditions must be distinguished from normal ones [13], reputation systems [14], for which it is essential to distinguish between genuine and malevolent scores.

A closer look reveals that a similar rationale exists behind some of the most popular techniques developed so far. This is the case for instance of the oracle attacks, i.e., attacks based on the information gathered by repeatedly querying the detector. In watermarking applications, oracle attacks have been successfully used to remove the watermark from watermarked contents, and/or illegally introduce the watermark in non-watermarked contents. The most popular cases are the sensitivity [56, 57] and BNSA attacks [58]. Hill-climbing attacks in biometric recognition systems are other example of oracle attacks [59, 60, 61] together with similar attacks in spam filtering and intrusion detection [62, 22], and, more in general, the ACRE attack in machine learning [63]. Active attackers that use a decoding oracle to detect the presence of hidden messages are also encountered in steganography [64].

Countermeasures also rely on similar approaches, starting from classical security by obscurity mechanisms, in which the access to the detector is denied to the attacker, to more sophisticated approaches like detector randomization [65, 66, 7], or the adoption of complicated detection regions [66, 67]. In addition to these early countermeasures, which are easily deflectable by the attackers and then poorly effective, researchers have also started to address the problem in a more systematic way [68, 69, 70]. Then, finding countermeasures against malicious attacks that query the classifier to gather information useful for the attack is becoming a common need in security-oriented applications. A novel direction for counteracting oracle attacks rely on the concept of *smart detectors* and has been recently explored in [71] and [72]. A smart detector is defined as a detector that is able to *learn from* and *react to* repeated query attacks. Notice that detectors producing a random output close to the decision boundary are not smart according to the previous definition, because they are not able to determine whether they are being attacked. To learn whether the system is being subject to an oracle attack, a *metadetector* is proposed that works at a higher level than the primary detector. While the operation of the latter is not modified, the former is specifically devoted to detect malicious queries and its definition is not affected by the specific purpose of the primary detector. Once the smart detector decides that an oracle attack is ongoing, effective countermeasures can be enforced, including the prevention of further accesses to the detector (banning), or the conservative switch to a more convoluted detection function. It is worth stressing that, although [71] and [72] consider watermark detection as motivating example, the

proposed metadetectors are higher-level detectors that can be applied to any binary decision problem where oracle attacks can be a threat, i.e., regardless of the underlying or primary detection problem. In this way, these works represent a first attempt to cast the problem of the oracle attacks under a unified framework.

A possible generalization of the basic binary decision problem regards the number of involved attackers. In attacks against reputation systems, for instance, several attackers may pool to degrade the performance of the system [73], leading to a multiple-player game. A similar situation is encountered in traitor tracing systems [74], with the noticeable difference that in this case active techniques like fingerprinting may be used to improve the performance of the system. Attacks against reputation systems introduce yet another perspective into the picture: the collaborative nature of the to-be-performed tasks and the attacks. In addition to the presence of multiple players, this requires that proper solutions are adopted to either encourage fair behaviors, e.g. through the definition of a suitable pay-off function, or to allow cross-checking between users, as commonly done in sentiment tagging applications [75]. In these cases, the presence of a large number of independent users, with a vast majority of fair users, ensures the proper behavior of the system.

Concerning the connection between adversarial binary detection and game theory, it is worth mentioning the relatively novel and interesting field of application of the *inspection games* [76]. An inspection game models a situation where an 'inspector' verifies that another party, called 'inspectee', adheres to certain legal rules, which found their main applications in arms control, economics and crime control. However, inspection games are also seen as a way to extend the classical statistical decision problem, when the distribution of the random variable is strategically controlled by another player, namely the 'inspectee'. The 'inspectee' can behave either legally or illegally, in which case he also chooses a violation procedure. Then, the statistician, namely the inspector, has to decide between the two cases. The field of the inspection games provide useful tools for handling practical problems that can be modeled in the above way. Inspection games are expected to play an interesting role in the study of some problems of adversarial signal processing. For example, they could be used for modeling the problem of the oracle attacks (see Section 2.1) in the practical scenarios where the detector works as an oracle and then an attacker may query the detector to learn the parameter of the systems.

2.2 Basics of Game Theory

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 interdepen-

dent, 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 a counterpart of *classical decision theory*.

Although examples of games occurred long before, the birth of modern Game Theory as a unique field traces back to 1944, with the book "Theory of Games and Economic Behavior" by John von Neumann and Oskar Morgenstern [77]. 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 over the outcomes of the game [1]. An intelligent player is able to act in a rational way and then always chooses the action which gives him the outcome he prefers most, given what he expects his opponents to do (his expectation on the other players). The goal of game-theoretic analysis is then to predict how the game will be played by rational players, or, relatedly, to give advice on how to play the game against rational opponents.

The models of Game Theory are highly abstract representations of classes of real-life situation for which equilibrium solutions are suggested with interesting (desirable) properties. Game Theory encompasses a great variety of situations depending on the number of players, the way the players interact, the knowledge that a player has on the strategies adopted by the others, the deterministic or probabilistic nature of 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, in which case they behave as individual entities (modeled as non-cooperative games). Hybrid games contain cooperative and non-cooperative elements. For instance, coalitions of players are formed in a cooperative game, but they play in a non-cooperative fashion.

Another common classification 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. The

[1] Axioms of rationality, Von Neumnn-Morgenstern utility theorem [77].

games in strategic form are discussed in the next chapter. 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. The extensive form of a game indeed is an explicit, highly descriptive, representation of a number of important aspects, like the sequencing of players' possible moves, their choices at every decision point, 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 [78].

In this thesis, we focus on non-cooperative, 2-players, strategic games.

2.2.1 Strategic games

The strategic form (also called normal form) is the basic type of game 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.

For a 2-player interaction, a strategic game is defined as a 4-tuple $G(\mathcal{S}_1, \mathcal{S}_2, u_1, u_2)$, where $\mathcal{S}_1 = \{s_{1,1} \dots s_{1,n_1}\}$ and $\mathcal{S}_2 = \{s_{2,1} \dots s_{2,n_2}\}$ are the sets of strategies (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 $(s_{1,i}, s_{2,j})$ is called profile and corresponds to an outcome of the game. Games in strategic form are compactly represented by matrices, namely payoff matrices.

In a *strictly-competitive* game, also referred to as *zero-sum* game, 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. In the particular case of a zero-sum game, then, only one payoff function needs to be defined. The payoff of the game (generally indicated by u) can be defined by adopting the perspective of one of the two players, e.g., without loss of generality, $u_1 = u$, with the understanding that the payoff of the second player u_2 is equal to $-u$. In the most common formulation (game with perfect information), the sets $\mathcal{S}_1$, $\mathcal{S}_2$ and the payoff functions are assumed to be known to both players. 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. The above discussion assumes that the players play *pure* strategies; however, it can be extended to the case in which the players make random choices over their set of actions, i.e., they play mixed strategies. Specifically, a *mixed* strategy for a player is defined as a distribution of probability over its set of actions.

Nash equilibrium

Given a game, the determination of the best strategy that each player should follow to maximize its payoff is not an easy task, all the more that a profile that is optimum for both the 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 both players. While there are many definitions of equilibrium, the most famous and commonly adopted is the one by Nash [79, 80]. For the particular case of a 2-player game, a profile (s_{1,i^*}, s_{2,j^*}) is a Nash equilibrium if:

$$\begin{array}{ll} u_1((s_{1,i^*}, s_{2,j^*})) \geq u_1((s_{1,i}, s_{2,j^*})) & \forall s_{1,i} \in \mathcal{S}_1 \\ u_2((s_{1,i^*}, s_{2,j^*})) \geq u_2((s_{1,i^*}, s_{2,j})) & \forall s_{2,j} \in \mathcal{S}_2, \end{array} \tag{2.1}$$

where for a zero-sum game $u_2 = -u_1$. In practice, a profile is a Nash equilibrium if no player can improve its payoff by changing its strategy unilaterally. It is known that every strategic game with finite sets of strategies for the players has a (at least one) Nash equilibrium in mixed strategies.

The notion of Nash equilibrium captures a steady-state of the play of a strategic game; the process by which such steady-state is reached is not examined.

For strictly competitive games, Nash equilibria have interesting properties. Let G be a zero-sum game and (s_{1,i^*}, s_{2,j^*}) be a Nash equilibrium; then, s_{1,i^*} is a maxminimizer for player 1, that is s_{1,i^*} is the action that maximizes the payoff of player 1 in the worst case scenario, i.e., assuming that player 2 plays his most profitable strategy (corresponding to the most damaging action for player 1). Similarly, s_{2,j^*} is a maxminimizer for player 2. We also have that

$$\max_{\mathcal{S}_1} \min_{\mathcal{S}_2} u_1(s_{1,i}, s_{2,j}) = \min_{\mathcal{S}_2} \max_{\mathcal{S}_1} u_1(s_{1,i}, s_{2,j}) = u_1(s_{1,i^*}, s_{2,j^*}) \tag{2.2}$$

As a consequence of relation (2.2), if many equilibrium points exist, they all yield the same payoff. A known result asserts that, if the two players are allowed to take randomized strategies over their set of actions (mixed strategies), finding the Nash equilibria for the game corresponds to solving one of the two Linear Programming (LP) problems in (2.2), (Von Neumann's Minimax Theorem, 1928 [81]).

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 [80]. The concept of dominance-solvability is directly related to the notion of dominant and dominated strategies. In

particular, 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 corresponds to the largest payoff, no matter how the other player decides to play. 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 choice made by the other player. Formally, in the 2-players case, we say that strategy $s_{1,i}$ is *strictly dominated* by strategy $s_{1,k}$ for player 1 (or, equivalently, that strategy $s_{1,k}$ strictly dominates $s_{1,i}$) if

$$u_1(s_{1,k}, s_{2,j}) > u_1(s_{1,i}, s_{2,j}) \quad \forall s_{2,j} \in \mathcal{S}_2. \tag{2.3}$$

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

The recursive elimination of dominated strategies is a possible technique for solving games. The basic idea is the following: all the strategies that a player should definitely not take can be eliminated from the set of possible actions. By this view, the recursive elimination works as follows: in the first step, all the dominated strategies are removed from the set of available strategies, since no rational player would ever play 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 remaining game, and hence are eliminated. The process goes on until no dominated strategy exists for any player. A *rationalizable equilibrium* is any profile which survives the iterative elimination of dominated strategies [82, 83]. 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 Nash equilibrium, and its practical meaning easier to grasp [84]: in fact, under the assumption of rational (and intelligent) players, we can anticipate that the players will choose the strategies corresponding to the unique rationalizable equilibrium. We notice again that, whereas *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 exist for dominance solvable games. An interesting, related 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.

2.3 Adversarial Hypothesis Testing (Adv-HT)

Hypothesis Testing is a widely studied topic with applications in virtually all technological and scientific fields. In its most basic form, an analyst is asked to decide which among two hypotheses, usually referred to as null hypothesis (H_0) and alternative hypothesis (H_1), is true based on a set of observables. Several versions of the problem are obtained according to the knowledge that the analyst has on the statistics of the observables under the two hypotheses.

Due to its importance, Hypothesis Testing has been extensively studied and a solid theoretical framework has been built permitting to analyze and understand its many facets [85, 86]. In the last years, though, many applications have emerged in which Hypothesis Testing is given a new twist, due to the presence of an adversary aiming at making the test fail. In all these applications, the analyst cannot neglect the presence of one or more adversaries explicitly aiming at decision error and the attacking behavior must be taken into account when defining the test. The reader may refer to the discussion in Section 2.1.1 for a review of hypothesis testing problems from various security-oriented fields.

Below, we introduce the main concepts of Hypothesis Testing in classical decision theory, when no attacker is present. Then, we adapt the test to encompass the presence of an attacker aiming at impeding a correct decision. We also provide the reader with an overview of the hypothesis testing problems studied in the first part of the thesis and introduce some of the choices we made in the definition of the various setups, when casting the defender-attacker interaction into a rigorous framework.

2.3.1 Hypothesis Testing

Here we review the main concepts of Hypothesis Testing [86].

Let S be an observed system, whose probabilistic model is different under two hypotheses, namely H_0 and H_1. We wish to test hypothesis H_0 against H_1 based on n observations. We denote with x_i be the i-th outcome of the system, belonging to the alphabet of symbols $\mathcal{X}$, and with $x^n = (x_1, x_2, ..., x_n)$ the observed sequence. As a result of the test, $\mathcal{X}^n$ is partitioned into two complementary regions Λ and $\bar{\Lambda}$, such that for $x^n \in \Lambda$ the Defender decides in favor of H_0, while for $x^n \in \bar{\Lambda}$,[2] H_1 is preferred. We say that a Type-I error occurs if H_1 is chosen when H_0 holds. In the same way, we say that a Type-II error occurs when H_1 holds but H_0 is chosen. In the following, we will refer to Type-I errors as false positive errors (or false alarms) and to Type-II as false negative (or missed detection), and will indicate the probabilities of such events as P_{FP} and P_{FN} respectively. The motivation for such a terminology comes from

[2] Given a set A, notation $\bar{A}$ denotes the complementary set.

applications in which H_0 is seen as a standard situation and its rejection in favor of H_1 raises an alarm since something unusual happened. This is the case, for instance, in multimedia forensics applications, where H_0 corresponds to the hypothesis that x^n was produced by a legitimate source, and an alarm is raised whenever this is not the case; or in radar applications, where H_0 corresponds to the absence of the target. It goes without saying that our derivation remains valid even in different scenarios where the false positive and false negative terms may not be appropriate. In our analysis we are mainly interested in the asymptotic behavior of P_{FP} and P_{FN}. In particular, we define the false positive (η) and false negative (ε) error exponents as follows:[3]

$$\eta = -\limsup_{n\to\infty} \frac{\log P_{\text{FP}}}{n}; \quad \varepsilon = -\limsup_{n\to\infty} \frac{\log P_{\text{FN}}}{n}, \tag{2.4}$$

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

2.3.2 Hypothesis Testing in adversarial setup

We now define the hypothesis testing problem in the presence of an adversary aiming at impeding the correct decision. To do so, we must assume that an analyst and an adversary, to whom we will refer as Defender (D) and Attacker (A), face each other in rigorously defined contexts.

A schematic representation of the adversarial binary decision test in its most basic form is depicted in Figure 2.1. Given the test sequence z^n, D must decide whether it has been generated under hypothesis H_0 or H_1.

When the Neyman Pearson (NP) setup is considered [86], as in the classical scenario, the Defender must choose the decision regions Λ and $\bar{\Lambda}$ in such a way to ensure that the Type-I error probability is lower than a certain prescribed value. Regarding the specific goal of the Attacker, we distinguish between *one-side* attack, when A is active under one of the two hypothesis only, and *two-side* attack, when A acts under both hypotheses. In the one-side attack case, the Attacker takes a sequence y^n generated under one of the two hypothesis, usually H_1, and transforms it into a modified sequence z^n so that when presented with the modified sequence D still accepts H_0. In doing so, the Attacker has to respect a distortion constraint, limiting the amount of modifications that can be introduced into the sequence. In

[3] We remind that the limit superior of a sequence x^n is defined as:

$$\limsup_{n\to\infty} x_n = \lim_{n\to\infty} \left(\sup_{m\geq n} x_m \right).$$

Differently from the limit, the lim sup always exists.

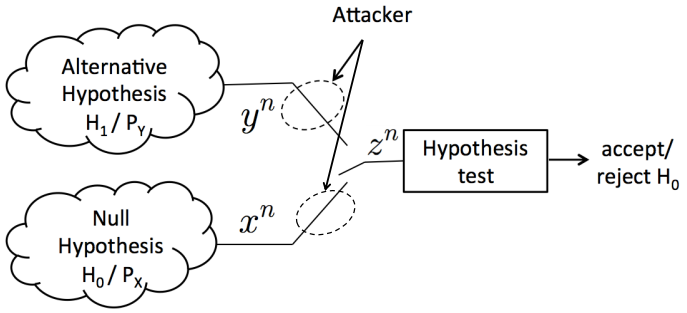


Figure 2.1: Basic adversarial decision setup considered in this thesis. P_X and P_Y denote the generation probabilities under H_0 and H_1 respectively.

such a scenario, the goal of the Attacker is causing a false negative decision error. Therefore, A aims at maximizing the Type-II error probability, while D's goal is to minimize it by taking into account the presence of the Attacker.

The scenario with one-side attack provides a suitable model for the decision problems found in many practical applications, when the adversary wants to pass a forgery off as an authentic signal (e.g. in biometric authentication, multimedia forensics, watermarking) and most of the analysis developed in this thesis focuses on this case. Clearly, in such situations, we implicitly assume that the Attacker knows the system status when he pursues the attack.

Given the general adversarial setup, depending on the knowledge that the Defender and the Attacker have on the statistical characterization under the two hypotheses, various versions of the problem can be defined. In the thesis, we studied both the case in which a full statistical knowledge of the system, i.e. of the *probability mass function* (pmf) underlying both hypotheses, is available to D and A (Chapter 3) and the case in which D and A know only some 'examples', i.e. sequences generated under the two hypotheses, namely training sequences (Chapter 4). Yet another variant is considered by assuming that the Attacker has the freedom of modifying, up to a certain extent, the training data available to D (Chapter 6). A completely different situation where the decision is based on multiple, possibly corrupted, observations is also investigated (Chapter 7).

In the two-side attack scenario, the goal of the Attacker is to mix up the decision, i.e. to cause a decision error of both types. This is a relevant for instance in applications of camera fingerprinting, where an attacker may be interested in removing the fingerprint from an image to hide the generating camera or implanting in the image the fingerprint of another camera to frame an innocent victim; another example comes from watermarking, where an attacker may want to remove the wa-

termark from an image or a video, to erase the information about the ownership, or to embed a watermark into the content, e.g., to redistribute a video with fake copyright. For the study of the detection problem in the two-side attack scenario, we consider two different setups: in the first setup, the Defender bases the decision on an adversary-aware Neyman-Pearson test; in the second one, a Bayesian approach is adopted, where the role of the two error probabilities is symmetrized, and the decision is based on the minimization of a Bayesian risk function. The analysis of the two side attack is limited to Chapter 8.

We analyze all the above variants, by adopting a game-theoretic approach, in which the Defender and the Attacker have opposite goals (zero-sum games) and operate by satisfying a different set of requirements, all together specifying the nature of the game. The final goal will be the derivation of the optimum strategies for the Defender and the Attacker in terms of game equilibrium points, and the study of the achievable performance at the equilibrium.

Part I

Theoretical Foundations of Adversarial Detection

Abstract

We theoretically study the binary detection problem in adversarial conditions. Many variants of the problem are considered depending on the decision setup, the attacking conditions and the knowledge that the Defender and the Attacker have about the sources and the status of the observed system. We focus mainly on the the scenario with one-side *attack, where the Attacker is active under one of the two hypotheses only, however the scenario with a* two sided *attack is also considered, where the Attacker acts under both hypotheses. Thanks to the adoption of a game-theoretic approach, under some limiting assumptions, namely the first order statistical analysis and the asymptotic setup, our analysis permits to derive the ultimate achievable performance when both players act rationally to maximize their payoff.*

Chapter 3

Detection Games with Known Sources

In this chapter we study the simple case of binary detection in a scenario with *one-side* attack, when both the Defender and the Attacker have full knowledge of the two sources.

First, we propose a rigorous framework based on game theory and information theory that can be used to model and analyze the interplay between the Defender and the Attacker, and we derive the equilibrium point of the game for some simple, yet meaningful cases. Specifically, we show that under certain assumptions on the set of strategies available to the Defender, the game admits an asymptotic rationalizable equilibrium, and derive the optimum strategies for the Defender and the Attacker at the equilibrium. As a second contribution, we analyze the asymptotic behavior of the payoff at the equilibrium. In this way we are able to distinguish the cases in which the Defender will succeed from those in which the Attacker will eventually win the game. The complete characterization of the game at the equilibrium is possible only by means of numerical analysis, except for some very simple cases in which the payoff at the equilibrium can be expressed in closed-form.

The chapter is structured as follows: first, we introduce the notation used throughout the thesis in Section 3.1; then, in Section 3.2 we provide a rigorous definition of the binary detection game with known sources. The game is solved by determining the equilibrium points in Section 3.3. Section 3.4 introduces an alternative and insightful characterization of the detection game by means of transportation theory, leading to an interesting interpretation of the optimum attack strategy (and of the outcome of the game). Then, in Section 3.5, we analyze the behavior of the payoff at the equilibrium. A closed form expression for the payoff is derived in the simple case of binary alphabet sources. In Section 5.3, we extend the analysis to the case in which the L_∞ metric is used in the definition of the Attacker's strategies (which is a case of particular interest in practical applications). The analysis is carried out for the case of memoryless sources; then, the chapter ends with some considerations on the extension of the analysis to the case of sources with memory, see Section 3.6.

3.1 Basic concepts, notation and definitions

In this section we summarize the notation and the definitions used in this chapter, to which we stick throughout the thesis. We also introduce some basic concepts of information theory that are needed to study the various versions of the detection problem.

We will use capital letters 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. Instantiations of random sequences are indicated by the corresponding lowercase letters, so x^n indicates a specific realization of the random sequence X^n, and X_i, x_i, $i = 1, ..., n$ indicate the $i-$th element of X^n and x^n, respectively. Information sources will also be defined by capital letters. The alphabet of an information source will be indicated 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 indicate, 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. Notation $X \sim P_X$ indicates that source X emits symbols according to P_X. Given an event A (be it a subset of $\mathcal{X}$ or $\mathcal{X}^n$), we will use the notation $P_X(A)$ to indicate the probability of the event A under the probability measure P_X. Given two sequences x^n and y^n, their *Hamming distance* is defined as the number of locations for which $x_i \neq y_i$, i.e.,

$$d_H(x^n, y^n) = n - \sum_{i=1}^{n} \delta(x_i, y_i), \tag{3.1}$$

with $\delta(x_i, y_i) = 1$ if $x_i = y_i$ and 0 otherwise.

Throughout the thesis, for a given quantity s, we adopt the following notation: $[s]_+ \stackrel{\triangle}{=} \max\{s, 0\}$. Equivalently, $[s]_+ = s$ if $s \geq 0$ and zero otherwise.

We also need to introduce the concept of distances between subsets and the definition of the *Hausdorff distance*, as a way to measure distance between subsets of a metric space [87]. Let S be a generic space and d a metric defined over S. For any point $x \in S$ and any non-empty subset $A \subseteq S$, the distance of x from the subset A is defined as:

$$d(x, A) = \inf_{a \in A} d(x, a). \tag{3.2}$$

Definition 1. *For any given pair (A, B) of subsets of S let us define $\delta_A(B) =$*

$\sup_{b\in B} d(b, A)$. *Let then* δ_H *be a function which associates to the pair of subsets* (A, B) *the quantity*

$$\delta_H(A, B) = \max\{\delta_A(B), \delta_B(A)\}. \tag{3.3}$$

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

It is worth observing that, according to the definition above, the Hausdorff distance is not a true metric, but only a pseudometric ($\delta(A, B) = 0$ implies that the closures of the sets coincide, namely $cl(A) = cl(B)$, but not necessarily that $A = B$). Then, in order for δ_H to be a metric we would need to restrict the definition to closed subsets.

Let $\mathcal{L}(S)$ denote the space of non-empty closed and limited subsets of S and let $\delta_H : \mathcal{L}(S) \times \mathcal{L}(S) \to [0, \infty)$. The assumption of boundedness of the sets [1] 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 [88]. We can give the following definition.

Definition 2. *Let* $\{K_n\}$ *be a sequence of closed and limited subsets of* (X, 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 limit in* $(\mathcal{L}(S), \delta_H)$ *and the limiting set is* K*.*

3.1.1 Basic information theory concepts

The mathematical machinery used to prove the main results of the thesis relies heavily on the methods of types [89, 90].

Then, throughout the thesis, we make extensive use of the concepts of type and type class defined in the following. Let x^n be a sequence with elements belonging to a discrete alphabet $\mathcal{X}$. The type P_{x^n} of x^n is the empirical probability distribution induced by the sequence x^n, i.e. $\forall a \in \mathcal{X}, P_{x^n}(a) = \frac{1}{n}\sum_{i=1}^{n} \delta(x_i, a)$.[2] In the following we indicate with $\mathcal{P}_n$ 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 indicate with $\mathcal{T}(P)$ 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 with $\mathcal{T}(P_{x^n})$, or simply $\mathcal{T}(x^n)$, the set of the sequences having the same type of x^n. Given a pair of sequences (x^n, y^n), 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) $P_{y^n|x^n}$.

For more insights into the use of type classes in information theory and statistics we refer to [90].

[1] We remind that the boundedness of the sets depends on the distance measure d defined in the metric space.

[2] P_{x^n} is often referred to as empirical probability distribution or type of the sequence x^n, for short.

Empirical distributions can be used to calculate empirical information theoretic quantities, thus the empirical entropy of a sequence will be denoted by:

$$H(P_{x^n}) = -\sum_{a\in\mathcal{X}} P_{x^n}(a)\log P_{x^n}(a), \tag{3.4}$$

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:

$$\mathcal{D}(P||Q) = \sum_{a\in\mathcal{X}} P(a)\log\frac{P(a)}{Q(a)}, \tag{3.5}$$

where, according to usual conventions, $0\log 0 = 0$ and $p\log p/0 = \infty$ if $p > 0$. When empirical distributions are considered, definition (3.5) is the empirical KL-divergence.

3.2 Definition of the detection game with known sources

3.2.1 Problem formulation

Let $\mathcal{C}$ be a class of the discrete memoryless sources (DMS), i.e., the class of the sources with finite alphabet $\mathcal{X}$ and let X and Y be two sources belonging to $\mathcal{C}$. Given their memoryless nature, the sources can be identified with their pmf's, respectively P_X and P_Y.

As already said in Section 2.3.2, the goal of D is the definition of a test to accept or reject the hypothesis that the sequence under analysis was generated by the source X. On the other hand, the goal of A is to take a sequence generated by Y and modify it in such a way that D accepts the hypothesis that the modified sequence has been generated by X. In doing so A may want to minimize the amount of modifications he has to introduce to deceive D.

In the scenario considered in this chapter, we assume that the probability measures P_X and P_Y ruling the emission of sequences by X and Y are known to both D and A. The assumption that the source Y is also known to D may seem a questionable choice in some practical applications, since it could be difficult for D to have full access to the source Y. We will see, however, that, at least asymptotically, the assumption that D knows Y can be removed, thus leading to a more realistic model. One may also argue that perfect knowledge of sources X and Y can never be reached

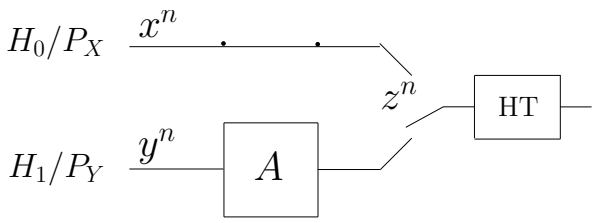


Figure 3.1: General scheme of the adversarial decision setup with one-side attack considered in Chapter 3 through 5.

in practice, yet we believe that the analysis of even this simplified version of the game can be extremely insightful and open the way to the analysis of more realistic and complex scenarios, like those studied in Chapters 4 and 6.

Let $x^n \in \mathcal{X}^n$, res. $y^n \in \mathcal{X}^n$, be a sequence drawn from X, res. Y and let $z^n \in \mathcal{X}^n$ denote the sequence observed by the Defender. With reference to the adversarial setup with one-side attack depicted in Figure 3.1, we have $z^n = x^n$ under H_0 (no attack occurs), whereas, under H_1, z^n is a modified version of y^n produced by the Attacker in the attempt to deceive the Defender.

3.2.2 The DT_{ks} game

We define the binary detection game with known source (DT_{ks}) as follows.

Definition 3. *The $DT_{ks}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a zero-sum, strategic, game played by D and A, defined by the following strategies and payoff.*

- *Defender's strategies.* By adopting a Neyman-Pearson approach, the set of strategies the Defender can choose from is the set of acceptance regions for H_0 for which the false positive probability is below a certain threshold:

$$\mathcal{S}_D = \{\Lambda^n : P_X(z^n \notin \Lambda^n) \leq P^*_{\text{FP}}\}, \tag{3.6}$$

  where Λ^n is the acceptance region for H_0 (similarly we indicate with $\bar{\Lambda}^n$ the rejection region for H_0) P^*_{FP} is a prescribed maximum false positive probability. The term $P_X(z^n \notin \Lambda^n)$ indicates the probability that a sequence generated by X does not belong to Λ^n, i.e., the false positive probability.

- *Attacker's strategies.* The set of strategies of 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:[3]

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n)) \leq nL\}, \tag{3.7}$$

where $d(\cdot,\cdot)$ is a proper distortion function and L is the maximum allowed average per-letter distortion.[4]

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

$$u(\Lambda^n, g) = -P_{\text{FN}} = -P_Y(z^n \in \Lambda^n) = -\sum_{y^n : g(y^n) \in \Lambda^n} P_Y(y^n), \tag{3.8}$$

  where the Defender's perspective is adopted, i.e., the Defender aims at maximizing u, while the Attacker wishes to minimize it.

Discussion

We pause to clarify some of the choices we made to formulate the DT_{ks} game.

First of all, we decided to limit the strategies available to A to deterministic functions of y^n. This may seem a limiting choice, however we will see that, at least asymptotically, the optimum strategy of D depends neither on the strategy chosen by A nor on P_Y, then, it does not make sense for A to adopt a randomized strategy to confuse D. Accordingly, everything would remain the same if we modeled the attack strategy in (3.7) as a channel, thus only complicating the notation.

The second comment regards the assumption that D knows P_Y. As it is evident from equation (3.8), 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. An alternative possibility could be to define the payoff under a worst case assumption on P_Y, however such a choice has two major drawbacks. First of all, if X and Y belong to the same class of sources $\mathcal{C}$, the worst case for D would always be $P_X = P_Y$, a condition under which no meaningful analysis can be made. One could require that X and Y belong to different source classes, however such classes should have to be known to D for a proper definition of the game, thus raising the same concerns raised by the assumption that D knows Y. Secondly, adopting a worst case analysis leads to the necessity of differentiating the payoffs of D and A, since for

[3] To avoid confusion between distortion and distance, we stress that, depending on the case, $d(\cdot,\cdot)$ may or may not be a distance (e.g., we will consider the case of measures L_1, L_2^2,...); that is why we prefer to refer to this quantity with the more general term of distortion.

[4] While L can be interpreted as the average per-letter distortion, A is not obliged to introduce a distortion that is lower than L for each sample of the sequence, since equation (3.7) defines only a global constraint.

D the worst case corresponds to the highest false negative error probability across all $Y \in \mathcal{C}$, while A knows P_Y and hence can compute the actual error probability. This observation would lead to the definition of a non-competitive version of the game in which two different payoffs are specified for D and A. In the sequel, we will focus on the asymptotic solution of the game for which the optimum strategy of D does not depend on P_Y, thus making the assumption that D knows P_Y irrelevant.

A last, even more basic, comment regards the overall structure of the game. Since the Attacker is interested in the false negative probability and does not intervene when H_0 holds (one side attack), his action has no impact on the false positive probability. In Chapter 8, we will consider a situation in which A modifies also the sequences generated under H_0 in the attempt to increase the false positive rate (two-side attack scenario). In this case, we will also depart from the Neyman-Pearson set up by considering the decision based on a Bayesian approach, and define the payoff in terms of the overall error probability.

3.2.3 DT_{ks} game with limited resources

Solving the DT_{ks} game as stated in Definition 3 is a cumbersome task, hence in this section we focus on the asymptotic optimum strategies that are obtained when the length n of the observed sequence tends to infinity. In order to make the problem tractable, we also limit the kind of acceptance regions D can choose from. We will do so by using an approach similar to that used in [91] to derive the optimal embedding and detection strategies for a general watermarking problem. Specifically, we limit the complexity of the analysis carried out by D by confining it to depend on a limited set of statistics computed on the test sequence. Given the memoryless nature of the sources, it makes sense to require that D bases its decision by relying only on P_{z^n}, i.e., on the empirical probability distribution induced by the test/observed 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, A could introduce some memory within the sequence as a result of the application of g. This is the reason why we need to introduce explicitly the requirement that D bases its decision only on the empirical distribution, that is, on first order statistics.

A fundamental consequence of this *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 density function of the sequences contained in it, we can redefine the acceptance region Λ^n as a union of types $P \in \mathcal{P}_n$, where $\mathcal{P}_n$ is the set of all possible types with denominator n.

With the above ideas in mind we can define the asymptotic DT_{ks}^{lr} game (where

lr stands for limited resources) as follows.

Definition 4. *The $DT_{ks}^{lr}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a game between D and A defined by the following strategies and payoff:*

$$\mathcal{S}_D = \{\Lambda^n \in 2^{\mathcal{P}_n} : P_{FP} \leq 2^{-\lambda n}\}, \tag{3.9}$$

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n)) \leq nL\}, \tag{3.10}$$

$$u(\Lambda^n, g) = -P_{FN}, \tag{3.11}$$

where in the definition of $\mathcal{S}_D$, $2^{\mathcal{P}_n}$ indicates the power set of $\mathcal{P}_n$, i.e., all the possible unions of types[5]. Note also that we now require that the false positive error probability decays exponentially fast with n, with exponent λ, thus opening the way to the asymptotic solution of the game.

As a final remark, we point out that so far we did not make any assumption on the distortion measure $d(\cdot, \cdot)$ adopted by the Attacker. However, we anticipate that for deriving some of the results of this chapter, that is, for computing the payoff at the equilibrium (Section 3.5), we will need to confine it to permutation-invariant distortion functions.

Since in this thesis we study only this version of the detection game with known sources, for the sake of readability, in the sequel we will omit the apex in the corresponding notation: then, from now on, the notation DT_{ks} will directly refer to the limited resources version of the game stated in Definition 4.

3.3 Solution of the DT_{ks} game

We start our derivation by proving the following lemma.

Lemma 1. *Let $\bar{\Lambda}^{n,*}$ be defined as follows:*

$$\bar{\Lambda}^{n,*} = \left\{ P \in \mathcal{P}_n : \mathcal{D}(P||P_X) \geq \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}, \tag{3.12}$$

and let $\Lambda^{n,*}$ be the corresponding acceptance region of the test.[6] *Then we have:*

1. *$P_{FP} \leq 2^{-n(\lambda - \delta_n)}$, with $\delta_n \to 0$ for $n \to \infty$,*

[5] In the rest of the chapter we will refer at Λ^n as a union of sequences or a union of types interchangeably, the two perspectives being equivalent and clearly understandable from the context.

[6] For convenience, sometimes the source pmf P_X and/or the threshold λ is made explicit in the notation for the acceptance region, and we refer to $\Lambda^{n,*}$ as $\bar{\Lambda}^{n,*}(P_X)$ or $\bar{\Lambda}^{n,*}(P_X, \lambda)$.

2. for every $\Lambda^n \in \mathcal{S}_D$ (with $\mathcal{S}_D$ defined as in (3.9)) we have $\bar{\Lambda}^n \subseteq \bar{\Lambda}^{n,*}$.

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

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

$$P_{\text{FP}}(\Lambda^{n,*}) = \sum_{P \in \bar{\Lambda}^{n,*}} P_X(\mathcal{T}(P)), \tag{3.13}$$

where $P_X(\mathcal{T}(P))$ indicates 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 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 [90] chapter 12), hence we have:

$$\begin{aligned} P_{\text{FP}}(\Lambda^{n,*}) &\leq (n+1)^{|\mathcal{X}|} \max_{P \in \bar{\Lambda}^{n,*}} P_X(\mathcal{T}(P)) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \bar{\Lambda}^{n,*}} \mathcal{D}(P||P_X)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\left(\lambda - |\mathcal{X}| \frac{\log(n+1)}{n}\right)} \\ &= 2^{-n\left(\lambda - 2|\mathcal{X}| \frac{\log(n+1)}{n}\right)}, \end{aligned} \tag{3.14}$$

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 (3.12).

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 [90] Chapter 12 for a justification of the last inequality):

$$\begin{aligned} 2^{-\lambda n} &\geq P_X(\bar{\Lambda}^n) \\ &\geq P_X(\mathcal{T}(P)) \\ &\geq \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P||P_X)}, \end{aligned} \tag{3.15}$$

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

The first relation proved in Lemma 1 says that, asymptotically, $\Lambda^{n,*}$ defines a valid strategy for D, while the second one implies the optimality of $\Lambda^{n,*}$. In fact, if for a certain strategy of 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 higher false negative error probability.

Some interesting consequences of the lemma are the following. The optimum strategy for the Defender does not depend on the strategy chosen by the Attacker. By adopting a game theoretic terminology, the best defence strategy is *dominant*, i.e., it is optimum regardless of the attacking strategy. As a further consequence, the optimum defence strategy does not depend on P_Y, meaning that the optimum 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 D knows P_Y irrelevant. In the same way, it is not necessary for D to know the probability distribution of the attacked sequences.

As a final notice, we observe that the strategy expressed by equation (3.12) has a simple heuristic interpretation: D will accept only the sequences whose empirical pmf is close enough, in divergence terms, to the known pmf of X: this result corresponds to the known Hoeffding test for the non-adversarial case [92].

We now pass to the determination of the optimum strategy of A. The existence of a dominant strategy for D significantly simplifies the search for the optimum attacking strategy. In fact, since a rationale Defender will surely play his dominant strategy $\Lambda^{n,*}$, A can choose her strategy by assuming that $\Lambda^n = \Lambda^{n,*}$. In this way, the derivation of the optimum attacking strategy becomes an easy task. By observing that the goal of 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:

$$\mathcal{D}(P_{g(y^n)}||P_X) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}. \tag{3.16}$$

In doing so A must only respect the constraint that $d(y^n, g(y^n)) \leq nL$. The optimum strategy for A can then be expressed as follows:[7]

$$g^*(y^n) = \arg \min_{z^n : d(z^n, y^n) \leq nL} \mathcal{D}(P_{z^n}||P_X). \tag{3.17}$$

Together with Lemma 1, the above observation permits to state our first fundamental result, summarized in the following theorem.

Theorem 1. (Equilibrium point of the DT_{ks} game). *The DT_{ks} game is a dominance solvable game and the profile $(\Lambda^{n,*}, g^*)$ is the only rationalizable equilibrium.*

Proof. Lemma 1 asserts that $\Lambda^{n,*}$ is a strictly dominant strategy for D, thus permitting us to eliminate all the other strategies in $\mathcal{S}_D$ (since they are strictly dominated by $\Lambda^{n,*}$). The theorem, then, follows by observing that g^* satisfies

$$-u(\Lambda^{n,*}, g^*) \geq -u(\Lambda^{n,*}, g) \quad \forall g \in \mathcal{S}_A, \tag{3.18}$$

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 (3.17) is not lower than the acceptance threshold, no other strategy will succeed in bringing y^n inside the acceptance region; hence, A maximizes the false negative probability, namely $P_Y(g(y^n) \in \Lambda^{n,*})$, by playing strategy g^*. □

[7] In principle, the minimization in (3.17) may have multiple solutions. However, for all the distortion functions considered in our analysis (see Section 3.4), the minimum is unique.

As a remark, we observe that, being a rationalizable equilibrium, profile $(\Lambda^{n,*}, g^*)$ has the desirable characteristic of being the only possible choice if the two players behave rationally. In fact, a rational Defender will surely adopt the acceptance region $\Lambda^{n,*}$, since any other choice will lead to a (asymptotically) higher P_{FN}, regardless of the choice made by A. On his side, a rational Attacker, knowing that D will behave rationally, will adopt the strategy g^* since this is the strategy that optimizes his payoff when D plays $\Lambda^{n,*}$.

It is worth noticing that, even if in the definition of the *DT* game the payoff corresponds to the average false negative error probability, the strategy defined by equation (3.17) represents the optimal attack that A can use for *each* sequence: if the minimization in (3.17) fails to bring a sequence y^n into $\Lambda^{n,*}$, any other attack will also fail.

As a final remark, we observe that, due to the universality of the defence strategy with respect to Y, with few modifications, Theorem 1 can be applied to a more general *composite hypothesis testing* scenario in which only the characterization under H_0 is known [86].

3.4 Characterization of the game by means of transportation theory

In deriving the results of the previous sections, we did not make any assumption on the distortion measure $d(\cdot, \cdot)$. However, as already said, in order to be able to compute the payoff of the game at the equilibrium, we will need to limit our analysis to the case of permutation-invariant distortion measures.[8] Since most of the commonly adopted distortions are permutation invariant, such a limitation is not a strict one. On the other hand, confining the analysis of the game to this class of distortion measures allows an interesting reformulation of the game, which is the purpose of this section. Specifically, we show that for the case of permutation-invariant distortion, we can look at the optimum Attacker's strategy from a different perspective, by drawing a parallelism with *optimal transport theory* [93].

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 formalised [94]. In its most ancient formulation, optimal transport theory deals with the problem of moving mass from a source location to a sink location by minimizing some cost function of the transportation per unit of mass. In one of its instance, OT searches for the transportation map that transforms a random variable with a given pmf into

[8] For any pair (y^n, z^n), a permutation invariant distortion d is such that $d(y^n, z^n) = d(\sigma(y^n), \sigma(z^n))$ for any permutation σ of the elements of the sequence.

another random variable with a different pmf, defined over the same alphabet, by minimizing the average cost of the transport, that can be interpreted as an expected distance between the variables.[9]

Drawing a parallelism with optimal transport theory allows us to derive a very intuitive and insightful interpretation of the optimum Attacker's strategy, opening the way to the analysis performed in Chapter 5.

Given a sequence y^n drawn from Y, the goal of A is to transform it into a sequence z^n belonging to the acceptance region chosen by D. Let us indicate 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 indicate 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 refer to S^n_{YZ} as *transportation map*. Once again, we explicitly indicate that S^n_{YZ} refers to n-long sequences by adding the superscript n. For any permutation-invariant distortion measure, the overall distortion introduced by the attack can be expressed in terms of S^n_{YZ}. In this thesis, we focus on distortion measures for which the average per-letter distortion between y^n and z^n can be written in the form

$$f(\{d(i,j)\}_{i=1}^{|\mathcal{X}|}, S^n_{YZ}),$$

where $d(i,j)$ is the distortion introduced when the symbol i is transformed into the symbol j and $f(\cdot)$ is an arbitrary function.

For example, for an *additive distortion measure*, we have $d(y^n, z^n) = \sum_i d(y_i, z_i) = \sum_{i,j} n(i,j)d(i,j)$, and hence the average per-symbol distortion depends only on S^n_{YZ}, i.e.

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

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

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

Finally, we observe that the Attacker can not change more symbols than there are in the sequence y^n; as a consequence a map S^n_{YZ} can be applied to a sequence y^n only if $S^n_Y(i) \triangleq \sum_j S^n_{YZ}(i,j) = P_{y^n}(i)$. The above reasoning suggests an interesting interpretation of S^n_{YZ}, which can be seen as the *joint empirical pmf* 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 of y^n and z^n.

[9] Further insights on optimal transportation can be found in Section 5.1.1.

By remembering that Λ^n depends only on the empirical pmf of the test sequence (i.e., on its type), and given that the empirical pmf of the attacked sequence depends on S_Z^n only through S_{YZ}^n, we can define the action of the Attacker as the choice of a transportation map among all *admissible* maps, a map being admissible if:

$$\begin{cases} S_Y^n = P_{y^n} \\ f(\{d(i,j)\}_{i=1}^{|\mathcal{X}|}, S_{YZ}^n) \leq L, \end{cases} \tag{3.19}$$

where, in the general case, the second condition expresses the average per-symbol distortion constraint the Attacker is subject to, and L is the maximum (average) allowable per-letter distortion. The set of the admissible maps is denoted by $\mathcal{A}^n(L, P_{y^n})$. For the case of additive distortion, the admissibility constraints can be rewritten as follows:

$$\begin{cases} S_Y^n = P_{y^n} \\ \sum_{i,j} S_{YZ}^n(i,j) d(i,j) \leq L. \end{cases} \tag{3.20}$$

Given the above definitions, 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 $S_{YZ}^n(y^n)$, or $S_{YZ}^n(i,j;y^n)$, 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, $S_Z^n(j;y^n)$ indicates the output marginal of $S_{YZ}^n(i,j;y^n)$. With regard to the input marginal, we always have $S_Y^n(i;y^n) = P_{y^n}(i)$.[10]. By adopting this symbolism, the space of strategies for the Attacker can be redefined as:

$$\mathcal{S}_A = \{S_{YZ}^n(y^n) : S_{YZ}^n(i,j) \in \mathcal{A}^n(L, P_{y^n})\}. \tag{3.21}$$

Accordingly, we can rewrite the payoff function, i.e., the opposite of the false negative probability of the test, as follows

$$u(\Lambda^n, S_{YZ}^n) = - \sum_{y^n : S_Z^n(y^n) \in \Lambda^n} P_Y(y^n). \tag{3.22}$$

By adopting the above transportation theory perspective, Theorem 1 can be rephrased as follows.

Corollary 1 (Equilibrium point of the DT_{ks} game)**.** *Let*

$$\Lambda^{n,*} = \left\{ P \in \mathcal{P}_n : \mathcal{D}(P||P_X) < \lambda - |\mathcal{X}| \frac{\log(n+1)}{n} \right\}, \tag{3.23}$$

[10]Similarly, we use notation $S_Y^n(y^n)$ to denote the pmf P_{y^n}.

and

$$S_{YZ}^{n,*}(y^n) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L,P_{y^n})} \mathcal{D}(S_Z^n||P_X). \tag{3.24}$$

Then $\Lambda^{n,*}$ is a dominant equilibrium for D and the profile $(\Lambda^{n,*}, S_{YZ}^{n,*}(y^n))$ is the only rationalizable equilibrium of the DT_{ks} game, which, then, is a dominance solvable game.

3.5 Analysis of the payoff at the equilibrium

The next step is the computation of the payoff at the equilibrium. Given the asymptotic nature of the solution we found, it makes sense to compute the asymptotic behavior of P_{FN} at the equilibrium. From the foregoing discussion it is easy to argue that P_{FN} will either tend to 0 or to 1 for $n \to \infty$ depending on the relationship between the maximum allowed distortion and the KL-divergence between P_X and P_Y. For a more accurate analysis, we will also evaluate the error exponent of the false negative error probability at the equilibrium[11]. Such evaluation will be carried out under the assumption that the set of admissible maps $\mathcal{A}$ is determined by a linear set of constraints, or equivalently, by assuming that the distortion measure d can be expressed as a *linear function* of the transportation map S_{YZ}^n (or equivalently, function f in (3.19) is linear in S_{YZ}^n). It is straightforward to see that for instance any additive distance measure meets this constraint.

Let us define Γ^n as the set of sequences generated by Y that can be moved into $\Lambda^{n,*}$ as a consequence of the attack. We can write:[12]

$$\Gamma^n(P_X, \lambda, L) = \{y^n : \exists\ z^n \in \Lambda^{n,*}(P_X, \lambda) \text{ s.t. } d(y^n, z^n) \leq nL\}. \tag{3.25}$$

Accordingly, the false negative error probability is equal to the probability that the sequence y^n belongs to this set, that is $P_{\text{FN}} = P_Y(y^n \in \Gamma^n)$. We observe that, under some very general assumptions, Γ^n is still a union of type classes.

Property 1. *The set $\Gamma^n(P_X, \lambda, L)$ defined in (3.25) is a union of type classes for any permutation invariant distance-measure.*

The above property 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 the Attacker is equivalent to the application

[11] We remind that, according to the Neyman-Pearson setup adopted, the false positive error exponent is always larger than or equal to λ (see (3.9)).

[12] We notice that when we write the constraint in the form $d(y^n, z^n) \leq nL$, we are implicitly assuming that an additive distortion measure is adopted.

of a transportation map $S_{YZ}^{n,*}(y^n)$. The set in (3.25) can then be easily redefined in terms of types instead of sequences: [13]

$$\Gamma^n(P_X, \lambda, L) = \{P \in \mathcal{P}_n : \exists\ S_{PV}^n \in \mathcal{A}^n(L, P) \text{ s.t. } V \in \Lambda^{n,*}(P_X, \lambda)\}. \tag{3.26}$$

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

We find convenient to introduce the asymptotic version of $\Gamma^n(P_X, \lambda, L)$, which is defined as follows:

$$\Gamma(P_X, \lambda, L) = \{P \in \mathcal{P} :\ \exists\ S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Lambda^*(P_X, \lambda)\}, \tag{3.27}$$

where

$$\Lambda^*(P_X, \lambda) = \{P \in \mathcal{P} : \mathcal{D}(P||P_X) \leq \lambda\}, \tag{3.28}$$

while 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.[14]

Theorem 2. (Asymptotic payoff of the DT_{ks} game at the equilibrium). *For the DT_{ks} game, the error exponent of the false negative error probability at the equilibrium is given by*[15]*:*

$$\varepsilon = \min_{P \in \Gamma(P_X, \lambda, L)} \mathcal{D}(P||P_Y), \tag{3.29}$$

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[16]

$$\varepsilon = -\lim_{n \to \infty} \frac{1}{n} \log\left(P_Y(P_n \in \Gamma^n)\right), \tag{3.30}$$

[13] 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 pmfs are not specified.

[14] For the definition of the error exponent of false negative and positive probability we remind to (2.4).

[15] The use of the minimum instead of the infimum is justified by the compactness of $\Gamma(P_X, \lambda, L)$ which will be actually demonstrated within the proof (the same in the following).

[16] We use directly the lim (and not the lim sup) because, as we will show, such limit exists.

namely, the error exponent of the probability of the sequence of sets Γ^n. The evaluation of the above limit can be carried out by applying the generalization of Sanov's theorem proven in Appendix A. In order to apply the theorem to our case, it is sufficient to show that, for a given distance measure $d : \mathcal{P} \times \mathcal{P} \to [0, \infty)$, $\Gamma^n \xrightarrow{H} \Gamma$, that is, Γ^n tends to Γ in the Hausdorff metric δ_H (see Corollary 4 in the appendix).

We first notice that, 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 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 which moves a distribution out of the acceptance region, that is

$$\Gamma(P_X, \lambda, L) = \{P \in \mathcal{P} : \exists S_{VP} \in \mathcal{A}(L, V), \text{for some } V \in \Lambda^*(P_X, \lambda)\}. \tag{3.31}$$

The equivalence of definition (3.31) and (3.27) 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[17]. A similar equivalence holds for 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 $S_{V_1 P_1} \in \mathcal{A}(L, V_1)$. 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 0 as $n \to \infty$, V_2 can be taken arbitrarily close to V_1 for large enough n. According to Theorem 25 (Appendix B), 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 25), for any given P_1 and $S_{V_1 P_1}$, the map $S^n_{P_2 V_2} \in \mathcal{A}^n(L, V_2)$ can be chosen in such a way that $P_2 \in B(P_1, e_n)$[18] 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 0, as $n \to \infty$, thus concluding the proof. □

The main consequence of Theorem 2 is that, given P_X, L and λ, the set of sources P_Y can be split into two distinct regions: the subset of sources for which the false

[17] 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 thesis.

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

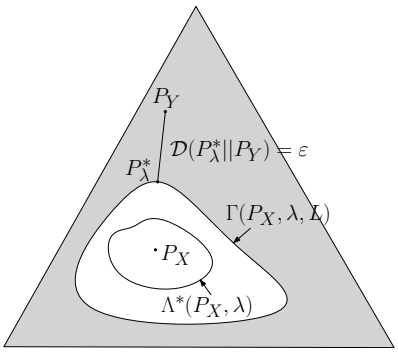


Figure 3.2: Geometric interpretation of $\Gamma(P_X, \lambda, L)$ and $\Lambda^*(P_X, \lambda)$ by the light of Theorem 2.

negative probability tends to zero exponentially fast ($P_Y \in \bar{\Gamma}(P_X, \lambda, L)$) and the sources for which, as a consequence of the attack, the false negative probability tends to 1. Stated in another way, given two pmf's P_X and P_Y, a maximum attacking distortion L and the desired false positive error exponent λ, Theorem 2 permits to understand whether D may ever succeed to make the false negative error probability vanishingly small and thus *win* the game. Then, $\Gamma(P_X, \lambda, L)$ 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, where by *reliably distinguished* we mean distinguished in such a way to grant a strictly positive error exponent for P_{FN}. Accordingly, $\Gamma(P_X, \lambda, L)$ represents the *indistinguishability region* of the adversarial detection test in the DT_{ks} setup. A geometric interpretation of Theorem 2 is given in Figure 3.2.

In general, the expression of Γ does not allow an analytic computation of the pmf's P_Y which the Defender 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.

3.5.1 Hamming distance and distinguishability of Bernoulli sources: a case study

In this section, 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. Given two sequences x^n and y^n, their Hamming distance $d_H(x^n, y^n)$ is defined as the number of locations for which $x_i \neq y_i$. It is easy to see that the Hamming distance is a particular case of additive distance for which the distance between a pair of alphabet symbols (i, j) is given by the one's complement of the

Kronecker delta,[19] namely, $d(i,j) = \overline{\delta}_{ij}$.

When the Hamming distance is considered, a closed-form expression can be found for Γ thus greatly simplifying the analysis. The simplification relies on the following lemma.

Lemma 2. *If $d(y^n, z^n) = d_H(y^n, z^n)$, the set Γ^n can be expressed as:*

$$\Gamma^{n,*}(P_X, \lambda, L) = \{P \in \mathcal{P}_n : \exists P' \in \Lambda^{n,*}(P_X, \lambda) \text{ s.t. } d_{L_1}(P, P') \leq 2L\} \tag{3.32}$$

where the L_1 distance between P and P' (sometimes called variational distance) is defined as:

$$d_{L_1}(P, P') = ||P - P'||_{L_1} = \sum_{a \in \mathcal{X}} |P(a) - P'(a)|. \tag{3.33}$$

Proof. We start by proving that a sequence whose type has a L_1 distance larger than $2L_H$ from all the types in $\Lambda^{n,*}$ cannot belong to Γ_H^n. 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:

$$\begin{aligned} ||P_{y^n} - P_{z^n}||_{L_1} &= \sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)] \\ &\quad + \sum_{a \in \mathcal{X}^-} [P_{z^n}(a) - P_{y^n}(a)] \\ &= 2 \sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P_{z^n}(a)], \end{aligned} \tag{3.34}$$

where $\mathcal{X}^+$ (res. $\mathcal{X}^-$, $\mathcal{X}^=$) indicates 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:

$$\sum_{a \in \mathcal{X}^-} P_{y^n}(a) = 1 - \sum_{a \in \mathcal{X}^+} P_{y^n}(a) - \sum_{a \in \mathcal{X}^=} P_{y^n}(a). \tag{3.35}$$

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:

$$||P_{y^n} - P_{z^n}||_{L_1} \leq \frac{2 d_H(y^n, z^n)}{n}. \tag{3.36}$$

[19] Given two variables i and j, the Kronecker delta δ_{ij} is equal to 1 if $i = j$, 0 otherwise.

For the sequences y^n whose type does not satisfy (3.32), we have $||P_{y^n} - P_{z^n}||_{L_1} > 2L_H \ \forall z^n \in \Lambda^{n,*}$, yielding

$$2L < ||P_{y^n} - P_{z^n}||_{L_1} \leq \frac{2d_H(y^n, z^n)}{n}, \tag{3.37}$$

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

$$\sum_{a \in \mathcal{X}^+} [P_{y^n}(a) - P'(a)] = \sum_{b \in \mathcal{X}^-} [P'(b) - P_{y^n}(b)]. \tag{3.38}$$

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:

$$\begin{aligned} d_H(y^n, z^n) &= \sum_{a \in \mathcal{X}^+} n[P_{y^n}(a) - P'(a)] \\ &= \frac{n||P_{y^n} - P'||_{L_1}}{2} \\ &\leq nL, \end{aligned} \tag{3.39}$$

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

Lemma 2 permits to rewrite the expression for the indistinguishability region in a simpler form:

$$\Gamma^* = \{P \in \mathcal{P} : \exists P' \in \Lambda_0^*(P_X) \text{ s.t. } d_{L_1}(P, P') \leq 2L\}. \tag{3.40}$$

The relation between Hamming distance and L_1 distance, investigated in the proof of the Lemma 2, will turn useful in other parts of the thesis (e.g. in Chapter 6).

Bernoulli sources

In order to exemplify the general concepts introduced in the previous section, we now apply them to the case of two Bernoulli sources. For the sequences emitted by these sources (binary alphabet sources), the Hamming distance is a natural choice to

define the distortion constraint, thus permitting to adopt the simplified definition of Γ given in (3.40).

Let X and Y be 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[20] function of $|\nu_x(1) - p|$, where we indicated with $\nu_x(1)$ 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:

$$\Lambda^{n,*}(p, \lambda) = \{P \in \mathcal{P}_n : P(1) \in (\nu_{inf}(\lambda), \nu_{sup}(\lambda))\}, \tag{3.41}$$

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

$$\mathcal{D}(P||P_X) = \lambda - |\mathcal{X}|\frac{\log(n+1)}{n}. \tag{3.42}$$

Note that in some cases we may have $\nu_{inf} = 0$ and/or $\nu_{sup} = 1$, since equation (3.42) may admit a solution only for $P(1) > p$, $P(1) < p$, or no solution at all.

The optimum strategy of A is also easy to define. Given the monotonic nature of the KL-divergence, 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. The Attacker 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 $d(y^n, z^n) \leq nL$, we clearly have:

$$\Gamma^n(p, \lambda, L) = \{P \in \mathcal{P}_n : P(1) \in (\nu_{inf}(\lambda) - L, \nu_{sup}(\lambda) + L)\}, \tag{3.43}$$

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 consider the asymptotic version of $\Lambda^{n,*}$ and Γ^n:

$$\Lambda^*(p, \lambda) = \{P \in \mathcal{P} : P(1) \in (\nu^\infty_{inf}(\lambda), \nu^\infty_{sup}(\lambda))\}, \tag{3.44}$$

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

$$\mathcal{D}(P||P_X) = \lambda; \tag{3.45}$$

and then the indistinguishability region is

$$\Gamma^*(p, \lambda, L) = \{P \in \mathcal{P} : P(1) \in [\nu^\infty_{inf}(\lambda) - L, \nu^\infty_{sup}(\lambda) + L]\}. \tag{3.46}$$

[20] 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 2, we can distinguish two cases:

$$\begin{aligned} q = P_Y(1) &\in [\nu_{inf}^{\infty}(\lambda) - L, \nu_{sup}^{\infty}(\lambda) + L] \\ q = P_Y(1) &\notin [\nu_{inf}^{\infty}(\lambda) - L, \nu_{sup}^{\infty}(\lambda) + L]. \end{aligned} \tag{3.47}$$

In the first case $\varepsilon = 0$. In the second case P_{FN} tends to 0 for $n \to \infty$ and the error exponent can be computed by resorting to equations (3.29) and (3.40). 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 $\varepsilon = \mathcal{D}(p^*||q)$.

3.5.2 Analysis of the game with L_∞ distance

In this section, we extend the analysis of the previous Sections 3.3–3.5 to the case in which the distortion measure constraining the Attacker is expressed in terms 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, thus calling for the adoption of near-lossless image coding schemes [95]. Another example in which the use of the L_∞ distance is recommended, is when it must be ensured that two versions of the same image, an original and a processed one, are visually indistinguishable. 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) [96]) 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 he can choose from, that is, to define the set of admissible maps in the form in (3.19). More specifically, we observe that the maximum distance between the sequences y^n and z^n can be rewritten as follows:

$$d_{L_\infty}(y^n, z^n) = \max_k |z_k - y_k| = \max_{(i,j):S_{YZ}^n(i,j)\neq 0} |i - j|. \tag{3.48}$$

Then, we can define the set of strategies of the Attacker as the set of rules associating an admissible map S_{YZ}^n to the to-be-attacked sequence y^n, where, the set of the

admissible maps $\mathcal{A}^n_{L_\infty}(L, P_{y^n})$ is given by

$$\begin{cases} S^n_Y(i; y^n) = P_{y^n} \\ \max_{(i,j):S^n_{YZ}(i,j)\neq 0} |i-j| \leq L, \end{cases} \tag{3.49}$$

where now the distortion constraint is imposed on a per-letter basis and not only on the average (and L is the maximum allowable per-symbol distortion).

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 the Defender does not depend on the set of strategies available to the Attacker. Let $\Gamma^n_{L_\infty}(P_X, \lambda, L)$ denote the set of the types for which the Defender 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 (3.27),

$$\begin{aligned} &\Gamma_{L_\infty}(P_X, \lambda, L) = \\ &\{P \in \mathcal{P} : \exists\ S_{YZ} \in \mathcal{A}_{L_\infty}(L, P) \text{ s.t. } S_Z \in \Lambda^*(P_X, \lambda)\}, \end{aligned} \tag{3.50}$$

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:

$$\max_{(i,j):S^n_{YZ}(i,j)\neq 0} |i-j| \leq L \longleftrightarrow S^n_{YZ}(i,j) = 0, \forall i,j : |i-j| \leq L, \tag{3.51}$$

we deduce that the admissible set in (3.49) is a linear set. Accordingly, Theorem 2 also holds in the L_∞ case and the asymptotic payoff can be computed as in (3.29), with the indistinguishability region given by equation (3.50).

3.6 Extension to sources with memory

The existence of an equilibrium for the DT_{ks} game has been proven by assuming that D bases its analysis on the empirical pmf of the test sequence, i.e., on first order statistics only. Although it might seem that this assumption is justified by the DMS nature of the sources, actually the memorylessness of the sources and the first order-based analysis are independent assumptions and we need to explicitly set both of them. The use of first order statistics to distinguish between two discrete memoryless sources, in fact, is optimum only when no attack is present [90]. In general, the Attacker could introduce memory within z^n, thus making the use of first order statistics sub-optimum. This makes the explicit requirement that the detector relies on first order analysis necessary. As an alternative path, we could have imposed that the attack corresponds to a memoryless channel. In that case, the use of first

order statistics by the Defender could be proven to be an optimal strategy, however we would have simply moved our constraint from the Defender to the Attacker[21].

A closer inspection to the methods used in Sections 3.3 and 3.5, however, reveals that the analysis carried out therein can be extended to sources with memory, as long as the concepts of type and type class can still be used. As a matter of fact, even if the method of types was initially developed to work with memoryless sources [89], it can be extended to more complex models as well. Given a class $\mathcal{C}$ of sources with alphabet $\mathcal{X}$, we say that a partition of $\mathcal{X}^n$ into N_n disjoint sets $T_1, \dots T_{N_n}$, is a partition into type classes if all the sequences in the same T_i are equiprobable for all the sources in $\mathcal{C}$. If the number N_n of type classes grows sub-exponentially with n, then the method of types can be applied to sources in $\mathcal{C}$, and the analysis we carried out in Sections 3.3 (and maybe also 3.5) can be extended to such sources, if we continue to assume that D is restricted to define the acceptance region as a union of type classes. Then, it turns out that the concept of types can be applied to some of the most commonly used source models, including Markov sources with finite order and renewal processes.

For Markov sources of finite order, a model that is commonly used to describe a wide variety of sources with memory, it is known that the number of type classes grows polynomially with n [89], hence making the extension of our analysis possible. For instance, in this case, the limited resources assumption is equivalent to ask that D bases its decision on the empirical transition probabilities induced by x^n plus the pmf of x_1. While the final form of the optimum acceptance region and the minimization problem to be solved by A will be much more complicated, the theoretical analysis will remain essentially the same.

Renewal processes are another class of sources that is amenable to be analyzed by relying on the concept of types. Given a binary source, let us indicate by $\tau_0, \tau_0 + \tau_1, \tau_0 + \tau_1 + \tau_2 \dots$ the positions of the 1's in the sequences produced by the source: τ_i's ($i \geq 1$) are called inter-arrival times, and τ_0 initial waiting time. If the τ_i's are independent and identically distributed random variables, the output of the source is called a renewal process. In the same way, if the τ_i sequence forms a $k-$order Markov chain, the output of the source is called a Markov renewal process of order k. Renewal processes can be used, for instance, to model run length sequences and hence could be of interest in forensic problems dealing with compressed streams adopting run-length coding (e.g. the JPEG coding standard). In [97], it is shown that the number of type classes of renewal processes and Markov renewal processes (of finite order) grows sub-exponentially with n, thus opening the way to the extension of our analysis to this class of sources.

[21] By adopting the Defender's point of view, avoiding to impose any additional constraint on the action of the Attacker may be interpreted as a worst case assumption.

Chapter 4
Detection Games 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. As an example, we can consider a multimedia forensic scenario in which D is asked to verify that a signal has been generated by a given acquisition device. It is very unlikely that a good statistical model of the device is available. On the contrary, it is likely that the analyst will build a suitable model to characterize H_0 by relying on a number of signals produced by the same acquisition device [98]. For these reasons, in this chapter, we remove the assumption that P_X and P_Y are known and study the detection game when training data is available to the players.

The chapter is organized as follow: we first formally define the binary detection game with training data in Section 4.1. Then, in Section 4.2 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 in Section 4.3, where we also compare the performance with those achieved by the game with known sources. Finally, Section 4.4 addresses the case of different training sequences available to the players.

4.1 Definition of the detection game with training data

4.1.1 Problem formalization

By sticking to the notation introduced in Chapter 3.1, 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 DT_{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, D relies on the knowledge of a training sequence of a given length N, namely t_D^N, drawn from X. On

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

his side, A 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 D thinks that the modified sequence was generated by the same source that generated t_D^N. As usual, the Attacker 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, A 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 A and one to 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 take the simplifying assumption that P_Y is known to neither D nor 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 $\mathcal{X}^n \times \mathcal{X}^N$, 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.

4.1.2 The $DT_{tr,a}$ game

With the above ideas in mind, and by paralleling the definition given in Chapter 3, we define a first version of the binary decision game with training sequences as follows:

Definition 5. *The $DT_{tr,a}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a zero-sum, strategic, game played by D and A, defined by the following strategies and payoff.*

- *Defender's strategies.* The set of strategies 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:[3]

$$\mathcal{S}_D = \{\Lambda_{tr}^n \subset \mathcal{X}^n \times \mathcal{X}^N : \max_{P_X \in \mathcal{P}} P_X\{(z^n, t_D^N) \notin \Lambda_{tr}^n\} \leq P_{\text{FP}}^*\}, \tag{4.1}$$

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

[2] To distinguish between the case of known sources and training data, we add the pedex 'tr' and 'ks' in the notation of the quantities Λ, Γ and ε.

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

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

$$\mathcal{S}_A = \{g(\cdot) : d(y^n, g(y^n, t_A^K)) \leq nL\}, \tag{4.2}$$

  where $d(\cdot,\cdot)$ is a proper distortion function and L is the maximum allowed per-letter distortion. Note that the function $g(\cdot)$ depends on t_A^K, since when performing his attack 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 probability, that is:

$$u(\Lambda_{tr}^n, g) = -P_{\text{FN}} = -\sum_{\substack{t_D^N \in \mathcal{X}^N,\ t_A^K \in \mathcal{X}^K \\ y^n : (g(y^n, t_A^K), t_D^N) \in \Lambda_{tr}^n}} P_Y(y^n) P_X(t_D^N) P_X(t_A^K), \tag{4.3}$$

  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. Again, the Defender's perspective is adopted in the definition of the payoff.

Before going on with the analysis, we pause to discuss some of the choices we implicitly made with the above definition.

A first observation regards the payoff function. As a matter of fact, the expression in (4.3) 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 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 optimum strategies of D and 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 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 D would always correspond to $P_Y = P_X$ for which no decision is possible[4].

As a second remark, we stress that, in the DT_{ks} setup, limiting the strategies of the Attacker to deterministic functions of the sequence is not a restrictive choice since, at least asymptotically, the optimum strategy of D does not depend either on the strategy chosen by A (hence on t_A^K) or on P_Y.

[4] Alternatively, we could assume that X and Y belong to two disjoint source classes $\mathcal{C}_X$ and $\mathcal{C}_Y$.

4.1.3 A variant of the game with equal training sequences: the $DT_{tr,b}$ game

An interesting variant of the $DT_{tr,a}$ game is obtained by assuming that the training sequence available to A is equal to that available to D, leading to the following definition.

Definition 6. *The $DT_{tr,b}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a zero-sum, strategic, game defined as the $DT_{tr,a}$ game with the only difference that $K = N$ and $t_A^K = t_D^N$ (simply indicated as t^N in the following). The set of strategies of D and A are the same as in the $DT_{tr,a}$ game, while the payoff is redefined as:*

$$u(\Lambda_{tr}^n, g) = -P_{FN} = - \sum_{\substack{t^N \in \mathcal{X}^N \\ y^n:(g(y^n,t^N),t^N)\in\Lambda}} P_Y(y^n) P_X(t^N). \tag{4.4}$$

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

4.1.4 $DT_{tr,b}$ game with limited resources

Studying the existence of an equilibrium point for the $DT_{tr,b}$ game is a prohibitive task, hence we use the same approach adopted for the known sources case and consider a simplified version of the game in which D can only base his decision on a limited set of statistics computed on the test and training sequences: specifically, we require that 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}. To be consistent with the terminology introduced in the previous chapter, we call this version of the game detection game with *limited-resources*, and refer to it with the notation $DT_{tr,b}^{lr}$ game. Note that P_{z^n} and P_{t^N} are not sufficient statistics for 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 D relies only on P_{z^n} and P_{t^N} as an explicit requirement.

As a consequence of the limited resources assumption, 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. 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.

In our analysis, we are interested in studying the asymptotic behavior of the game when n and N tend to infinity. Rather than considering two limits with n and N tending to infinity independently, we will express N 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 (2.4) (Section 2.3.1). Note that this assumption does not reduce the generality of our analysis, however it destroys the symmetry of the testing problem with respect to the two sequences z^n and t^N. The consequences of this loss of symmetry will be discussed at the end of Section 4.2.1.

We are now ready to define the asymptotic version of the $DT_{tr,b}$ game under the limited resources assumption for the Defender. We will do it by directly rewriting the set of strategies for the Attacker in terms of transportation maps (adopting the transportation theoretic formalism introduced in the previous chapter).[5]

Definition 7. *The $DT_{tr,b}^{lr}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a zero-sum, strategic, game played by D and A, defined by the following strategies and payoff:*

$$\mathcal{S}_D = \{\Lambda_{tr}^n \subset \mathcal{P}_n \times \mathcal{P}_N : \max_{P_X \in \mathcal{P}} P_X\{(z^n, t^{N(n)}) \notin \Lambda_{tr}^n\} \leq 2^{-\lambda n}\}, \tag{4.5}$$

$$\mathcal{S}_A = \{S_{YZ}^n(y^n, t^{N(n)}) : S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})\}, \tag{4.6}$$

$$u(\Lambda_{tr}^n, S_{YZ}^n) = - \sum_{\substack{(y^n, t^{N(n)}) \in \mathcal{X}^n \times \mathcal{X}^N: \\ (S_Z^n(j; y^n, t^{N(n)}), t^{N(n)}) \in \Lambda_{tr}^n}} P_Y(y^n) P_X(t^{N(n)}), \tag{4.7}$$

where in the definition of $\mathcal{S}_A$, we have explicitly indicated that the choice of the transportation map depends on $t^{N(n)}$.[6] By using the transportation theoretic formalism introduced in the previous chapter, the set of strategies of the Attacker consists of all the possible ways of choosing an admissible transportation map to transform y^n into z^n. The constraint on the exponential decay velocity for the false positive probability suggests the asymptotic solution for the game.

A similar definition can be given for version a of the game.

Since we will study only the versions of the game under the limited resources assumption for the Defender, as we did in the previous chapter, in the sequel we will omit the apex lr in the corresponding notation.

[5] Note that, in formulating the game with the transportation approach, we implicitly assume that the distance measure d defining the distortion introduced by the Attacker is invariant to permutation.

[6] Notation $S_{YZ}^n(y^n, t^{N(n)})$ corresponds to S_{YZ}^n where the dependence on the sequences is made explicit, and should not be confused with $S_{YZ}^n(i, j)$, which corresponds to the value taken for a pair of bins (i, j), extensively $S_{YZ}^n(i, j; y^n, t^{N(n)})$.

4.2 Asymptotic equilibrium of the $DT_{tr,b}$ game.

We start the analysis of the asymptotic equilibrium point of the $DT_{tr,b}$ game defined in Definition 7, by determining the optimum acceptance region for D.

To do so we will make an analysis similar to that carried out in [99] to study hypothesis testing with observed statistics. The main difference between our analysis and [99] is the presence of the Attacker, i.e. the game-theoretic nature of our problem. The derivation of the optimum strategy for D passes through the definition of the generalized log-likelihood ratio function $h(z^n, t^N)$ ([100], ch. 24, [99] pg.403).

Given a 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:[7]

$$h(z^n, t^N) = \mathcal{D}(P_{z^n}||P_{r^{n+N}}) + \frac{N}{n}\mathcal{D}(P_{t^N}||P_{r^{n+N}}), \tag{4.8}$$

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

$$r_i = \begin{cases} z_i & i \leq n \\ t_{i-n} & n < i \leq n+N \end{cases} . \tag{4.9}$$

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 $DT_{tr,b}$ game passes through the following lemmas.

Lemma 3. *For any P_X we have:*

$$\begin{aligned} n\mathcal{D}(P_{z^n}||P_{r^{n+N}}) + N\mathcal{D}(P_{t^N}||P_{r^{n+N}}) \leq \\ n\mathcal{D}(P_{z^n}||P_X) + N\mathcal{D}(P_{t^N}||P_X), \end{aligned} \tag{4.10}$$

with equality holding if only if $P_X = P_{r^{n+N}}$.

Proof. We rewrite (4.10) by moving all the non zero terms to the left-hand side:

$$\begin{aligned} n\mathcal{D}(P_{x^n}||P_{r^{n+N}}) + N\mathcal{D}(P_{t^N}||P_{r^{n+N}}) \\ - n\mathcal{D}(P_{x^n}||P_X) - N\mathcal{D}(P_{t^N}||P_X) \leq 0. \end{aligned} \tag{4.11}$$

By using the definition of empirical KL divergence stated in (3.5) and grouping the first term with the third and the second with the fourth, the left hand side of (4.11) is equivalent to

$$n\sum_{a\in\mathcal{X}} P_{x^n}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)} + N\sum_{a\in\mathcal{X}} P_{t^N}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)}. \tag{4.12}$$

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

Being r^{n+N} the concatenation of x^n and t^N, we argue that $nP_{x^n}(a) + NP_{t^N}(a) = (n+N)P_{r^{n+N}}(a)\ \forall a \in \mathcal{X}$, which permits to rewrite the sum in (4.12) as follows:

$$(n+N)\sum_{a\in\mathcal{X}} P_{r^{n+N}}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)} = -(n+N)\mathcal{D}(P_{r^{n+N}}||P_X). \tag{4.13}$$

Hence, the proof of relation (4.11) follows from the positivity of the divergence function, which equals zero if and only if $P_X = P_{r^{n+N}}$.

In hindsight, relation (4.11) 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 (4.13) we write:

$$\sum_{a\in\mathcal{X}} N_{r^{n+N}}(a)\log\frac{P_X(a)}{P_{r^{n+N}}(a)} \leq 0. \tag{4.14}$$

Exploiting the properties of the logarithm, relation (4.14) is equivalent to the following

$$\log\prod_{a\in\mathcal{X}} P_X(a)^{N_{r^{n+N}}(a)} \leq \log\prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)}, \tag{4.15}$$

which implies

$$P_X(r^{n+N}) \leq \prod_{a\in\mathcal{X}} P_{r^{n+N}}(a)^{N_{r^{n+N}}(a)} = P_{r^{n+N}}(r^{n+N}). \tag{4.16}$$

□

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

Lemma 4. *Let $\Lambda_{tr}^{n,*}$ be defined as follows:*

$$\Lambda_{tr}^{n,*} = \left\{(P_{z^n}, P_{t^N}):\quad h(P_{z^n}, P_{t^N}) < \lambda - |\mathcal{X}|\frac{\log(n+1)(N+1)}{n}\right\}, \tag{4.17}$$

with

$$\lim_{n\to\infty}\frac{\log(N(n)+1)}{n} = 0. \tag{4.18}$$

Then:

1. $\max_{P_X} P_X\{(z^n, t^N) \notin \Lambda_{tr}^{n,*}\} \leq 2^{-n(\lambda-\nu_n)}$, *with* $\nu_n \to 0$, *for* $n \to \infty$,
2. $\forall \Lambda_{tr}^n \in \mathcal{S}_D$, *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:

$$\begin{aligned}\max_{P_X} P_{\text{FP}} &= \max_{P_X} \sum_{(z^n,t^N)\in\bar{\Lambda}_{tr}^{n,*}} P_X(z^n,t^N) \\ &= \max_{P_X} \sum_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} P_X(\mathcal{T}(P_{z^n})\times\mathcal{T}(P_{t^N})). \end{aligned} \tag{4.19}$$

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 [90], so we can write:

$$\begin{aligned}\max_{P_X} P_{\text{FP}} &\leq \max_{P_X} \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{n,*}} \\ &\quad [(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}P_X(\mathcal{T}(P_{z^n})\times\mathcal{T}(P_{t^N}))] \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\cdot\max_{P_X} \\ &\quad \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} 2^{-n[\mathcal{D}(P_{z^n}||P_X)+\frac{N}{n}\mathcal{D}(P_{t^N}||P_X)]}, \end{aligned} \tag{4.20}$$

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 [90]). By exploiting Lemma 3, we can write:

$$\begin{aligned}\max_{P_X} P_{\text{FP}} &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\cdot \max_{(P_{z^n},P_{t^N})\in\bar{\Lambda}_{tr}^{*}} \\ &\quad 2^{-n[\mathcal{D}(P_{z^n}||P_{r^{n+N}})+\frac{N}{n}\mathcal{D}(P_{t^N}||P_{r^{n+N}})]} \\ &\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}\ 2^{-n(\lambda-|\mathcal{X}|\frac{\log(n+1)(N+1)}{n})} \\ &= 2^{-n(\lambda-2|\mathcal{X}|\frac{\log(n+1)(N+1)}{n})}, \end{aligned} \tag{4.21}$$

where the last inequality derives from the definition of $\Lambda_{tr}^{n,*}$. Together with (4.18), equation (4.21) proves the first part of the lemma with $\nu_n = 2|\mathcal{X}|\frac{\log(n+1)(N+1)}{n}$.[8]

For any $\Lambda_{tr}^{n} \in \mathcal{S}_D$, 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

[8] We notice that $\nu_n \to 0$ as $n \to \infty$ thanks to the condition in (4.18).

$\mathcal{T}(P_{t^N})$ will be entirely contained in $\bar{\Lambda}^n_{tr}$. Then we have:

$$\begin{aligned} 2^{-\lambda n} &\geq \max_{P_X} P_X(\bar{\Lambda}) \\ &\overset{(a)}{\geq} \max_{P_X} P_X(\mathcal{T}(P_{z^n}) \times \mathcal{T}(P_{t^N})) \\ &\overset{(b)}{\geq} \max_{P_X} \frac{2^{-n[\mathcal{D}(P_{z^n}||P_X)+\frac{N}{n}\mathcal{D}(P_{t^N}||P_X)]}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}} \\ &\overset{(c)}{=} \frac{2^{-n[\mathcal{D}(P_{z^n}||P_{r^{n+N}})+\frac{N}{n}\mathcal{D}(P_{t^N}||P_{r^{n+N}})]}}{(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|}}, \end{aligned} \tag{4.22}$$

where (a) is due to the limited resources assumption, (b) follows from the independence of z^n and t^N and a lower bound on the probability of a pair of type classes [90], and (c) derives from Lemma 3. By taking the logarithm of both sides we find that $(z^n, t^N) \in \bar{\Lambda}^{n,*}_{tr}$, thus completing the proof. □

The first part of Lemma 4 shows that, at least asymptotically, $\Lambda^{n,*}_{tr}$ is an admissible strategy for the Defender; in fact, the constraint in (4.5) is fullfilled asymptotically and then $\Lambda^{n,*}_{tr}$ belongs to $\mathcal{S}_D$ for sufficiently lartge n. Then, the optimality of $\Lambda^{n,*}_{tr}$ follows from the second part of the lemma.

An important observation is that the optimum strategy of 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 D. According to the lemma, the optimum strategy of D does not depend on the strategy chosen by the A (then, neither on the training sequence available to him), that is $\Lambda^{n,*}_{tr}$ is a *strictly dominant strategy* for D. As a consequence, $\Lambda^{n,*}_{tr}$ is the optimum Defender's strategy even for version a of the DT_{tr} game.

As it happened for the DT_{ks} game, due to the existence of a dominant strategy for the Defender, the derivation of the optimum attacking strategy is an easy task. We only need to observe that the goal of 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

$$h(S^n_Z(y^n, t^N), P_{t^N}) < \lambda - |\mathcal{X}|\frac{\log(n+1)(N+1)}{n}, \tag{4.23}$$

The optimum attacking strategy, then, can be expressed as a minimization problem, i.e.,

$$S^{n,*}_{YZ}(y^n, t^N) = \arg \min_{S^n_{YZ} \in \mathcal{A}^n(D_{max}, P_{y^n})} h(S^n_Z, P_{t^N}). \tag{4.24}$$

Note that to implement this strategy A needs to know t^N, i.e., (4.24) determines the optimum strategy only for version b of the game. Since $S^{n,*}_{YZ}(y^n, t^N)$ (res. $S^n_Z(y^n, t^N)$)

depends on the sequences y^n and t^N only through their empirical pmf, we can also use the notation $S^{n,*}_{YZ}(P_{y^n}, P_{t^N})$ (res. $S^n_Z(P_{y^n}, P_{t^N})$).

Having determined the optimum strategies for D and A, we can state the first main result of this chapter.

Theorem 3 (Equilibrium point of the $DT_{tr,b}$ game)**.** *The $DT_{tr,b}$ game is a dominance solvable game and the profile $(\Lambda^{n,*}_{tr}, S^{n,*}_{YZ}(y^n, t^N))$ is the only rationalizable equilibrium.*

4.2.1 Comparison between the test functions in the DT_{ks} and $DT_{tr,b}$ setup

To get a better insight into the meaning of the equilibrium point of the $DT_{tr,b}$ game, it is instructive to compare it with the equilibrium of the corresponding game with known sources, namely the DT_{ks} game.

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 $DT_{tr,b}$ case, D must ensure that the false positive probability stays below the desired threshold for all possible discrete memoryless sources (DMS's). 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 (4.16)) 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 4.1. For large N compared to n, $P_{r^{n+N}}(r^{n+N})$ is closer to P_{t^N} than to P_{z^n}. Another observation regards the optimum 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: both are positive and convex functions achieving the absolute minimum when $P_{z^n} = P_{t^N}$,[9] so one may be tempted to think that from A'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, e.g. when the absolute minimum can be reached, in general the two minimization problems yield different solutions.

To further compare the $DT_{tr,b}$ and the DT_{ks} games, it is useful to rewrite the generalized log-likelihood function in a more convenient way. By applying some

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

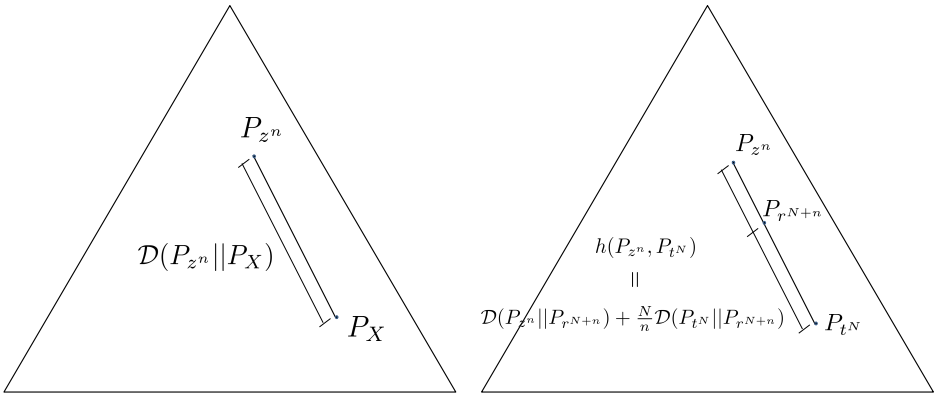


Figure 4.1: Geometric interpretation of the difference between $\mathcal{D}$ (left) and h (right) functions. The position of $P_{r^{N+n}}$ in the segment joining P_{z^n} to P_{t^N} depends on the ratio between the lengths N and n.

algebra, it is easy to prove the following equivalent expression for h:

$$h(P_{z^n}, P_{t^N}) = \mathcal{D}(P_{z^n}||P_{t^N}) - \frac{N+n}{n}\mathcal{D}(P_{r^{n+N}}||P_{t^N}), \tag{4.25}$$

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 A 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 4.3.1, where we will provide a rigorous proof that the $DT_{tr,b}$ game is actually more favorable to the Attacker than the DT_{ks} game.

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

$$P_{r^{n+N}} = c_z P_{z^n} + c_t P_{t^N}. \tag{4.26}$$

In the simplest case, n and N will tend 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 D is dictated by (4.17) and no particular behavior can be noticed. This is not the case when N/n tends to 0 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 (4.8) tends to $\mathcal{D}(P_{z^n}||P_{t^N})$. To understand the behavior of the second term of (4.8) when $n \to \infty$, we can use the Taylor expansion of $\mathcal{D}(P||Q)$ when P approaches Q (see [101], Chapter 4), which applied to the second term of the h function yields:

$$\begin{aligned}\frac{N}{n} \cdot \mathcal{D}(P_{t^N}||P_{r^{n+N}}) &\approx \frac{N}{2n} \cdot \sum_x \frac{(P_{t^N}(x) - P_{r^{n+N}}(x))^2}{P_{r^{n+N}}(x)} \\ &= \frac{N}{2n} \cdot \sum_x \frac{(c_x P_{t^N}(x) - c_x P_{z^n}(x))^2}{P_{r^{n+N}}(x)} \\ &= \frac{\frac{n}{N}}{2(\frac{n}{N}+1)^2} \sum_x \frac{(P_{t^N}(x) - P_{z^n}(x))^2}{P_{r^{n+N}}(x)}. \end{aligned} \tag{4.27}$$

When $N/n \to \infty$, the above expression clearly tends to 0, and hence $h(P_{z^n}, P_{t^N}) \to \mathcal{D}(P_{z^n}||P_{t^N})$. In other words, the optimum 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.

One may wonder the reason behind the asymmetric behavior of the optimum decision strategy when the length of one between the two sequences under analysis grows much faster than the other. This apparent anomaly derives from the choice of analyzing the asymptotic behavior by letting n tend to infinity, a choice that breaks the symmetry between the test and training sequences (if we had defined the false positive and false negative error exponents in terms of N, the situation would have been completely reversed).

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

4.3 Analysis of the payoff at the equilibrium

Having derived the equilibrium point of the $DT_{tr,b}$ 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 $\varepsilon_{tr,b}$.[10] Specifically, we would like to know whether it is possible for 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.[11]

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 A is able to bring y^n within $\Lambda^{n,*}_{tr}$. By adopting the transportation formulation of the attacking 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:

$$\Gamma^n_{tr,b}(\lambda, L) = \{(P,Q) \in \mathcal{P}^n \times \mathcal{P}^N : \exists S^n_{PV} \in \mathcal{A}^n(L,P) \text{ s.t. } (V,Q) \in \Lambda^{n,*}_{tr}(\lambda)\}. \quad (4.28)$$

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 set $\Lambda^{n,*}_{tr}$ and $\Gamma^n_{tr,b}$, that is:

$$\Lambda^{n,*}_{tr}(Q, \lambda) = \{P \in \mathcal{P}^n : (P,Q) \in \Lambda^{n,*}_{tr}(\lambda)\}, \quad (4.29)$$

$$\Gamma^n_{tr,b}(Q, \lambda, L) = \{P \in \mathcal{P}^n : \exists S_{PV} \in \mathcal{A}^n(L,P) \text{ s.t. } V \in \Lambda^{n,*}_{tr}(Q,\lambda)\}. \quad (4.30)$$

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 not necessarily belonging to $\mathcal{P}_n$ or $\mathcal{P}_N$. By remembering that $N/n = c$, we introduce the following definition:

$$h_c(P,Q) = \mathcal{D}(P||U) + c\mathcal{D}(Q||U), \quad (4.31)$$

with

$$U = \frac{1}{1+c}P + \frac{c}{1+c}Q. \quad (4.32)$$

Note that when $P \in \mathcal{P}_n$ and $Q \in \mathcal{P}_N$, the above definition is equivalent to (4.8). By using h_c instead of h, we can generalize definitions (4.30) and (4.29) to a generic pmf Q in $\mathcal{P}$ (not necessarily belonging to $\mathcal{P}_N$).

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

$$\Gamma_{tr,b}(Q, \lambda, L) = \{P \in \mathcal{P} : \exists S_{PV} \in \mathcal{A}(L,P) \text{ s.t. } V \in \Lambda^*_{tr}(Q,\lambda)\}, \quad (4.33)$$

where

$$\Lambda^*_{tr}(Q, \lambda) = \{P : h_c(P,Q) < \lambda\}. \quad (4.34)$$

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

[11] As for the known sources case, in order to derive the expression of the error exponent at the equilibrium of the game, we must require that the admissible set $\mathcal{A}$ is a convex polytope (i.e., the set of constraints defining $\mathcal{A}$ is a linear set) with the considered permutation-invariant distortion d.

Of course, when P and Q are not empirical pmf's, the meaning of $\Lambda_{tr}^{n,*}$ as 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 $DT_{tr,b}$ game, i.e. the set of all the pmf's for which D does not succeed in distinguishing between H_0 and H_1 ensuring a false negative error probability that tends to zero exponentially fast. In other words, if $P_Y \in \Gamma_{tr,b}(P_X,\lambda,L)$, no strictly positive false negative error exponent can be achieved by D. The above conclusions follow from the following theorem:

Theorem 4 (Asymptotic payoff of the $DT_{tr,b}$ game at the equilibrium)**.** *For the $DT_{tr,b}$ game, with $N/n = c$, the false negative error exponent at the equilibrium is given by*

$$\varepsilon_{tr,b}(\lambda) = \min_{Q}[c \cdot \mathcal{D}(Q||P_X) + \min_{P\in\Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)]. \tag{4.35}$$

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 extended Sanov theorem proven in the Appendix (see A).

The false negative error probability at the equilibrium, for a given n, can be written as

$$\begin{aligned} P_{\text{FN}} &= \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q)) P_Y(\Gamma_{tr,b}^n(Q,\lambda,L)) \\ &= \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q)) \sum_{P\in\Gamma_{tr,b}^n(Q,\lambda,L)} P_Y(\mathcal{T}(P)). \end{aligned} \tag{4.36}$$

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 $\mathcal{P}_n$ [90], we can write:

$$\begin{aligned}
P_{\text{FN}} &\leq \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q)) \sum_{P\in\Gamma^n_{tr,b}(Q,\lambda,L)} 2^{-n\mathcal{D}(P||P_Y)} \\
&\leq \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma^n_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)} \\
&\leq \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q))(n+1)^{|\mathcal{X}|} 2^{-n \min\limits_{P\in\Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)} \\
&\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \cdot 2^{-n \min\limits_{Q\in\mathcal{P}_N}[\frac{N}{n}\mathcal{D}(Q||P_X)+\min\limits_{P\in\Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)]} \\
&\leq (n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \cdot 2^{-n \min\limits_{Q}[c\mathcal{D}(Q||P_X)+\min\limits_{P\in\Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)]}, \quad (4.37)
\end{aligned}$$

where the last inequality is obtained by minimizing over all 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:

$$-\frac{\log P_{\text{FN}}}{n} \geq \min_{Q\in\mathcal{C}} \left[c\mathcal{D}(Q||P_X) + \min_{P\in\Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)\right] + \alpha_n, \quad (4.38)$$

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 (4.35). 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[12]. We can write:

$$\begin{aligned}
P_{\text{FN}} &= \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q))P_Y(\Gamma^n_{tr,b}(Q,\lambda,L)) \\
&\geq P_X(\mathcal{T}(Q_N))P_Y(\Gamma^n_{tr,b}(Q_N,\lambda,L)), \\
&\geq \frac{2^{-N\mathcal{D}(Q_N||P_X)}}{(N+1)^{|\mathcal{X}|}} P_Y(\Gamma^n_{tr,b}(Q_N,\lambda,L)), \quad (4.39)
\end{aligned}$$

where in the first inequality we have replaced the sum with the single element of the subsequence Q_N defined previously, and the the second inequality derives from the usual lower bound on the probability of a type class [90]. From (4.39), by taking the log and dividing by n we obtain

$$-\frac{\log P_{\text{FN}}}{n} \leq c\mathcal{D}(Q_N||P_X) - \frac{1}{n}\log P_Y(\Gamma^n_{tr,b}(Q_N,\lambda,L)) + \alpha'_n, \quad (4.40)$$

[12] 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.

where, as in (4.38), $\alpha'_n = |\mathcal{X}|\frac{\log(N+1)}{n}$ tends to 0 when n tends to infinity.

We now apply the extended Sanov limit (see Appendix A) for computing the term $P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L))$ in (4.40). To do so, we must show that $\Gamma^n_{tr,b}(Q_N, \lambda, L) \to \Gamma_{tr,b}(Q^*, \lambda, L)$, where the convergence is intended in the Hausdorff distance.[13] This can be done by reasoning as in the proof of Theorem 2 (when we proved that $\Gamma^n_{ks}(P_X, \lambda, L) \stackrel{H}{\to} \Gamma_{ks}(P_X, \lambda, L)$). 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 same arguments used for the know sources case.

Then, from the generalized Sanov theorem, we get:

$$-\lim_{n\to\infty} \frac{1}{n} \log P_Y(\Gamma^n_{tr,b}(Q_N, \lambda, L)) = \min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)} \mathcal{D}(P||P_Y). \tag{4.41}$$

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

$$-\frac{\log P_{\text{FN}}}{n} \leq c\mathcal{D}(Q^*||P_X) + \beta'_n + \min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)} \mathcal{D}(P||P_Y) + \beta''_n + \alpha'_n, \tag{4.42}$$

where all the sequences α'_n, β'_n and β''_n tend to zero when n tends to infinity.

By coupling equations (4.38) and (4.42) and by letting $n \to \infty$, we eventually obtain:

$$-\lim_{n\to\infty} \frac{\log P_{\text{FN}}}{n} = \min_Q [c \cdot \mathcal{D}(Q||P_X) + \min_{P\in\Gamma_{tr,b}(Q^*,\lambda,L)} \mathcal{D}(P||P_Y)], \tag{4.43}$$

thus proving the theorem. □

According to Theorem 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 (4.35) is obtained by letting $Q = P_X$, it is not possible for 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, it is not possible that the two divergences in (4.35) are simultaneously equal to zero, hence P_{FN} tends to 0 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 despite the presence of the adversary. As anticipated, then, $\Gamma_{tr,b}(P_X, \lambda, L)$ is the indistinguishability region of the $DT_{tr,b}$ game. A pictorial representation of the sets $\Lambda^{n,*}_{tr}$ and $\Gamma_{tr,b}$ is given in Fig. 4.2.

[13] We remind that, for computing the Hausdorff distance, a distance measure between pmf's must be specified, such that $\mathcal{P}$ is bounded (see discussion in the Appendix).

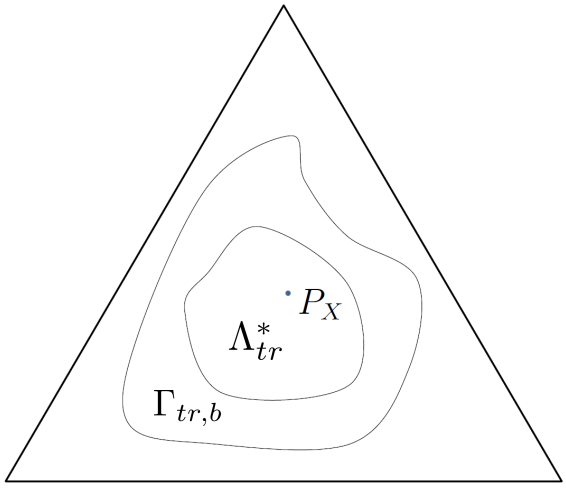


Figure 4.2: Geometric interpretation of the sets Λ^*_{tr} and $\Gamma_{tr,b}$. When $P_Y \in \Gamma_{tr,b}$, a reliable distinction between H_0 and H_1 is not possible and the Attacker *wins* the game.

4.3.1 Comparison between the DT_{ks} and $DT_{tr,b}$ games

In this section we compare the performance achievable by D for the DT_{ks} and $DT_{tr,b}$ games. We start the analysis by comparing the indistinguishability regions of the two games, namely $\Gamma_{ks}(P_X, \lambda, L)$ and $\Gamma_{tr,b}(P_X, \lambda, L)$, reported below for convenience:

$$\Gamma_{ks}(P_X, \lambda, L) = \{P \in \mathcal{P} : \exists S_{PV} \in \mathcal{A}(L, P), \text{ s.t. } V \in \Lambda^*_{ks}(P_X, \lambda)\}, \tag{4.44}$$

with

$$\Lambda^*_{ks}(P_X, \lambda) = \{P \in \mathcal{P} : \mathcal{D}(P||P_X) \leq \lambda\}; \tag{4.45}$$

and

$$\Gamma_{tr,b}(P_X, \lambda, L) = \{P \in \mathcal{P} : \exists S_{PV} \in \mathcal{A}(L, P), \text{ s.t. } V \in \Lambda^*_{tr}(P_X, \lambda)\}, \tag{4.46}$$

with

$$\Lambda^*_{tr}(P_X, \lambda) = \{P \in \mathcal{P} : h_c(P, P_X) \leq \lambda\}. \tag{4.47}$$

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 (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,*

$$h_c(P, P_X) < \mathcal{D}(P||P_X). \tag{4.48}$$

Proof. By rewriting $h_c(P, P_X)$ as in (4.25), we have:

$$h_c(P, P_X) = \mathcal{D}(P||P_X) - (1+c)\mathcal{D}(U||P_X), \tag{4.49}$$

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, it follows immediately the strict inclusion between the acceptance regions, that is $\Lambda^*_{ks}(P_X, \lambda) \subset \Lambda^*_{tr}(P_X, \lambda)$. From Lemma 5, we can prove the following theorem:

Theorem 5 ($DT_{tr,b}$ vs DT_{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* [14] *of $\Gamma_{tr,b}(P_X, \lambda, L)$ there exists a positive value ε such that $B(P, \varepsilon) \subset \overline{\Gamma_{ks}(P_X, \lambda, L)}$* [15]*, where $B(P, \tau)$ is a ball centered in P with radius τ.*
- *For any pmf P belonging to the boundary of $\Gamma_{ks}(P_X, \lambda, L)$ there exists a positive value τ such that $B(P, \tau) \subset \Gamma_{tr,b}(P_X, \lambda, L)$.*
- $\Gamma_{ks}(P_X, \lambda, L) \subset \Gamma_{tr,b}(P_X, \lambda, L)$.

Proof. As an immediate consequence of Lemma 5, we observe that the non-strict inclusion between the indistinguishability regions holds, i.e.,

$$\Gamma_{ks}(P_X, \lambda, L) \subseteq \Gamma_{tr,b}(P_X, \lambda, L). \tag{4.50}$$

Point 1.
Let then 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 $B(P', \varepsilon) \subset \overline{\Gamma_{ks}(P_X, \lambda, L)}$ for some $\varepsilon > 0$, by simply showing that $P' \in \overline{\Gamma_{ks}(P_X, \lambda, L)}$ and then apply the definition of open set.

[14] We remark that, by confining the space to the set of probabilities, i.e., the probability simplex, the following definition of boundary is adopted: given a set $A \in \mathcal{P}$, the *boundary* of A is the set of the points that can be approached both from A and from $\bar{A}$, where $A \cup \bar{A} \equiv \mathcal{P}$. Accordingly, a point in A which lies on the boundary of the simplex is an internal point of A, except for the case in which such point can be approached from $\bar{A}$. Concerning our case, it is worth observing that there could be points inside the set Γ_{ks} (or $\Gamma_{tr,b}$) that lie on the boundary of the simplex, when the distortion constraint the Attacker is subject to is less constraining than the limitation provided by the simplex. According to the above definition, such points (except those that can be approached from the outside/complementary set) do not belong to the boundary.

[15] The relation is equivalent to $B(P, \varepsilon) \cap \Gamma_{ks}(P_X, \lambda, L) = \emptyset$.

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, $h_c(R', P_X) < \lambda$, that is R' is an internal point of Λ^*_{tr}. Let δ be such that $B(R', \delta) \subset \Lambda^*_{tr}$. By exploiting Theorem 25 in Appendix B, it is possible to fix $\varepsilon > 0$ such that for any $P \in B(P', \varepsilon)$ a map $S_{PR} \in \mathcal{A}(L, P)$ can be found such that $R \in B(R', \delta)$ (specifically, we can choose $\varepsilon = \delta/|\mathcal{X}|^2$).

Then, by construction, $B(P', \varepsilon) \subset \Gamma_{tr,b}$, that is, P' is 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 in the boundary of Γ_{ks} is an internal point of $\Gamma_{tr,b}$ [16]. By definition of internal point, there exists $\tau > 0$ such that $B(P, \tau) \subset \Gamma_{tr,b}$, thus concluding the proof.

Point 3.
From the above points it follows that there is at least one point (actually an infinite number of points) that belongs to $\Gamma_{tr,b}$ but not to Γ_{ks}, thus proving that the inclusion relation in (4.50) is indeed strict. □

Theorem 5 has the following corollary.

Corollary 2. *Let ε_{ks} and $\varepsilon_{tr,b}$ denote the error exponents at the equilibrium for the DT_{ks} and $DT_{tr,b}$ games. Then we have:*

$$\varepsilon_{tr,b} \leq \varepsilon_{ks}, \tag{4.51}$$

where the equality holds if and only if $P_Y \in \Gamma_{ks}(P_X, \lambda, L)$, when 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 $DT_{tr,b}$ game we have:

$$\begin{aligned}
\varepsilon_{tr,b} &= \min_{Q}[c \cdot \mathcal{D}(Q||P_X) + \min_{P \in \Gamma_{tr,b}(Q,\lambda,L)} \mathcal{D}(P||P_Y)] \\
&\leq c\mathcal{D}(P_X||P_X) + \min_{P \in \Gamma_{tr,b}(P_X,\lambda,L)} \mathcal{D}(P||P_Y) \\
&\overset{(a)}{=} \min_{P \in \Gamma_{tr,b}(P_X,\lambda,L)} \mathcal{D}(P||P_Y) \\
&< \min_{P \in \Gamma_{ks}(P_X,\lambda,L)} \mathcal{D}(P||P_Y) = \varepsilon_{ks},
\end{aligned} \tag{4.52}$$

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

where the last strict inequality is justified by observing 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 value of P satisfying the minimization on the right-hand side of equality (a) belongs to the *boundary* of $\Gamma_{tr,b}(P_X, \lambda, L)$ which, by Theorem 5, lies outside the closed set $\Gamma_{ks}(P_X, \lambda, L)$. □

Theorem 5 and Corollary 2 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 $DT_{tr,b}$ game.

4.4 Detection game with independent training sequences ($DT_{tr,a}$)

We now pass to the analysis of version a of the DT_{tr} game. Again, we consider the limited resources version of the game. We remind that in this case D and A rely on independent training sequences, namely t_D^N and t_A^K. As for 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. This is the case, for instance, in source identification for multimedia forensics, where we can assume that the analyst has a better knowledge of the statistics of authorized sources. 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 4.2, the strategy $\Lambda_{tr}^{n,*}$ identified by Lemma 4 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 A does not know. In this case, finding the optimum attack strategy is an hard 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 4.3 (equation (4.29)), 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 $DT_{tr,a}$ game is equal to that of the $DT_{tr,b}$ game, since outside the indistinguishability region, the false negative error exponent for case a may be different (actually larger) than that of case b.

4.4.1 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 $DT_{tr,a}$ 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:

$$\tilde{S}_{YZ}^n(P_{y^n}, P_{t_A^N}) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} h(S_Z^n, P_{t_A^N}). \tag{4.53}$$

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

By following the same flow of ideas used in Section 4.3, we consider the set of pmf's for which the Attacker is able to move P_{y^n} within the acceptance region, that is

$$\tilde{\Gamma}_{tr,a}^n(\lambda, L) = \{(P_{y^n}, P_{t_D^N}, P_{t_A^N}) : (\tilde{S}_Z^n(P_{y^n}, P_{t_A^N}), P_{t_D^N}) \in \Lambda_{tr}^{n,*}(\lambda)\}. \tag{4.54}$$

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

$$\tilde{\Gamma}_{tr,a}^n(P_{t_D^N}, P_{t_A^N}, \lambda, L) = \{P_{y^n} \in \mathcal{P}_n : \tilde{S}_Z^n(P_{y^n}, P_{t_A^N}) \in \Lambda_{tr}^{n,*}(P_{t_D^N}, \lambda)\}. \tag{4.55}$$

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:

$$\tilde{\Gamma}_{tr,a}^n(Q, R, \lambda, L) = \{P \in \mathcal{P}_n : \tilde{S}_Z^n(P, R) \in \Lambda_{tr}^{n,*}(Q, \lambda)\}. \tag{4.56}$$

It is easy to see that:

$$\begin{aligned} &\tilde{\Gamma}_{tr,a}^n(Q, R, \lambda, L) \subseteq \tilde{\Gamma}_{tr,a}^n(Q, Q, \lambda, L) \\ &\tilde{\Gamma}_{tr,a}^n(Q, Q, \lambda, L) = \Gamma_{tr,b}^n(Q, \lambda, L), \end{aligned} \tag{4.57}$$

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

$$\tilde{\Gamma}_{tr,a}(Q, R, \lambda, L) = \{P \in \mathcal{P} : \tilde{S}_Z(P, R) \in \Lambda^*_{tr}(Q, \lambda)\}, \tag{4.58}$$

where $\Lambda^*_{tr}(Q, \lambda)$ is the same set defined in (4.34). Straightforwardly, the relations in (4.57) also holds for $\tilde{\Gamma}_{tr,a}$.

We are now ready to prove the following result.

Theorem 6 (Asymptotic payoff of the $DT_{tr,a}$ game)**.** *The error exponent of the payoff associated to the profile* $(\Lambda^{*,n}_{tr}(P_{t_D^N}), \tilde{S}^n_Z(P_{y^n}, P_{t_A^N}))$ *is lower (res. upper) bounded as follows*[17]

$$\tilde{\varepsilon}_{tr,a} \geq \min_{Q,R} \big\{ c[\mathcal{D}(Q||P_X) + \mathcal{D}(R||P_X)] + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,R,\lambda,L)} \mathcal{D}(P||P_Y)) \big\}, \tag{4.59}$$

$$\tilde{\varepsilon}_{tr,a} \leq \min_{Q} \big[2c \cdot \mathcal{D}(Q||P_X) + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P||P_Y) \big]. \tag{4.60}$$

Proof. The proof is similar to the proof of Theorem 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

[17] We adopt the definition $\tilde{\varepsilon}_{tr,a} = -\limsup_{n\to\infty} \frac{1}{n} \log(P_{\text{FN}})$, since the lim may not exists (see (2.4)).

recalling the definition of the false negative error probability, for any n we can write:

$$\begin{aligned}
P_{\text{FN}} &= \sum_{t_D^N}\sum_{t_A^N} P_X(t_D^N)P_X(t_A^N)P_Y(\tilde{\Gamma}^n_{tr,a}(P_{t_D^N}, P_{t_A^N}, \lambda, L)) \\
&= \sum_{t_D^N}\sum_{t_A^N} P_X(t_D^N)P_X(t_A^N) \sum_{P\in\tilde{\Gamma}^n_{tr,a}(P_{t_D^N},P_{t_A^N},\lambda,L)} P_Y(\mathcal{T}(P)) \\
&= \sum_{Q\in\mathcal{P}_N}\sum_{R\in\mathcal{P}_N} P_X(\mathcal{T}(Q))P_X(\mathcal{T}(R)) \sum_{P\in\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)} P_Y(\mathcal{T}(P)) \\
&\leq \sum_{Q\in\mathcal{P}_N}\sum_{R\in\mathcal{P}_N} P_X(\mathcal{T}(Q))P_X(\mathcal{T}(R))\cdot(n+1)^{|\mathcal{X}|}2^{-n\min\limits_{P\in\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)}\mathcal{D}(P||P_Y)} \\
&\leq \sum_{Q\in\mathcal{P}_N} P_X(\mathcal{T}(Q))(n+1)^{|\mathcal{X}|}(N+1)^{|\mathcal{X}|} \\
&\qquad \cdot 2^{-n\min\limits_{R\in\mathcal{P}_N}[c\mathcal{D}(R||P_X)+\min\limits_{P\in\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)}\mathcal{D}(P||P_Y)]} \\
&\leq (n+1)^{|\mathcal{X}|}(N+1)^{2|\mathcal{X}|} \\
&\qquad \cdot 2^{-n\min\limits_{Q,R}[c\mathcal{D}(Q||P_X)+c\mathcal{D}(R||P_X)+\min\limits_{P\in\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)}\mathcal{D}(P||P_Y))]},
\end{aligned} \tag{4.61}$$

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, the lower bound in (4.59) is easily proved.

We now turn the attention to the upper bound. To do so, let Q^* be the pmf achieving the minimum in (4.60). 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 to write the following chain of inequalities:

$$\begin{aligned}
P_{\text{FN}} &= \sum_{Q\in\mathcal{P}_N}\sum_{R\in\mathcal{P}_N} P_X(\mathcal{T}(Q))P_X(\mathcal{T}(R))P_Y(\tilde{\Gamma}^n_{tr,a}(Q,R,\lambda,L)) \\
&\geq P_X(\mathcal{T}(Q_N))P_X(\mathcal{T}(R_N))P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L)) \\
&\geq \frac{2^{-N(\mathcal{D}(Q_N||P_X)+\mathcal{D}(R_N||P_X))}}{(N+1)^{2|\mathcal{X}|}}\cdot P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N,R_N,\lambda,L)),
\end{aligned} \tag{4.62}$$

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 ([90], chapter 12). By taking the logarithm of each side in (4.62) and dividing by n, we get:

$$-1/n \log(P_{\text{FN}}) \leq c\mathcal{D}(Q_N||P_X) + c\mathcal{D}(R_N||P_X) \\ - 1/n \log(P_Y(\tilde{\Gamma}^n_{tr,a}(Q_N, R_N, \lambda, L))) - \beta_n, \tag{4.63}$$

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

In order to apply the generalized Sanov theorem for evaluating the the probability term in (4.63), we need to prove that

$$\tilde{\Gamma}^n_{tr,a}(Q_N, R_N, \lambda, L) \xrightarrow{H} \Gamma_{tr,b}(Q^*, \lambda, L). \tag{4.64}$$

We observe that the proof of such convergence is more involved with respect to similar proofs in Theorem 2 and 4, due to the inveolved 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 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 $\tilde{\Gamma}^n_{tr,a}$ cannot be written in a form similar to (4.30) (and (3.26)), thus preventing us from directly using the same arguments used therein.

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.

Let $\lambda'_n = \max\{\lambda' : \Lambda^{n,*}_{tr}(R_N, \lambda') \subseteq \Lambda^{n,*}_{tr}(Q_N, \lambda)\}$.[18] We can define the following set:

$$\dot{\Gamma}^n_{tr,a}(Q_N, R_N, \lambda, L) = \{P \in \mathcal{P}^n : \exists S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Lambda^{n,*}_{tr}(R_N, \lambda'_n)\}. \tag{4.65}$$

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:

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

Since $\Gamma^n_{tr,b}(Q_N, \lambda, L) \xrightarrow{H} \Gamma_{tr,b}(Q^*, \lambda, L)$ (see the proof of Theorem 4), by applying the squeeze theorem, (4.64) is proven if we show that[19]

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

By reasoning as in the proof of Theorem 2 and 4, the above relation follows by proving that $\delta_H(\Lambda^{n,*}_{tr}(R_N, \lambda'_n), \Lambda^*(Q^*, \lambda)) \to 0$ as $n \to \infty$, which derives easily from

[18] 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 4.3).

[19] The squeeze theorem (known also as sandwich theorem) also holds in the case of Hausdorff convergence [102].

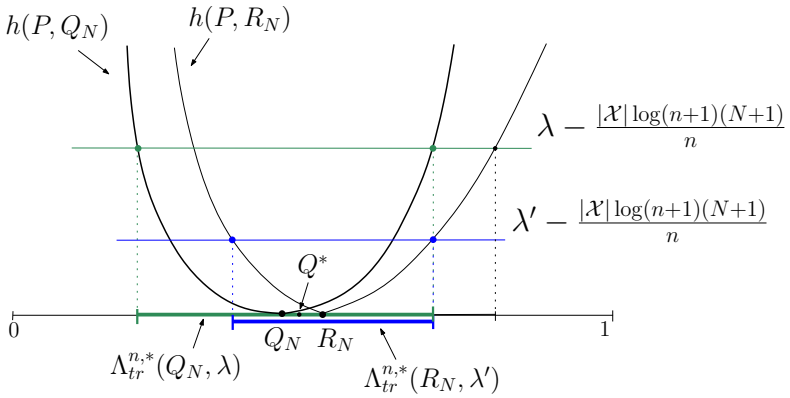


Figure 4.3: Geometric construction of set $\Lambda_{tr}^{n,*}(R_N, \lambda')$. For ease of graphical representation, the sketch refers to the case $|\mathcal{X}| = 2$ (one-dimensional case).

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.

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

$$-\lim_{n\to\infty} 1/n \log(P_Y(\tilde{\Gamma}_{tr,a}^n(Q_N, R_N, \lambda, L)) = \min_{P\in\tilde{\Gamma}_{tr,a}(Q^*,Q^*,\lambda,L)} \mathcal{D}(P||P_Y). \tag{4.66}$$

Therefore, by going on from (4.63), letting $n \to \infty$ and exploiting the continuity of $\mathcal{D}$ with respect to its arguments, we have

$$\tilde{\varepsilon}_{tr,a} \leq c\mathcal{D}(Q^*||P_X) + c\mathcal{D}(Q^*||P_X) + \min_{P\in\tilde{\Gamma}_{tr,a}(Q^*,Q^*)} \mathcal{D}(P||P_Y). \tag{4.67}$$

By recalling that

$$Q^* = \arg\min_Q \left[c \cdot \mathcal{D}(Q||P_X) + \min_{P\in\tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P||P_Y)\right], \tag{4.68}$$

we get the upper bound in (4.60). □

Theorem 6 has an important corollary.

Corollary 3 (Indistinguishability region for $DT_{tr,a}$)**.** *The false negative error exponent associated to the profile $(\Lambda_{tr}^{n,*}(P_{t_D^N}), \tilde{S}_{YZ}(\cdot, P_{t_A^N}))$ 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 $DT_{tr,a}$ game is equal to that of the $DT_{tr,b}$ game.*

Proof. From the upper bound in Theorem 6, 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 that $P_Y \in \tilde{\Gamma}_{tr,a}(P_X, P_X, \lambda, L)$. Being $\tilde{\Gamma}_{tr,a}(P_X, P_X, \lambda, L) = \Gamma_{tr,b}(P_X, L, \lambda)$, the corollary is proven. □

Corollary 3 provides an interesting insight into the achievable performance of the $DT_{tr,a}$ 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 side, a strategy that allows the Attacker to reach the same payoff as for version b may not exists.

4.4.2 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 A knows at least the form the h_c function used by D. We focus on the following attacking strategy: use the training sequence t_A^K to estimate $P_{t_D^N}$ and use the estimate to attack the sequence y^n. Specifically, the Attacker may use the following estimate of $P_{t_D^N}$:

$$\begin{aligned}\tilde{P}_{t_D^N}(i) &= \frac{1}{N}\left\lfloor P_{t_A^K}(i) \cdot N \right\rfloor \quad \forall i = 1 \dots |\mathcal{X}| - 1, \\ \tilde{P}_{t_D^N}(|\mathcal{X}|) &= 1 - \sum_{i=1}^{|\mathcal{X}|-1} \tilde{P}_{t_D^N}(i),\end{aligned} \tag{4.69}$$

to implement the attacking function:

$$\tilde{S}_{YZ}^n(P_{y^n}, P_{t_A^K}) = \arg \min_{S_{YZ}^n \in \mathcal{A}^n(L, P_{y^n})} h_c(P_{z^n}, \tilde{P}_{t_D^N}). \tag{4.70}$$

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 6 can be rewritten as:

$$\tilde{\varepsilon}_{tr,a} \leq \min_Q \left[(c+d) \cdot \mathcal{D}(Q||P_X) + \min_{P \in \tilde{\Gamma}_{tr,a}(Q,Q,\lambda,L)} \mathcal{D}(P||P_Y)\right], \tag{4.71}$$

whose proof is practically identical to the proof of Theorem 6 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 A knows exactly the acceptance region adopted by D and hence his attacks will surely be more effective), equation (4.71) permits to conclude that the indistinguishability region is equal to that obtained for the case $c = d$.

4.5 Detection game with dependent training data

We end this section by considering yet another version of the DT_{tr} game, of practical interest in some situations. In certain cases, in fact, we may assume that D has the possibility to observe the output of a system under H_0, and hence the output of the source X, for a longer time than A (see [103] for a multimedia forensics scenario in which such an assumption holds quite naturally). In this framework, the Attacker knows a subpart of the training set available to D.[20] We can model such a situation by assuming that the sequence t_A^K is a subsequence of t_D^N, leading to the following definition.

Definition 8. *The $DT_{tr,sub}(\mathcal{S}_D, \mathcal{S}_A, u)$ game is a zero-sum, strategic, game defined as the $DT_{tr,a}^{lr}$ game with the only difference that $t_A^K = (t_{D,l+1}, t_{D,l+2} \dots t_{D,l+K})$ with l and K known to D.*

In some sense, we can say that this new version of the game is halfway between versions a and b. Like in version a, the Attacker does not have a perfect knowledge of the training sequence used by the Defender and hence he must resort to an estimate of the true acceptance region. On the other hand, the situation is more favorable to the Attacker with respect to version a with $d < c$, since now A knows at least part of the training samples used by D. Given that versions a and b of the game have the same indistinguishability region, we can conjecture that the indistinguishability region of this latest version of the game will also be the same.[21]

[20] We are considering a particuar dependence between the training sequences, which is of interest in some practical applications.

[21] We remind that $d > c$ is not a case of interest because of the arguments discussed at the beginning of Section 4.4.

Chapter 5

Limiting Performance of the Adversarial Detection: Source Distinguishability

The game theoretical analysis of the previous chapters permitted us to study the distinguishability of two sources in a Neyman-Pearson setup under adversarial conditions, that is, when an attacker modifies the output of one of the two sources subject to a distortion constraint. In this chapter, we overcome the lack of symmetry inherent in the NP approach by symmetrizing the role of the two kinds of error probabilities. This permits to study the *ultimate achievable performance* of the detection in adversarial setting. By exploiting the parallelism with Optimal Transport Theory, 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 an adversarial setting.

We derive the $\mathcal{SM}$ for a wide class of pmf's in both the discrete and the continuous case and, by relying on some results in the field of optimal transport theory, we present a numerical algorithm for its efficient computation. We also derive general bounds on $\mathcal{SM}$ assuming that the distortion is measured in terms of the mean square error between the original and the attacked sequence.

The chapter is organized as follows: in Section 5.1 we study the limiting performance of the game and introduce the Security Margin concept for adversarial detection with known sources (DT_{ks} game) and training data (DT_{tr} game), when an additive distortion measure is adopted by the Attacker. In Section 5.2, we derive the Security Margin for several classes of sources, and provide an efficient algorithm to compute it when a close form solution can not be found. Section 5.3 extends the Security Margin concept to a situation in which the distortion is defined in terms of L_∞ distance.

5.1 The Security Margin

A drawback with the analysis carried out in Chapter 3 and 4 is the asymmetric role of the false positive and false negative error exponents, namely μ and ε [1]. In that cases, in fact, the Defender aims at ensuring a given value for μ, namely λ (i.e., the Defender imposes that $\mu \geq \lambda$), 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 Stein's lemma [90] for the non adversarial version of the test, we analyze the behavior of the indistinguishability regions of the tests, namely $\Gamma_{ks}(P_X, \lambda, L)$ and $\Gamma_{tr}(P_X, \lambda, L)$, when the false positive decay rate λ tends to 0, to see whether, given a maximum allowable distortion L, it is possible for 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.

5.1.1 Security Margin for the DT_{ks} game

In this section, we adopt an optimal transport interpretation, to introduce a measure of source distinguishability in the set-up defined by the DT_{ks} game when an additive distortion measure is used by the Attacker. This is the most common and interesting category of permutation invariant measures. Another interesting case of permutation invariant distance which does not fall into this category is the case of maximum distance, which is treated separately in Section 5.3.

Characterization of the indistinguishability region using Optimal Transportation

To start with, we find it convenient to rephrase the results obtained in Chapter 3 as an *optimal transport problem* [93].

Let P and Q be two pmf's defined over the same finite alphabet, and let $c(i, j)$ be the cost of transporting the i-th symbol into the j-th one. In one of its instances, optimal transport theory looks for the transportation map that transforms P into Q by minimizing the average cost of the transport. By using the notation introduced in the previous section, this corresponds to solving the following minimization problem:

$$\min_{S_{YZ}: S_Y = P, S_Z = Q} \sum_{i,j} S_{YZ}(i,j) c(i,j). \tag{5.1}$$

[1] To distinguish between the DT_{ks} and the DT_{tr} games, we will use the subscript ks and tr in the notation of the main quantities.

A nice interpretation of the problem defined by equation (5.1) is obtained by interpreting the pmf's P and Q as two different ways of piling up a certain amount of soil, and $c(i,j)$ as the cost necessary to move a unitary amount of earth from position i to position j. In this case, the minimum cost achieved in (5.1) can be seen as the minimum effort required to turn one pile into the other. Due to such a viewpoint, in computer vision applications, the minimum in equation (5.1) is usually known as Earth Mover Distance (*EMD*) between P and Q [104]. However, while the definition of the *EMD* given in [104] refers in general to signatures (non-normalized distributions with unequal masses), here the earth piles P and Q are probability mass functions. In this case, when $c(i,j) = l(i,j)^p$ for some distance measure l (with $p \geq 1$), the *EMD* has a more general statistical meaning. Given 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 the random variables X and Y taken over all joint probability distributions P_{XY} with marginal distributions respectively equal to P_X and P_Y:

$$EMD_{l^p}(P_X, P_Y) = \min_{P_{XY}: \substack{\sum_y P_{XY}=P_X \\ \sum_x P_{XY}=P_Y}} E_{XY}[l(X,Y)^p]. \tag{5.2}$$

In transport theory terminology, expression (5.2) is the p-th power of the Wasserstein distance [105], [93] (or the Monge-Kantorovich metric of order p [106], [107]). In particular, when $c(i,j) = |i-j|^2$ (i.e. $l(i,j) = |i-j|$ and $p=2$) the earth mover distance $EMD_{L_2^2}(P_X, P_Y)$ is equivalent to the squared Mallows distance between P_X and P_Y [108]. In the following, we will continue to refer to (5.1) as *EMD(P,Q)*. We also observe that even if we introduced the *EMD* by considering finite-alphabet sources, there is no need to restrict the definition in (5.2) to discrete random variables. In fact, in the sequel, we will extend our analysis and use the *EMD* to measure the distinguishability of continuous sources.

Optimal transport theory permits to rewrite the indistinguishability region in a more compact and easier-to-interpret way. In fact, it is immediate to see that equation (3.27) can be rewritten as:

$$\Gamma_{ks}(P_X, \lambda, L) = \{P \in \mathcal{P} : \ \exists\, Q \in \Lambda^*_{ks}(P_X, \lambda) \text{ s.t } EMD(P,Q) \leq L\}, \tag{5.3}$$

where, in the definition of the *EMD*, $c(\cdot,\cdot)$ corresponds to the distortion metric $d(\cdot,\cdot)$ used to constraint the strategies available to the Attacker.

Such insightful rewriting of the indistinguishability region is useful in the subsequent analysis, which leads to the Security Margin definition.

Security Margin definition

We now consider the sequence of DT_{ks} games as λ decreases and study the behavior of $\Gamma_{ks}(P_X, \lambda, L)$ when $\lambda \to 0$. Doing so will allow us to investigate whether two

sources X and Y are *ultimately* distinguishable in the setting defined by the DT_{ks} game. The rationale behind such analysis derives directly from the definition of the acceptance region. In fact, from the definition of $\mathcal{S}_D$, 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 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 0 exponentially fast.

More formally, we start by proving the following property.

Property 2. *For any two values λ_1 and λ_2 such that $\lambda_2 < \lambda_1$, $\Gamma_{ks}(P_X, \lambda_2, L) \subseteq \Gamma_{ks}(P_X, \lambda_1, L)$.*

Proof. The property follows immediately from (5.3) by observing that $\Gamma_{ks}(P_X, \lambda, L)$ depends on λ only through the acceptance region $\Lambda_{ks}(P_X, \lambda)$, for which we obviously have $\Lambda_{ks}(P_X, \lambda_2)^* \subseteq \Lambda_{ks}(P_X, \lambda_1)^*$ whenever $\lambda_2 < \lambda_1$. □

Thanks to Property 2, we can compute the limit of the false negative error exponent when λ tends to zero, as summarized in the following theorem (somewhat resembling Stein's Lemma [90]).

Theorem 7. *Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum average per-letter distortion L, let us adopt the following definition:*[2]

$$\Gamma(P_X, L) = \{P \in \mathcal{P} : \text{EMD}(P, P_X) \leq L\}; \tag{5.4}$$

then the maximum achievable false negative error exponent for the DT_{ks} game is

$$\lim_{\lambda \to 0} \lim_{n \to \infty} -\frac{1}{n} \log P_{FN} = \min_{P \in \Gamma(P_X, L)} \mathcal{D}(P||P_Y). \tag{5.5}$$

Proof. The innermost limit in the left-hand side of (5.5) defines the error exponent for a fixed λ, say it $\varepsilon_{ks}(\lambda)$. From Theorem 2, we know that

$$\lim_{n \to \infty} -\frac{1}{n} \log P_{FN} = \varepsilon(\lambda) = \min_{P \in \Gamma_{ks}(P_X, \lambda, L)} \mathcal{D}(P||P_Y). \tag{5.6}$$

Then, according to Property 2, the sequence $\varepsilon_{ks}(\lambda)$ is monotonically non decreasing as λ decreases. In addition, since $\Gamma(P_X, L) \subseteq \Gamma_{ks}(P_X, \lambda, L)\ \forall \lambda$, for any $\lambda > 0$, we have:

$$\varepsilon_{ks}(\lambda) \leq \min_{P \in \Gamma(P_X, L)} \mathcal{D}(P||P_Y). \tag{5.7}$$

[2] We avoid the subscript ks in the definition of the ultimate indistinguishability region $\Gamma(P_X, L)$, because, as we will see in the sequel, this is the same for both the DT_{ks} and DT_{tr} game.

Being $\varepsilon_{ks}(\lambda)$ 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 (5.5) and P_λ^* the point achieving the minimum on the set $\Gamma_{ks}(P_X, \lambda, L)$, i.e., the point achieving the minimum in equation (3.29) (see Figure 3.2 for a pictorial representation of P_λ^*). Due to Lemma 10 (Appendix C.1), for any arbitrarily small τ, we can choose a small enough λ such that, for any P in $\Gamma_{ks}(P_X, \lambda, L)$, 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

$$\mathcal{D}(P'||P_Y) \leq \min_{P\in\Gamma_{ks}(P_X,\lambda,L)} \mathcal{D}(P||P_Y) + \delta(\tau), \tag{5.8}$$

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 (5.8) holds for $P' = P_0^*$ and then we can write

$$\varepsilon_{ks}(\lambda) \geq \min_{P\in\Gamma(P_X,L)} \mathcal{D}(P||P_Y) - \delta(\tau). \tag{5.9}$$

where $\delta(\tau)$ can be made arbitrarily small by decreasing λ.
Equation (5.9), together with equation (5.7), show 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 (5.5) is the limit of the sequence $\varepsilon_{ks}(\lambda)$ as $\lambda \to 0$. □

Figure 5.1 gives a geometric interpretation of Theorem 7. The figure is obtained from Figure 3.2 in Chapter 3 by observing that when $\lambda \to 0$ the optimum acceptance region collapses into the single pmf P_X, i.e., $\Lambda^* = \{P_X\}$.

By the light of Theorem 7, $\Gamma(P_X, L)$ is the smallest indistinguishability region for the DT_{ks} game. Moreover, from equation (5.4), we see that the distinguishability of two pmf's (in the DT_{ks} setting) ultimately depends on their *EMD*. In fact, if $EMD(P_Y, P_X) > L$, the Defender 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.

By adopting a different perspective, given two sources X and Y, one may ask which is the maximum attacking distortion for which D can distinguish X and Y. The answer to this question follows immediately from Theorem 7 and leads naturally to the following definition.

Definition 9 (Security Margin in the DT_{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 DT_{ks} setup is called Security Margin and is given by*

$$\mathcal{SM}(P_Y, P_X) = \mathrm{EMD}(P_Y, P_X). \tag{5.10}$$

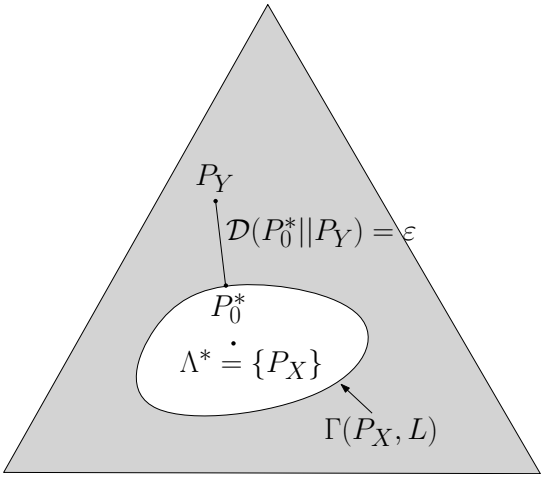


Figure 5.1: Geometric interpretation of $\Gamma(P_X, L)$ and P_0^* by the light of Theorem 7.

Since the *EMD* is a symmetric function of P_X and P_Y [104], the Security Margin does not depend on the role of X and Y in the test, i.e. $\mathcal{SM}(P_X, P_Y) = \mathcal{SM}(P_Y, P_X)$.

Discussion

The Security Margin is a powerful measure summarizing in a single quantity *how securely* two sources can be distinguished in (a given) 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 optimal choices. To better 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:

$$\sum_{i,j} S_{YX}(i, j)d(i, j) = S_{YX}(0, 1) + S_{YX}(1, 0). \tag{5.11}$$

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 $|p - q|$. 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 will still be distinguishable from the sequences emitted by X. Consider, for example, a simple case in which $q = 1/2$ and $p > 1/2$. Regardless of the crossover probability, the output of the channel will still be a binary source with equiprobable symbols, which is distinguishable from X given that $p > 1/2$. In other words, in the set up defined by the DT_{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 even if the output of the source Y passes through a noisy channel introducing the same average distortion introduced by the Attacker.

5.1.2 Security Margin for the DT_{tr} game

We now study the behavior of the DT_{tr} game studied in Chapter 4 when $\lambda \to 0$ so to investigate the *best achievable* performance for the Defender in this case.

By proceeding as in the previous section, we first rewrite the set $\Gamma_{tr}(Q, \lambda, L)$ in (4.33)–(4.34) by exploiting the optimal transport interpretation:

$$\Gamma_{tr}(Q, \lambda, L) = \{P \in \mathcal{P} : \exists R \in \Lambda^*_{tr}(Q, \lambda) \text{ s.t. } EMD(P, R) \leq L\}, \tag{5.12}$$

where

$$\Lambda^*_{tr}(Q, \lambda) = \{P \in \mathcal{P} : h_c(P, Q) \leq \lambda\}. \tag{5.13}$$

We remind that (from the analysis in Chapter 4) when $Q = P_X$ the above set corresponds to the indistinguishability region for the DT_{tr} game, namely $\Gamma_{tr}(P_X, \lambda, L)$[3].

To start with, we observe that the divergence and the h_c function share 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, Property 2 can be extended to the set $\Gamma_{tr}(Q, \lambda, L)$, yielding:

Property 3. *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 10 (Appendix C.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 7 for the detection game with training data. In doing so, we focus on the

[3]We remind that, with regard to the indistinguishability region, we have proven that the games studied in Chapter 4 are all equivalent, so we do not need to differentiate the notation.

game with equal training sequences; the extension of the analysis to the other cases is straightforward and is discussed afterwards.

Theorem 8. *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 $DT_{tr,b}$ game is*

$$\lim_{\lambda \to 0} \varepsilon_{tr}(\lambda) = \min_{Q} \left[c \cdot \mathcal{D}(Q||P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P||P_Y) \right], \tag{5.14}$$

where $\Gamma(Q, L)$ is defined as in (5.4) by replacing P_X with Q[4]*.*

Proof. The proof goes along the same line of the proof of Theorem 7. We know from Theorem 4 (in Chapter 4) that the expression of the false negative error exponent of the $DT_{tr,b}$ game for the payoff at the equilibrium is given by

$$\varepsilon_{tr}(\lambda) = \min_{Q} [c \cdot \mathcal{D}(Q||P_X) + \min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P||P_Y)], \tag{5.15}$$

where c is the ratio between the lengths of training and test.

From Property 3, we see immediately that $\varepsilon(\lambda)$ is non-increasing when λ decreases, since the innermost minimization in equation (5.15) is taken over a smaller set when λ decreases. Then, by the same token, we have:

$$\varepsilon_{tr}(\lambda) \leq \min_{Q} \left(c\mathcal{D}(Q||P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P||P_Y) \right). \tag{5.16}$$

This implies that $\lim_{\lambda \to 0} \varepsilon_{tr}(\lambda)$ exists and is finite. Given that Lemma 10 still holds for the set $\Gamma_{tr}(Q, \lambda, L)$ $\forall Q$, we can reason as in the proof of Theorem 7 to conclude that:

$$\min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P||P_Y) \geq \min_{P \in \Gamma(Q,L)} \mathcal{D}(P||P_Y) - \delta(\tau), \tag{5.17}$$

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 (5.17) and considering that the relation holds for any $Q \in \mathcal{P}$, we can write:

$$\begin{aligned} \varepsilon_{tr}(\lambda) &= \min_{Q} \left[c\mathcal{D}(Q||P_X) + \min_{P \in \Gamma_{tr}(Q,\lambda,L)} \mathcal{D}(P||P_Y) \right] \\ &\geq \min_{Q} \left[c\mathcal{D}(Q||P_X) + \min_{P \in \Gamma(Q,L)} \mathcal{D}(P||P_Y) \right] - \delta(\tau), \end{aligned} \tag{5.18}$$

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

[4] We will see afterwards that, when λ tends to 0, we do not need to differentiate anymore between the DT_{ks} and DT_{tr} games in the definition of Γ.

A consequence of Theorem 8 is that $\lim_{\lambda \to 0} \varepsilon(\lambda) = 0$ if and only if $P_Y \in \Gamma(P_X, L)$, which then can be seen as the smallest indistinguishability region for the DT_{tr} game. We conclude that the smallest indistinguishability regions for the two cases are the same thus implying that the Security Margin for the DT_{tr} setting, say $\mathcal{SM}_{tr}$, is the same of the DT_{ks} game, that is

$$\mathcal{SM}_{tr}(P_X, P_Y) = EMD(P_X, P_Y). \tag{5.19}$$

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 DT_{tr} setting is lower than that in the DT_{ks} setting. However, the difference between the two settings regards the decay rate of the error probabilities, not the ultimate distinguishability of the sources.

We conclude this section with a brief discussion of the DT_{tr} game with different training sequences ($t_D^N \neq t_A^K$), namely the $DT_{tr,a}$ game. We known from the analysis developed in Chapter 4 that, as long as the length of both sequences grows linearly with n, the indistinguishability region is equal to that of the game with equal training sequences. By relying on this result, it is straightforward to prove that the Security Margin remains the same even for such version of the game.

5.2 Security Margin computation

We now focus on the computation of the Security Margin for two generic sources. By following the analysis given so far, we first consider the case of discrete sources, then, at the end of the section, we extend the analysis to the case of continuous sources.

Given two discrete sources $X \sim P_X$ and $Y \sim P_Y$, the computation of the Security Margin requires the evaluation of $EMD(P_X, P_Y)$. A closed form solution can be found only in some simple cases. More generally, the *EMD* between two sources can be computed by resorting to numerical analysis, and in fact, due to its wide use as a similarity measure in computer vision applications, several efficient algorithms have been proposed (see [109] for example). In the following, we describe a fast iterative algorithm for the computation of the *EMD* between any two sources assuming that the distortion (or cost) function has the general form $d(i, j) = |i - j|^p$, with $p \geq 1$. 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 distance.

5.2.1 Hoffman's greedy algorithm for computing $\mathcal{SM}$

Let us assume that X and Y are discrete sources with alphabets $\mathcal{X}$ and $\mathcal{Y}$. The transportation problem we have to solve for computing $\mathcal{SM}(P_Y, P_X)$, i.e. $EMD(P_Y, P_X)$,

is known in modern literature as *Hitchcock transportation problem* [110][5], which, in turn, can be formulated as a linear programming problem in the following way:

$$EMD(P_X, P_Y) = \min_{S_{XY}} \sum_{i,j} d(i,j) S_{XY}(i,j), \tag{5.20}$$

where S_{XY} must satisfy the linear constraints:

$$\begin{aligned} &\sum_j S_{XY}(i,j) = P_X(i) && \forall i \in \mathcal{X} \\ &\sum_i S_{XY}(i,j) = P_Y(j) && \forall j \in \mathcal{Y} \\ &S_{XY}(i,j) \geq 0 \quad \forall i,j, \end{aligned} \tag{5.21}$$

and where, by referring to the original Monge formulation [94], $S_{XY}(i,j)$ denotes the quantity of soil shipped from location (source) i to location (sink) j and $d(i,j)$ is the cost for shipping a unitary amount of soil from i to j.

A transportation problem (TP) like the one defined by equations (5.20) and (5.21) is a particular minimum cost flow problem [111] which, being linear, can be solved through the simplex method [112]. 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 [113], allows to solve the transportation problem whenever $d(\cdot,\cdot)$ satisfies the so called Monge property [114], that is when:

$$d(i,j) + d(r,s) \leq d(i,s) + d(r,j), \tag{5.22}$$

$\forall (i,j,r,s)$ 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 in general, 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* [113] 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) non-empty bins of $\mathcal{X}$ and $\mathcal{Y}$ respectively. A pseudocode description of the *NWC* rule is given below.

1. Initialize: $i := i^{low}$, $j := j^{low}$.

[5] This is the discrete version of the Monge-Kantorovich mass transportation problem [105].

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 5.2. In the figure, we chose two distributions with disjoint supports for sake of clarity, however the procedure is valid regardless of how the two distributions are spread along the real line. 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 $|\mathcal{X}| + |\mathcal{Y}|$ [6]. This represents a dramatic simplification with respect to the complexity required to solve a general Hitchcock transportation problem (see [115]).

As detailed below, in some cases, it is also possible to derive a closed form expression for the Security Margin.

Uniform sources with different cardinalities

Let X and Y be two uniform pmf's with alphabets $\mathcal{X}$ and $\mathcal{Y}$ such that $|\mathcal{X}| = \alpha|\mathcal{Y}|$, with $\alpha \in \mathbb{N}$. In this case, thanks to Hoffman's algorithm we can express $\mathcal{SM}(P_X, P_Y)$ as:

$$\mathcal{SM}_{L_p^p}(P_X, P_Y) = \frac{1}{|\mathcal{Y}|} \sum_{i=0}^{|\mathcal{X}|-1} \sum_{j=0}^{\alpha-1} (|i^{low} - j^{low}| - j - (\alpha - 1)i)^p, \tag{5.23}$$

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

Security Margin under the L_1 distance

If the distortion function corresponds to the L_1 distance, the *EMD* (and hence the Security Margin) assumes a particularly simple form. Specifically, by applying the flow decomposition principle [116], the Security Margin between P and Q can be

[6] For sake of simplicity, the iterative algorithm described by the pseudocode spans all the bins between the minimum and the maximum non-empty bins. However, in principle, only the values $i \in \mathcal{X}$ and $j \in \mathcal{Y}$ must be considered given that for all the empty bins i and j we have $S_{XY}(i,j) = 0$.

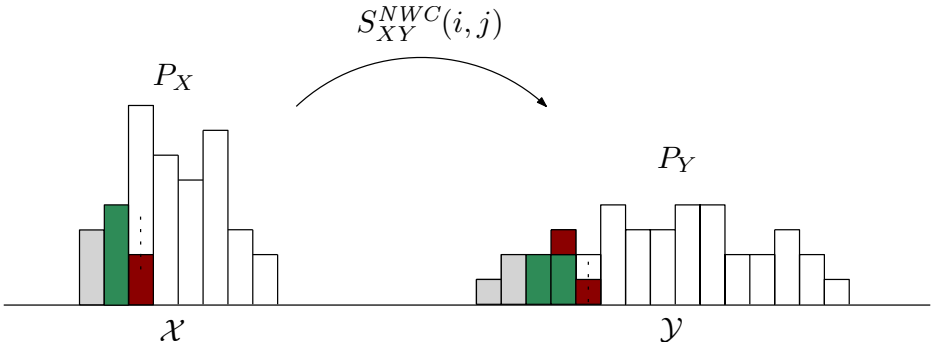


Figure 5.2: Graphical representation of the north-west corner rule for the earth mover transportation problem (Monge problem). P_X and P_Y are two generic earth piles (source and sink) $\mathcal{X}$ and $\mathcal{Y}$, while $S_{XY}^{NWC}(i,j)$ denotes the amount of soil moved from location i to j.

calculated as follows:

$$\mathcal{SM}_{L_1}(P,Q) = \sum_{i=\min\{i^{low},j^{low}\}}^{\max\{i^{up},j^{up}\}} \left| \sum_{s=1}^{i} (P(s) - Q(s)) \right|. \tag{5.24}$$

5.2.2 Security Margin for continuous sources

The analysis carried out in the previous sections is limited to discrete sources. When continuous sources are considered, we can quantize the probability density functions of the sources and apply the analysis for discrete sources. By letting the quantization step tend to zero, the *EMD* between P_X and P_Y can still be regarded as the Security Margin between the two sources. In this case, a general expression for the $\mathcal{SM}$ can be derived by considering the *continuous transportation problem* (CTP), known as Monge-Kantorovic formulation of the mass transportation problem, [105], which is formulated as follows:

$$\mathcal{SM}(P_X,P_Y) = \min_{S_{XY}(x,y)} \iint c(x,y) S_{XY}(x,y) dxdy, \tag{5.25}$$

subject to the constraints

$$\begin{aligned} \int S_{XY}(x,y)dx &= P_Y(y) \\ \int S_{XY}(x,y)dy &= P_X(x) \\ S_{XY}(x,y) &\geq 0 \quad \forall x,y, \end{aligned} \tag{5.26}$$

where $c(x,y)$ is a continuous cost function, $c : X \times Y \to \mathbb{R}$ If $c(x,y)$ satisfies the continuous Monge property, [114], that is if

$$c(x,y) + c(x',y') \leq c(x',y) + c(x,y'), \tag{5.27}$$

for all $x \leq x'$, $y \leq y'$, the optimum transportation map is defined as follows. Let $C_X(x)$ and $C_Y(y)$ be the cumulative distributions of X and Y respectively, and let $C_{XY}(x,y)$ be the cumulative transportation map, that is:

$$C_{XY}(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} S_{XY}(u,v)dudv. \tag{5.28}$$

The *optimum cumulative transportation map* is obtained by letting:

$$C^*_{XY}(x,y) = \min\{C_X(x), C_Y(y)\}, \quad \forall (x,y) \in \mathbb{R}^2, \tag{5.29}$$

which corresponds to the Hoeffding distribution [107]. The continuous map in (5.29) generalizes the *NWC* rule. Therefore, one can compute $\mathcal{SM}(P_Y, P_X)$ by evaluating the mean value $E_{XY}[c(x,y)]$ over the continuous distribution $C^*_{XY}(x,y)$. In general, however, finding a closed form expression is not an easy task.

A particularly simple and insightful formula can be obtained when the cost function corresponds to the squared Euclidean distance. Let us assume, then, that $c(x,y) = (x-y)^2$ and let X and Y be two continuous sources with means μ_X and μ_Y, variances σ_X and σ_Y and covariance $covXY$. As shown in [117] (decomposition theorem), the expectation in (5.2) can be rewritten as follows:

$$E_{XY}[(X-Y)^2] = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2[\sigma_X\sigma_Y - covXY], \tag{5.30}$$

where the three terms express, respectively, the difference in *location*, *spread* and *shape* between the variables X and Y [118]. Interestingly, the covariance $covXY$ is the only term in (5.30) which depends on the joint pdf of X and Y. Then, in order to find the Security Margin, we only need to compute the maximum covariance over all the possible joint pdf's:

$$\mathcal{SM}_{L_2^2}(P_X, P_Y) = (\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 + 2[\sigma_X\sigma_Y - \max_{P_{XY}:\substack{\sum_y P_{XY}=P_X \\ \sum_x P_{XY}=P_Y}} covXY]. \tag{5.31}$$

Since $0 \leq covXY \leq \sigma_X\sigma_Y$, the Security Margin can be bounded as follows:[7]

$$(\mu_X - \mu_Y)^2 + (\sigma_X - \sigma_Y)^2 \leq \mathcal{SM}_{L_2^2}(P_X, P_Y) \leq (\mu_X - \mu_Y)^2 + \sigma_X^2 + \sigma_Y^2. \tag{5.32}$$

[7] We point out that relation (5.30), as well as the bounds in (5.32), holds for the discrete case too.

Accordingly, given two sources X and Y with means μ_X and μ_Y and variances σ_X^2 and σ_Y^2, whenever the distortion introduced by the Attacker is less than the quantity on the left-hand side of (5.32), X and Y are distinguishable asymptotically, regardless of their specific distribution. Instead, if the distortion is above this value, the distinguishability of the two sources depends on their specific probability distributions. Finally, for distortions greater than the quantity on the right-hand side of (5.32), there is no way to distinguish X and Y. When P_X and P_Y have the same form, for instance when the random variables X and Y are both distributed according to a Gaussian or a Laplacian distribution, the Security Margin takes the minimum value and the lower bound in (5.32) holds with equality. In this case, in fact, it is possible to turn P_X into P_Y by imposing a deterministic relationship between X and Y, namely $Y = \frac{\sigma_Y}{\sigma_X}X + (\mu_Y - \frac{\sigma_Y}{\sigma_X}\mu_X)$; in this way, the covariance term is equal to $\sigma_X\sigma_Y$, and hence the contribution of the shape term in the Security Margin vanishes. This is a remarkable result stating that the distinguishability of two sources belonging to the same class depends only on their means and variances, regardless of their particular pdf.

5.3 The Security Margin with L_∞ distance

So far, we have restricted the analysis to the case of additive distortion measures. In this section, we extend the definition of the Security Margin to the case in which the distortion introduced by the Attacker is measured by relying on the L_∞ distance, due to its relevance in practical applications. In our analysis, we will refer to the binary detection game with known sources (DT_{ks} game), the extension to the case with training data being immediate.

By following the same steps of the previous, 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 d_{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 property.

Property 4. *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*

$$\min_{S_{YZ}:S_Y=P,S_Z=Q} \left(\max_{(i,j):S_{YZ}(i,j)\neq 0} |i-j| \right). \tag{5.33}$$

Proof. Let $S^* \neq S_{PQ}^{NWC}$ be a generic transformation mapping 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:

$$\begin{aligned} S'(t,v) &= S^*(t,v) + S^*(t,s) \\ S'(t,s) &= 0 \\ S'(r,v) &= S^*(r,v) - S^*(t,s) \\ S'(r,s) &= S^*(r,s) + S^*(t,s). \end{aligned} \tag{5.34}$$

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:

$$\max_{(i,j):S^*(i,j)\neq 0} |i-j| \geq \max_{(i,j):S'(i,j)\neq 0} |i-j|. \tag{5.35}$$

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 (5.33). □

Thanks to Property 4, the set $\Gamma_{L_\infty}(P_X, \lambda, L)$ in (3.50) can be rewritten as follows:[8]

$$\begin{aligned} \Gamma_{L_\infty}(P_X, \lambda, L) = \{P \in \mathcal{P} :\ &\exists\ Q \in \Lambda^*(P_X, \lambda) \text{ s.t} \\ &\max_{(i,j):S_{PQ}^{NWC}(i,j)\neq 0} |i-j| \leq L\}. \end{aligned} \tag{5.36}$$

By letting λ tend to 0, we obtain the smallest indistinguishability region, thus extending Theorem 7 to the DT_{ks} game with d_{L_∞} distance.

Theorem 9. *Given two sources $X \sim P_X$ and $Y \sim P_Y$ and a maximum allowable per-letter distortion L, and given:*

$$\Gamma(P_X, L) = \{P \in \mathcal{P} : \max_{(i,j):S_{PP_X}^{\mathrm{NWC}}\neq 0} |i-j| \leq L\}, \tag{5.37}$$

the maximum achievable false negative error exponent ε for the DT_{ks} game with L_∞ distance is

$$\lim_{\lambda\to 0} \lim_{n\to\infty} -\frac{1}{n} \log P_{fn} = \min_{P\in\Gamma_{L_\infty}(P_X,L)} \mathcal{D}(P||P_Y). \tag{5.38}$$

[8]The same arguments hold for both the ks and tr cases, so we do not specify the game in the notation of the acceptance region.

Proof. The proof relies on the extension of Property 2 and Lemma 10 to the L_∞ case. The extension of Property 2 is immediate since, once again, 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 10 requires some more care and is proven in Appendix C.2 (Lemma 11). For the rest, the theorem can be proven by reasoning as in the proof of Theorem 7. □

As a consequence of Theorem 9, 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

$$\max_{(i,j)\in S^{NWC}_{P_Y P_X}(i,j)\neq 0} |i-j| > L. \tag{5.39}$$

Condition (5.39) can be used to determine the maximum attacking distortion for which D is able to distinguish the two sources X and Y, i.e. the Security Margin.

Definition 10 (Security Margin for the L_∞ case)**.** *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 DT_{ks} setting with L_∞ distance is given by*

$$\mathcal{SM}_{L_\infty}(P_Y, P_X) = \max_{(i,j):S^{\mathrm{NWC}}_{P_Y P_X}(i,j)\neq 0} |i-j|, \tag{5.40}$$

where $S^{\mathrm{NWC}}_{P_Y P_X}$ is obtained by applying the NWC *rule to map P_Y into P_X.*

It is easy to see that, even if we proved Theorem 9 for the case of known sources, it is possible to extend it to the DT_{tr} game, yielding the same ultimate indistinguishability region.

5.4 Concluding remarks

By interpreting the Attacker's optimum strategy as the solution of an optimum transport problem, in this chapter we have analyzed the ultimate performance of the game, obtained through symmetrization of the decision setup. Then, we have introduced the concept of Security Margin, which is a powerful concept which permits to summarise into a single quantity the the distinguishability of two sources under adversarial conditions. We also described an efficient algorithm to compute the Security Margin between several classes of sources. By relying on the Security Margin concept, we can understand who between the Attacker and the Defender is going to asymptotically *win* the binary detection game. Some insights into the practical use of the $\mathcal{SM}$ in the multimedia forensic scenario are given in Chapter 12.

It is interesting to observe that the analysis carried out in this chapter can be extended in several directions, with different difficulty levels. As a first extension, we mention a scenario in which the system under analysis is observed through a noisy (memoryless) channel. If the Attacker acts after the channel, then $\mathcal{SM}$ can be calculated both at the input and the output of the channel, to measure the security loss caused by the channel. In case the Attacker acts before the channel, the situation is slightly more involved, since the Attacker must take into account the presence of the channel when devising the optimum attack. To calculate the Security Margin, then, we must consider the backward channel having at the input the sequence observed by the Defender and at the output the attacked sequence (a similar approach is used in [119] for biometric identification). As an alternative setup we could consider the effect that the maximum transmission rate allowed by the channel has on source distinguishability, linking the Security Margin to the degradation introduced by the channel in a typical rate distortion setup.

Chapter 6

Detection Games with Corruption of the Training Data

In this chapter, we extend the analysis of the detection game with training data by considering a situation in which the Attacker interferes with the learning phase by corrupting part of the training sequence. Adversarial learning is a rather novel concept, which has been studied for some years from a machine learning perspective [17, 7]. 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 statistics of the to-be-distinguished sources can be seen as a very simple learning mechanism, and 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.

More specifically, we extend the game-theoretic framework introduced in Chapter 4 to model a situation in which the Attacker is given the possibility of corrupting part of the training sequence. 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 analyse the best achievable performance when the Type I and II error probabilities tend to zero exponentially fast. Specifically, we study the distinguishability of the 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. The results of the analysis are summarised in terms of *blinding percentage*, defined as the percentage of corrupted samples making a reliable distinction between the two sources impossible, and Security Margin, i.e., the maximum distortion of the observed sequence for which a reliable distinction is possible (see Chapter 5). The analysis is applied to 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 favourable to the Attacker, since a lower distortion and a lower number of corrupted training samples

are enough to prevent a correct decision.

The organization of the chapter is the following: we formalize the detection problem with addition of corrupted training samples to the original training set and define the game in Section 6.1; the game is solved in Section 6.2. Then, Section 6.3 investigates the ultimate achievable performance of the game. The analysis is extended to the case in which the Attacker replaces part of the training set in Sections 6.4 and 6.5.

6.1 Formalization of the detection game with addition of training samples

In this section, we give a rigorous definition of the detection game with addition of corrupted training samples.

Given a discrete and memoryless source $X \sim P_X$ and a test sequence z^n, the goal of the Defender 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 D must ensure that the false positive error probability of rejecting H_0 when H_0 holds (Type I error) is lower than a given threshold. Similarly to the previous versions of the game, we assume that D relies only on first order statistics to make a decision. For mathematical tractability, likewise in the previous cases, we study the asymptotic version of the game when $n \to \infty$, 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 its side, the Attacker aims at inducing a Type II error. Specifically, A takes a sequence y^n drawn from a source $Y \sim P_Y$ and modifies it in such a way that D decides that the modified sequence z^n has been generated by X. In doing so, A must respect a distortion constraint requiring that the average per-letter distortion between y^n and z^n is lower than L.

Players A and D know the statistics of X through a training sequence, however the training sequence can be partly corrupted by A. Depending on how the training sequence is modified by the Attacker, we can define different versions of the game. We focus on two possible cases: in the first case, hereafter referred to as source identification game with addition of corrupted samples $DT^a_{c\text{-}tr}$, the Attacker can add some fake samples to the original training sequence. In the second case, analysed in Section 6.5, the Attacker can replace some of the training samples with fake values (source identification game with replacement of training samples, namely $DT^r_{c\text{-}tr}$). 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 D knows P_Y. This may seem a too strong assumption, however we will show later on that the optimum strategy of D does not depend on P_Y thus allowing us to relax the assumption that D knows P_Y.

With the above ideas in mind, we are now ready to give a formal definition of the $DT^a_{c\text{-}tr}$ game.

6.1.1 Structure of the $DT^a_{c\text{-}tr}$ game

A schematic representation of the scenario addressed by the $DT^a_{c\text{-}tr}$ game 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 A, who corrupts it by concatenating to it a sequence of fake samples τ^{m_2}. He then reorders the overall sequence in a random way so 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[1]. In the following, 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 portion of fake samples. The corrupted training sequence observed by D is indicated 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. The goal of D is to decide if an observed sequence z^n has been drawn from the same source that generated t^m (H_0) or not (H_1). We assume that D knows that a certain percentage of samples in the training sequence are corrupted, but he has no clue about the position of the corrupted samples. The Attacker can also modify the sequences generated by Y so to induce a decision error. The corrupted sequence is indicated by z^n. With regard to the two phases of the attack, we assume that 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.

6.1.2 Definition of the $DT^a_{c\text{-}tr}$ game

Let us define the set of strategies available to D and A (respectively $\mathcal{S}_D$ and $\mathcal{S}_A$) and the corresponding payoffs.

[1] By using the terminology introduced in [7], the above scenario can be referred to as a *causative* attack with control over training data.

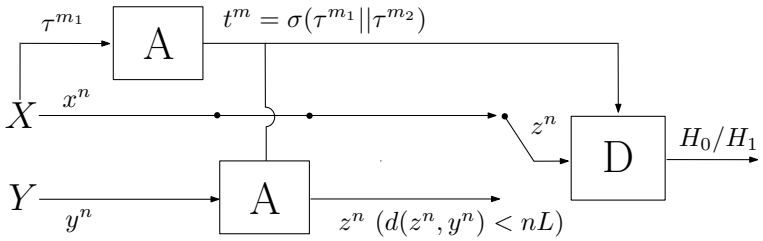


Figure 6.1: Schematic representation of the $DT^a_{c\text{-}tr}$ game. Symbol $||$ denotes concatenation of sequences and σ is a random permutation of sequence samples. Under H_0, $z^n = x^n$, whereas, under H_1, z^n is a distorted version of y^n.

Defender's strategies

The basic assumption behind the definition of the space of strategies available to D is that to make his decision, D relies only on the first order statistics of z^n and t^m. This assumption is equivalent to requiring that the acceptance region for hypothesis H_0, hereafter referred to as $\Lambda^{n\times m}$, is a union of pairs of type classes[2], or equivalently, pairs of types (P, R), where $P \in \mathcal{P}^n$ and $R \in \mathcal{P}^m$. To define $\Lambda^{n\times m}$, D follows a Neyman-Pearson approach, requiring that the false positive error probability is lower than a certain threshold. Specifically, we require 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 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. Given the above setting, the space of strategies of D is defined as follows:

$$\mathcal{S}_D = \{\Lambda^{n\times m} \subset \mathcal{P}^n \times \mathcal{P}^m : \max_{P_X\in\mathcal{P}} \max_{s\in\mathcal{S}_A} P_{\text{FP}} \leq 2^{-\lambda n}\}, \tag{6.1}$$

where the inner maximization is performed over all the strategies available to the Attacker. We will refine this definition at the end of the next section, after the exact definition of the space of strategies of the Attacker ($\mathcal{S}_A$).

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

Attacker's strategies

With regard to A, the attack consists of two parts. Given a sequence y^n drawn from P_Y, and the 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 D. Then he transforms y^n into z^n, 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 proper distortion function d.

Let us consider the corruption of the training sequence first. Given that the Defender bases his decision only on the type of t^m, we are only interested in the effect that the addition of the fake samples in τ^{m_2} has on P_{t^m}. By considering the different length of τ^{m_1} and τ^{m_2}, we find that:

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

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, it is reasonable to assume that 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 in letting the Defender think that a sequence generated by Y has been drawn by the same source that generated t^m.

To describe the part of the attack applied to the test sequence, we follow the usual approach based on transportation theory. Let $S^n_{YZ}(i,j) = n(i,j)/n$ be the relative frequency with which a move from i to j occurs; for any additive distortion measure, we know that the distortion introduced by the attack can be expressed in terms of $n(i,j)$ and S^n_{YZ} as follows:

$$d(y^n, z^n) = \sum_{i,j} n(i,j)d(i,j), \tag{6.3}$$

$$\frac{d(y^n, z^n)}{n} = \sum_{i,j} S^n_{YZ}(i,j)d(i,j). \tag{6.4}$$

where $d(i,j)$ is the distortion introduced when symbol i is transformed into symbol j.

By remembering that $\Lambda^{n\times m}$ depends only on the type of the test sequence, and given that the type of the attacked sequence depends on P_{y^n} only through S^n_{YZ}, we can define the second phase of the attack as the choice of a transportation map among

[3]While reordering is essential to hide the position of fake samples to 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 its decision only on the first order statistic of the observed sequences.

all *admissible* maps, where the set of admissibility maps is defined as in Section 3.4 (equation (3.19)).

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

$$\mathcal{S}_A = \mathcal{S}_{A,T} \times \mathcal{S}_{A,O}, \tag{6.5}$$

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

$$\mathcal{S}_{A,T} = \left\{ Q(P_{\tau^{m_1}}) \in \mathcal{P}^{m_2} \right\}, \tag{6.6}$$

$$\mathcal{S}_{A,O} = \left\{ S^n_{YZ}(P_{y^n}, P_{t^m}) \in \mathcal{A}^n(L, P_{y^n}) \right\}. \tag{6.7}$$

Note that the first part of the attack ($\mathcal{S}_{A,T}$) is applied regardless of whether H_0 or H_1 holds, while the second part ($\mathcal{S}_{A,O}$) 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 both on y^n and τ^{m_1} (through t^m). Finally, we observe that with these 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 $Q(\cdot) \in \mathcal{S}_{A,T}$, since this is the only part of the attack affecting P_{FP}. Specifically, we can rewrite (6.1) as

$$\mathcal{S}_D = \{\Lambda^{n\times m} \subset \mathcal{P}^n \times \mathcal{P}^m : \max_{P_X} \max_{Q(\cdot)\in\mathcal{S}_{A,T}} P_{\text{FP}} \leq 2^{-\lambda n}\}. \tag{6.8}$$

Payoff

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

$$u(\Lambda^{n\times m}, (Q(\cdot), S^n_{YZ}(\cdot,\cdot))) = -P_{\text{FN}}, \tag{6.9}$$

where P_{FN} is the false negative error probability. Of course, D aims at maximising u while A wants to minimise it.

6.1.3 $DT^a_{c\text{-}tr}$ game with targeted corruption ($DT^{a,t}_{c\text{-}tr}$ game)

The $DT^a_{c\text{-}tr}$ game is difficult to solve directly, because of the 2-step attacking strategy. We will work around this difficulty by tackling first with a slightly different version of the game, namely the source identification game with target corruption of the training sequence, $DT^{a,t}_{c\text{-}tr}$, depicted in Figure 6.2.

Whereas the strategies of 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

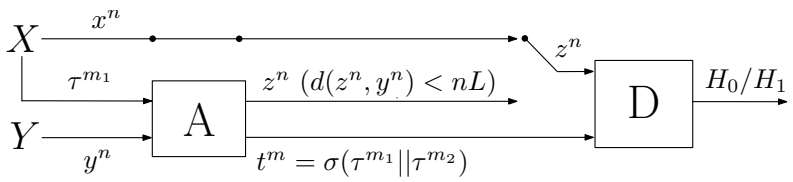


Figure 6.2: $DT^{a}_{c\text{-}tr}$ game with targeted corruption of the training sequence, named $DT^{a,t}_{c\text{-}tr}$ game.

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:

$$\mathcal{S}_{A,T} = \left\{ Q(P_{\tau^{m_1}}, P_{y^n}) \in \mathcal{P}^{m_2} \right\}. \tag{6.10}$$

Even if this setup is not very realistic and is more favourable to the Attacker which can exploit the exact knowledge of y^n, rather than its statistical properties, also for the corruption of the training sequence, it is possible to show a posteriori that, at least for large n, the $DT^{a,t}_{c\text{-}tr}$ game depicted in Figure 6.2 is equivalent to the non-targeted version of the game (see Figure 6.1).

With the above ideas in mind, the $DT^{a,t}_{c\text{-}tr}$ game is formally defined as follows.

Defender's strategies.

$$\mathcal{S}_D = \{\Lambda^{n\times m} \subset \mathcal{P}^n \times \mathcal{P}^m : \max_{P_X} \max_{Q(\cdot,\cdot)\in\mathcal{S}_{A,T}} P_{\text{FP}} \leq 2^{-\lambda n}\}. \tag{6.11}$$

Attacker's strategies.

$$\mathcal{S}_A = \mathcal{S}_{A,T} \times \mathcal{S}_{A,O} \tag{6.12}$$

with:

$$\mathcal{S}_{A,T} = \left\{ Q(P_{\tau^{m_1}}, P_{y^n}) \in \mathcal{P}^{m_2} \right\} \tag{6.13}$$

$$\mathcal{S}_{A,O} = \left\{ S^n_{YZ}(P_{y^n}, P_{\tau^{m_1}}) \in \mathcal{A}^n(L, P_{y^n}) \right\}, \tag{6.14}$$

We notice that, in this game, contrarily to the non-targeted case, the two phases of the attack are coupled.

Payoff.

$$u(\Lambda^{n\times m}, (Q(\cdot,\cdot), S^n_{YZ}(\cdot,\cdot))) = -P_{\text{FN}}. \tag{6.15}$$

6.2 Asymptotic equilibrium and payoff of the $DT^{a,t}_{c\text{-}tr}$ and $DT^{a}_{c\text{-}tr}$ games

In this section, we derive the asymptotic equilibrium point of the $DT^{a,t}_{c\text{-}tr}$ and the $DT^{a}_{c\text{-}tr}$ games when the length of the test and training sequences tends to infinity and evaluate the payoff at the equilibrium.

For the evaluation of the false positive probability, we will see that it is immaterial that the strategy of corruption of the training is targeted or non-targeted, then the optimum strategy for the Defender is the same for both versions of the game.

6.2.1 Optimum defender's strategy

To start with, we look for an explicit expression of the false positive error probability. Such a probability depends on P_X and on the strategy used by A to corrupt the training sequence. In fact, the mapping of y^n into z^n does not have any impact on D's decision under H_0. In the following derivations, we focus on the game with targeted corruption; by showing that the dependence on y^n has no impact on P_{FP}, we deduce, a posteriori, that the same result holds 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 is equal to

$$P_{\text{FP}} = \sum_{P_{y^n}\in\mathcal{P}^n} P_Y(\mathcal{T}(P_{y^n})) \sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} P_X(\mathcal{T}(P_{x^n}))\cdot \tag{6.16}$$

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

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 by X. Given the above formulation, the set of strategies available to D can be rewritten a:

$$\mathcal{S}_D = \Bigg\{\Lambda^{n\times m} : \max_{P_X}\max_{Q(\cdot,\cdot)} \sum_{P_{y^n}\in\mathcal{P}^n} P_Y(\mathcal{T}(P_{y^n}))\cdot \sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}^{n\times m}} \tag{6.17}$$

$$P_X(\mathcal{T}(P_{x^n}))\cdot \sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}})) \le 2^{-\lambda n}\Bigg\}.$$

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

Lemma 6. *Let $\Lambda^{n\times m,*}$ be defined as follows:*

$$\Lambda^{n\times m,*}=\left\{(P_{z^n},P_{t^m})\colon \min_{Q\in\mathcal{P}^{m_2}} h\left(P_{z^n},\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\leq\lambda-\delta_n\right\} \tag{6.18}$$

with

$$\delta_n=|\mathcal{X}|\frac{\log(n+1)((1-\alpha)nc+1)}{n}, \tag{6.19}$$

where $|\mathcal{X}|$ is the cardinality of the source alphabet and where the minimisation 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. $\max\limits_{P_X}\max\limits_{s\in\mathcal{S}_A} P_{FP}\leq 2^{-n(\lambda-\nu_n)}$, *with* $\lim\limits_{n\to\infty}\nu_n=0$,
2. $\forall\Lambda^{n\times m}\in\mathcal{S}_D$, *we have* $\bar{\Lambda}^{n\times m}\subseteq\bar{\Lambda}^{n\times m,*}$,

for both the $DT^{a,t}_{c\text{-}tr}$ and $DT^{a}_{c\text{-}tr}$ games.

Proof. To prove the first part of the lemma we rewrite the false positive error probability as in equation (6.16). We have:

$$\begin{aligned}\max_{P_X}\max_{Q(\cdot,\cdot)} P_{\text{FP}} &= \max_{P_X}\max_{Q(\cdot,\cdot)}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(P_{y^n})\cdot\sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(\mathcal{T}(P_{x^n}))\cdot\\ &\qquad\qquad\sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}}))\\ &\leq \max_{P_X}\sum_{P_{y^n}\in\mathcal{P}^n} P_Y(P_{y^n})\cdot\sum_{\substack{(P_{x^n},P_{t^m})\\ \in\bar{\Lambda}^{n\times m,*}}} P_X(\mathcal{T}(P_{x^n}))\cdot\\ &\qquad\qquad\max_{Q(\cdot,\cdot)}\sum_{\substack{P_{\tau^{m_1}}\in\mathcal{P}^{m_1}:\\ \alpha Q(P_{\tau^{m_1}},P_{y^n})+(1-\alpha)P_{\tau^{m_1}}=P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}})).\end{aligned} \tag{6.20}$$

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]

$$P_X(\mathcal{T}(P_{\tau^{m_1}}))\leq\max_{Q\in\mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right), \tag{6.21}$$

[4]It is easy to see that the bound (6.21) holds also for the non-targeted game, when Q depends on the training sequence only (that is, $Q=Q(P_{\tau^{m_1}})$).

with the understanding that the maximisation 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 this observation, we can upper bound the false positive error probability as follows:

$$\begin{aligned}\max_{P_X} \max_{Q(\cdot,\cdot)} P_{\text{FP}} &\leq \max_{P_X} \sum_{P_{y^n} \in \mathcal{P}^n} P_Y(P_{y^n}) \cdot \sum_{\substack{(P_{x^n}, P_{t^m}) \\ \in \bar{\Lambda}^{n\times m,*}}} P_X(\mathcal{T}(P_{x^n})) \cdot |\mathcal{P}^{m_1}| \cdot \\ &\qquad \max_{Q \in \mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m} - \alpha Q}{1-\alpha}\right)\right) \\ &= \max_{P_X} \sum_{\substack{(P_{x^n}, P_{t^m}) \\ \in \bar{\Lambda}^{n\times m,*}}} P_X(\mathcal{T}(P_{x^n}))|\mathcal{P}^{m_1}| \max_{Q \in \mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m} - \alpha Q}{1-\alpha}\right)\right) \\ &\leq |\mathcal{P}^{m_1}| \sum_{\substack{(P_{x^n}, P_{t^m}) \\ \in \bar{\Lambda}^{n\times m,*}}} \max_{Q \in \mathcal{P}^{m_2}} \max_{P_X} P_X(\mathcal{T}(P_{x^n})) \cdot P_X\left(T\left(\frac{P_{t^m} - \alpha Q}{1-\alpha}\right)\right),\end{aligned} \tag{6.22}$$

where in the equality we exploited the fact that the inner summation does not depend on y^n. It is straightforward to see that the same steps can be repeated for the non-targeted case. From this point, the proof goes along the same line of the proof of Lemma 4 in Chapter 4, by observing that $\max_{P_X} P_X(\mathcal{T}(P_{x^n}))P_X\left(T\left(\frac{P_{t^m}-\alpha Q}{1-\alpha}\right)\right)$ 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 $\bar{\Lambda}^{n\times m,*}$, $h(P_{x^n}, \frac{P_{t^m}-\alpha Q}{1-\alpha})$ is larger than $\lambda - \delta_n$ for every Q.

We now pass to the second part of the lemma. Let $\Lambda^{n\times m}$ be a strategy in $\mathcal{S}_D$, 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, we have (admissibility constraint)

$$2^{-\lambda n} \geq \max_{P_X} \max_{Q(\cdot,\cdot)} \sum_{P_{y^n} \in \mathcal{P}^n} P_Y(P_{y^n}) \cdot \sum_{(P_{x^n}, P_{t^m}) \in \bar{\Lambda}^{n \times m}} P_X(\mathcal{T}(P_{x^n})) \cdot \sum_{\substack{P_{\tau^{m_1}}: \\ \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha) P_{\tau^{m_1}} = P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}})) \tag{6.23}$$

$$\geq \max_{P_X} \max_{Q(\cdot,\cdot)} \sum_{P_{y^n} \in \mathcal{P}^n} P_Y(P_{y^n}) \cdot \left(P_X(\mathcal{T}(P_{x^n})) \cdot \sum_{\substack{P_{\tau^{m_1}}: \\ \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha) P_{\tau^{m_1}} = P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}})) \right)$$

$$= \max_{P_X} \sum_{P_{y^n} \in \mathcal{P}^n} P_Y(P_{y^n}) \cdot \left(P_X(\mathcal{T}(P_{x^n})) \cdot \max_{Q(\cdot,\cdot)} \sum_{\substack{P_{\tau^{m_1}}: \\ \alpha Q(P_{\tau^{m_1}}, P_{y^n}) + (1-\alpha) P_{\tau^{m_1}} = P_{t^m}}} P_X(\mathcal{T}(P_{\tau^{m_1}})) \right)$$

$$\geq \max_{P_X} P_X(\mathcal{T}(P_{x^n})) \max_{Q \in \mathcal{P}^{m_2}} P_X\left(T\left(\frac{P_{t^m} - \alpha Q}{1-\alpha} \right) \right),$$

where the maximization over Q is restricted to the Q's for which $P_{t^m} - \alpha Q \geq 0$ for all the symbols in $\mathcal{X}$ [5]. Since the expression in brackets in the second to the last line is the same for all P_{y^n}, for any $P_{\tau^{m_1}}$ the maximizing $Q^*(\cdot, P_{y^n})$ is independent of P_{y^n}. Then, the same lower bound also holds for the non targeted case.

By exploiting the usual lower bound on 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

$$\begin{aligned} 2^{-\lambda n} &\geq \frac{\max_{P_X} \max_{Q \in \mathcal{P}^{m_2}} 2^{-n\left[\mathcal{D}(P_{x^n}||P_X) + \frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m} - \alpha Q}{1-\alpha}||P_X\right)\right]}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}} \\ &\geq \frac{2^{-n \min_{Q \in \mathcal{P}^{m_2}} \min_{P_X} \left[\mathcal{D}(P_{x^n}||P_X) + \frac{m_1}{n}\mathcal{D}\left(\frac{P_{t^m} - \alpha Q}{1-\alpha}||P_X\right)\right]}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}} \\ &\overset{(a)}{=} \frac{2^{-n \min_{Q \in \mathcal{P}^{m_2}} h\left(P_{x^n}, \frac{P_{t^m} - \alpha Q}{1-\alpha}\right)}}{(n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|}}, \end{aligned} \tag{6.24}$$

[5] The last inequality follows from the fact that the summation over $P_{\tau^{m_1}}$ in the second last line may consists of more than one element.

where (a) derives from the minimization properties of the generalised log-likelihood ratio function h (see Lemma 3 in Chapter 4). By taking the log of both terms we have:

$$\min_{Q\in\mathcal{P}^{m_2}} h\left(P_{x^n}, \frac{P_{t^m}-\alpha Q}{1-\alpha}\right) \geq \lambda - \frac{|\mathcal{X}|\log_2(n+1)(m_1+1)}{n},$$

thus completing the proof of the lemma.

□

Lemma 1 shows that the strategy $\Lambda^{n\times m,*}$ is asymptotically admissible (point 1.) and optimal (point 2.), regardless of the attack. From a game-theoretic perspective, $\Lambda^{n\times m,*}$ is a dominant strategy for D and then the game is dominance solvable.

Despite the existence of a dominant strategy for the Defender, the identification of the optimum Attacker's strategy for the $DT^{a}_{c\text{-}tr}$ 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; then, we will use such a result to analyze the performance in the case of non non-targeted attack.

6.2.2 The $DT^{a,t}_{c\text{-}tr}$ game: asymptotic equilibrium

Given the dominant defense strategy, it is easy to show that, for any given τ^{m_1} and y^n, the optimum Attacker's strategy for the $DT^{a,t}_{c\text{-}tr}$ game boils down to the following double minimisation:

$$(Q^*(P_{\tau^{m_1}}, P_{y^n}),\ S^{n,*}_{YZ}(P_{y_n}, P_{\tau^{m_1}})) = \tag{6.25}$$
$$\underset{\substack{Q\in\mathcal{P}^{m_2}\\ S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})}}{\arg\min} \left(\min_{Q'} h\left(P_{z^n}, \frac{(1-\alpha)P_{\tau^{m_1}}+\alpha Q-\alpha Q'}{1-\alpha}\right)\right),$$

where P_{z^n} (i.e. S^n_Z) is obtained by applying the transformation map S^n_{YZ} to P_{y^n}. As usual, the minimisation over Q' is limited to the Q' such that all the entries of the resulting pmf are nonnegative.

As a remark, for $L=0$ (corruption of the training only), we get:

$$Q^*(P_{\tau^{m_1}}, P_{y^n}) = \underset{Q\in\mathcal{P}^{m_2}}{\arg\min}\min_{Q'} h\left(P_{y^n}, P_{\tau^{m_1}}+\frac{\alpha}{1-\alpha}(Q-Q')\right), \tag{6.26}$$

while, for $\alpha=0$ (classical setup, without corruption of the training sequence) we have:

$$S^{n,*}_{YZ}(P_{y^n}, P_{t^m}) = \underset{S^n_{YZ}\in\mathcal{A}^n(L,P_{y^n})}{\arg\min} h(P_{z^n}, P_{t^m}), \tag{6.27}$$

and we fall back to the known case of detection game with uncorrupted training, studied in Chapter 4. Given the optimum strategies of both players, it is immediate to state the following:

Theorem 10. *The $DT^{a,t}_{c\text{-}tr}$ game is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile $(\Lambda^{n\times m,*}, (Q^*(P_{\tau^{m_1}}, P_{y^n}),\ S^{n,*}_{YZ}(\cdot,\cdot)))$.*

Proof. the theorem is a direct consequence of the fact that $\Lambda^{n\times m,*}$ is a dominant strategy for D. □

Given the optimum attack strategy in (6.25), it is straightforward to define the set of pairs $(P_{y^n}, P_{\tau^{m_1}})$ for which, as a consequence of A's action, D accepts H_0:

$$\Gamma^n(\lambda,\alpha,L) = \{(P_{y^n}, P_{\tau^{m_1}}) : \exists S^n_{YZ} \in \mathcal{A}(L, P_{y^n}) \text{ and } \exists Q \in \mathcal{P}^{m_2} \text{ s.t. } (P_{z^n}, (1-\alpha)P_{\tau^{m_1}} + \alpha Q) \in \Lambda^{n\times m,*}\}. \tag{6.28}$$

By fixing the type of the non-corrupted training sequence $(P_{\tau^{m_1}})$ we get:

$$\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L) = \left\{P_{y^n} \in \mathcal{P}^n : \exists S^n_{YZ} \in \mathcal{A}(L, P_{y^n}) \text{ and } \exists Q \in \mathcal{P}^{m_2} \text{ s.t. } P_{z^n} \in \Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}} + \alpha Q)\right\}, \tag{6.29}$$

where $\Lambda^{n,*}(P)$ denotes the acceptance region for a fixed training type P in $\mathcal{P}^m$. It is interesting to notice that, since in the current setting 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 $\Lambda^{n,*}(P_{t^m})$, for a given P_{t^m}, and the acceptance region itself is modified so to facilitate the former action.

6.2.3 The $DT^{a,t}_{c\text{-}tr}$ game: payoff at the equilibrium

In this section we study the asymptotic payoff of the $DT^{a,t}_{c\text{-}tr}$ game at the equilibrium, thus trying to understand who and under which conditions is going to *win* the game. To do so, we first reformulate the set in (6.29) in a more convenient way. We rewrite region $\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L)$ as follows:

$$\Gamma^n(P_{\tau^{m_1}},\lambda,\alpha,L) = \{P \in \mathcal{P}^n : \exists S^n_{PV} \in \mathcal{A}(L,P) \text{ s.t. } V \in \Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha)\}, \tag{6.30}$$

where

$$\Gamma^n_0(P_{\tau^{m_1}},\lambda,\alpha) = \{P \in \mathcal{P}^n : \exists Q \in \mathcal{P}^{m_2} \text{ s.t. } P \in \Lambda^{n,*}((1-\alpha)P_{\tau^{m_1}} + \alpha Q)\}, \tag{6.31}$$

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 (6.30) when A

cannot modify the sequence drawn from Y (i.e. $L = 0$) and then tries to hamper the decision only by corrupting the training sequence.

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:

$$\Gamma_0^n(P_{\tau^{m_1}}, \lambda, \alpha) = \{P \in \mathcal{P}^n : \exists Q, Q' \in \mathcal{P}^{m_2} \text{ s.t.} \quad h\left(P, P_{\tau^{m_1}} + \frac{\alpha}{(1-\alpha)}(Q - Q')\right) \leq \lambda - \delta_n\Big\}, \tag{6.32}$$

where the second argument of h denotes the generic 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. It is worth observing that since we are treating the $DT_{c\text{-}tr}^{a,t}$ game, 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 a path similar to that used in the analysis of the DT_{ks} and DT_{tr} game in Chapter 3 and 4, respectively. First of all we generalize the definition of the sets $\Lambda^{n\times m,*}$, and then Γ_0^n, so that set Γ^ncan be evaluated for generic pmf's in $\mathcal{P}$ (that is, without requiring that the pmf is induced by sequences of a given 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:

$$h_c(P, P') = \mathcal{D}(P||U) + c\mathcal{D}(P'||U); \tag{6.33}$$
$$U = \frac{1}{1+c}P + \frac{c}{1+c}P'.$$

where $c \in [0, 1]$. The asymptotic versions of Γ^n is then obtained from (6.30) and (6.31) (or (6.32)) by considering h_c and letting $n \to \infty$ as follows:

$$\Gamma(R,\lambda, \alpha, L) = \{P \in \mathcal{P} : \exists S_{PV} \in \mathcal{A}(L, P) \text{ s.t. } V \in \Gamma_0(R, \lambda, \alpha)\}, \tag{6.34}$$

where[6]

$$\begin{aligned}\Gamma_0(R, \lambda, \alpha) =& \{P \in \mathcal{P} : \exists Q \in \mathcal{P} \text{ s.t. } P \in \Lambda^*((1-\alpha)R + \alpha Q)\} \\ =& \{P \in \mathcal{P} : \exists Q, Q' \in \mathcal{P} \text{ s.t. } h_c\left(P, R + \frac{\alpha}{(1-\alpha)}(Q - Q')\right) \leq \lambda\}. \end{aligned} \tag{6.35}$$

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

[6]We remind that the definitions of $S_{PV}(i, j)$ ad $\mathcal{A}(L, P)$ derive from those of $S_{PV}^n(i, j)$ and $\mathcal{A}^n(L, P)$ by relaxing the requirement that the terms $S_{PV}(i, j)$ and $P(i)$ are rational number with denominator n.

Theorem 11 (Asymptotic payoff of the $DT^{a,t}_{c\text{-}tr}$ game)**.** *For the $DT^{a,t}_{c\text{-}tr}$ game, the false negative error exponent at the equilibrium is given by*

$$\varepsilon = \min_{R}[(1-\alpha)c\mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)]. \tag{6.36}$$

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. Let us consider

$$\begin{aligned} P_{\text{FN}} &= \sum_{(P_{y^n},P_{\tau^{m_1}})\in\Gamma^n(\lambda,\alpha,L)} P_X(\mathcal{T}(P_{\tau^{m_1}}))P_Y(\mathcal{T}(P_{y^n})) \\ &= \sum_{R\in\mathcal{P}^{m_1}} P_X(\mathcal{T}(R)) \sum_{P\in\Gamma^n(R,\lambda,\alpha,L)} P_Y(\mathcal{T}(P)) \qquad (6.37) \\ &= \sum_{R\in\mathcal{P}^{m_1}} P_X(\mathcal{T}(R))P_Y(P\in\Gamma^n(R,\lambda,\alpha,L)). \qquad (6.38) \end{aligned}$$

We start by deriving an upper-bound of the false negative error probability. We can write:

$$\begin{aligned} P_{\text{FN}} &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(\mathcal{T}(R)) \sum_{P\in\Gamma^n(R,\lambda,\alpha,L)} 2^{-n\mathcal{D}(P||P_Y)} \\ &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(\mathcal{T}(R))(n+1)^{|\mathcal{X}|}2^{-n\min\limits_{P\in\Gamma^n(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)} \\ &\leq \sum_{R\in\mathcal{P}^{m_1}} P_X(\mathcal{T}(R))(n+1)^{|\mathcal{X}|}2^{-n\min\limits_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)} \\ &\leq (n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|} \\ &\quad \cdot 2^{-n\min\limits_{R\in\mathcal{P}^{m_1}}[\frac{m_1}{n}\mathcal{D}(R||P_X)+\min\limits_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)]} \\ &\leq (n+1)^{|\mathcal{X}|}(m_1+1)^{|\mathcal{X}|} \\ &\quad \cdot 2^{-n\min\limits_{R\in\mathcal{P}}[(1-\alpha)c\mathcal{D}(R||P_X)+\min\limits_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)]}, \qquad (6.39) \end{aligned}$$

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:

$$-\frac{\log P_{\text{FN}}}{n} \geq \min_{R\in\mathcal{P}}\left[(1-\alpha)c\mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)}\mathcal{D}(P||P_Y)\right] + \beta_n, \tag{6.40}$$

with $\beta_n = |\mathcal{X}|\frac{\log(n+1)((1-\alpha)nc+1)}{n}$ tending to 0 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 in (6.36). Due to the density of rational numbers within real numbers, we can find a sequence of pmf's $R_n \in \mathcal{P}_n$ that tends to R^* when n tends to infinity. By remembering that $m_1 = (1-\alpha)nc$, the subsequence $R_{m_1} = R_{(1-\alpha)cn}$ will also tend to R^* when n (and hence m_1) tends to infinity [7]. We can write:

$$\begin{aligned} P_{\text{FN}} &= \sum_{R \in \mathcal{P}^{m_1}} P_X(\mathcal{T}(R)) P_Y(\Gamma^n(R, \lambda, \alpha, L)) \\ &\geq P_X(\mathcal{T}(R_{m_1})) P_Y(\Gamma^n(R_{m_1}, \lambda, \alpha, L)), \\ &\geq \frac{2^{-m_1 \mathcal{D}(R_{m_1}||P_X)}}{(m_1+1)^{|\mathcal{X}|}} P_Y(\Gamma^n(R_{m_1}, \lambda, \alpha, L)), \end{aligned} \tag{6.41}$$

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 from the usual lower bound on the probability of a type class [90]. From (6.41), by taking the log and dividing by n we obtain

$$-\frac{\log P_{\text{FN}}}{n} \leq (1-\alpha)c\mathcal{D}(R_{m_1}||P_X) - \frac{1}{n}\log P_Y(\Gamma^n(R_{m_1}, \lambda, \alpha, L)) + \beta'_n, \tag{6.42}$$

where, as in (6.40), $\beta'_n = |\mathcal{X}|\frac{\log((1-\alpha)cn+1)}{n}$ tends to 0 when n tends to infinity. In order to compute the probability term $P_Y(P \in \Gamma^n(R_{m_1}, \lambda, \alpha, L))$ in (6.42), we resort to the extension of Sanov limit given by Theorem A8 (see Appendix A). To do so, we must show that $\Gamma^n(R_{m_1}, \lambda, \alpha, L) \xrightarrow{H} \Gamma(R^*, \lambda, \alpha, L)$.

By exploiting the continuity of the h_c function and the density of rational numbers into the real ones, it is not difficult to see that $\Gamma_0^n(R_{m_1}, \lambda, \alpha, L) \xrightarrow{H} \Gamma_0(R^*, \lambda, \alpha, L)$. The Hausdorff convergence of $\Gamma^n(R_{m_1}, \lambda, \alpha, L)$ to $\Gamma(R^*, \lambda, \alpha, L)$ follows from the regularity property of the admissibility set (see Appendix B). Therefore, we can apply the generalized Sanov theorem and obtain:

$$-\lim_{n\to\infty} \frac{1}{n}\log P_Y(\Gamma^n(R_{m_1}, \lambda, \alpha, L)) = \min_{P \in \Gamma(R^*, \lambda, \alpha, L)} \mathcal{D}(P||P_Y). \tag{6.43}$$

Then we have

$$-\frac{1}{n}\log P_Y(\Gamma^n(R_{m_1}, \lambda, \alpha, L)) \leq \min_{P \in \Gamma(R^*, \lambda, \alpha, L)} \mathcal{D}(P||P_Y) + \alpha'_n, \tag{6.44}$$

[7] Similarly to what we did in the case of uncorrupted training data studied in Chapter 4, we assume that αc (an then also $(1-\alpha)c$), is a non-null integer value. The analysis can be extended, with some care, to the case in which such an assumption does not hold.

where $\alpha'_n \to 0$ as $n \to \infty$. By exploiting the continuity of the divergence function, we can write

$$\mathcal{D}(R_{m_1}||P_X) \leq \mathcal{D}(R^*||P_X) + \alpha''_n, \tag{6.45}$$

where α''_n is arbitrarily small for large enough n. Then, going on from (6.42), we get

$$-\frac{\log P_{\text{FN}}}{n} \leq (1-\alpha)c\mathcal{D}(R^*||P_X) + (1-\alpha)\alpha''_n + \min_{P\in\Gamma(R^*,\lambda,\alpha,L)} \mathcal{D}(P||P_Y) + \alpha'_n + \beta'_n, \tag{6.46}$$

where the quantity $(1-\alpha)\alpha''_n + \alpha'_n + \beta'_n$ tends to zero when n tends to infinity.

By coupling equations (6.40) and (6.46) and by letting $n \to \infty$, we eventually obtain:

$$-\lim_{n\to\infty} \frac{\log P_{\text{FN}}}{n} = \min_{R}[(1-\alpha)c \cdot \mathcal{D}(R||P_X) + \min_{P\in\Gamma(R,\lambda,\alpha,L)} \mathcal{D}(P||P_Y)], \tag{6.47}$$

thus proving the theorem.

□

As an immediate consequence of Theorem 11, 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 A is able to induce D to decide in favour of H_0 even if H_1 holds.

We conclude this section by observing that the asymptotic version of the optimum Attacker's strategy does not depend anymore on the to-be-attacked sequence y^n. In fact, the Attacker needs only to find a pmf Q' which modifies the acceptance region $\Lambda^n(P_X)$ in such a way that it is possible to find an admissible transportation map moving P_Y within it. Accordingly, the optimum corruption strategy depends on P_Y rather than P_{y^n}. In hindsight, the reason for such a result is that, due to the law of large numbers, the type of the sequences generated by Y will tend to P_Y in probability hence making it possible to the Attacker to rely only on the knowledge of P_Y. This suggests that the asymptotic performance of the game remains the same in the case of non-targeted attack (depicted in Figure 6.1). Therefore, the indistinguishability region of the game $DT^{a}_{c\text{-}tr}$ game and $DT^{a,t}_{c\text{-}tr}$ game are the same. The rigorous proof of such argument is very technical and is omitted.

In the other version of the game with corrupted training studied in Section 6.5, we will focus on the case of targeted attack only, keeping in mind that the performance of the game in the non-targeted case are asymptotically equivalent.

6.3 Source distinguishability for the $DT^{a}_{c\text{-}tr}$ (and $DT^{a,t}_{c\text{-}tr}$) game

In this section we study the behaviour of the game when we vary the decay rate of the false positive probability, that is λ. In this way, we derive the best achievable performance of the Defender, by requiring only that P_{FP} tends to zero exponentially fast. Then, we use such a result to derive the limit conditions for which the reliable distinction between two sources is possible in terms of percentage of corrupted training samples α and maximum allowed distortion L.

We point out that, since the $DT^{a}_{c\text{-}tr}$ and $DT^{a,t}_{c\text{-}tr}$ games have the same indistinguishability region, the arguments of this section hold for both versions of the game.

6.3.1 Ultimately achievable performance of the game

As we said, the goal of this section is to study the limit of the indistinguishability region when $\lambda \to 0$. As we know from Chapter 5, this limit determines all the pmf's P_Y that cannot 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).

To this aim, we first observe that optimal transport theory permits us to rewrite the indistinguishability region $\Gamma(P_X, \lambda, \alpha, L)$ as:

$$\Gamma(P_X, \lambda, \alpha, L) = \{P : \exists V \in \Gamma_0(P_X, \lambda, \alpha) \text{ s.t. } \mathit{EMD}(P, V) \leq L\}, \tag{6.48}$$

where *EMD* is the Earth Mover Distance, i.e., the minimum transportation cost (see Chapter 5), that is

$$\mathit{EMD}(P, V) = \min_{S_{PV}: S_P = P, S_V = V} \sum_{i,j} S_{PV}(i, j) d(i, j). \tag{6.49}$$

With this definition, the main result of this section is stated by the following theorem.

Theorem 12. *Given two sources $X \sim P_X$ and $Y \sim P_Y$, a maximum allowed average per-letter distortion L and the fraction α of training samples provided by the Attacker, the maximum achievable false negative error exponent ε for the $DT^{a}_{c\text{-}tr}$ game is:*

$$\lim_{\lambda \to 0} \lim_{n \to \infty} -\frac{1}{n} \log P_{FN} = \min_{R} [\mathcal{D}(R||P_X) + \min_{P \in \Gamma(R, \alpha, L)} \mathcal{D}(P||P_Y)], \tag{6.50}$$

where $\Gamma(R, \alpha, L) = \Gamma(R, \lambda = 0, \alpha, L)$. Then the ultimate indistinguishability region is

$$\Gamma(P_X, \alpha, L) = \{P : \exists V \in \Gamma_0(P_X, \alpha) \text{ s.t. } \mathrm{EMD}(P, V) \leq L\}, \tag{6.51}$$

where $\Gamma_0(P_X,\alpha)=\Gamma_0(P_X,\lambda=0,\alpha)$. *Moreover, such a region can be rewritten as*

$$\begin{aligned}\Gamma(P_X,\alpha,L)=&\left\{P:\min_{V:\mathrm{EMD}(P,V)\leq L}\sum_i[V(i)\text{-}P_X(i)]_+\leq\frac{\alpha}{(1-\alpha)}\right\}\\=&\left\{P:\min_{V:\mathrm{EMD}(P,V)\leq L}d_{L_1}(V,P_X)\leq\frac{2\alpha}{(1-\alpha)}\right\}.\end{aligned}\tag{6.52}$$

Proof. The proof of the first part goes along the same steps used in the proof of Theorem 8 in Chapter 5 and is not repeated here. Instead, we show that the set $\Gamma(P_X,\alpha,L)$ in (6.51) can be rewritten as in (6.52). From the previous analysis, by observing that $h_c(P,Q)=0$ if and only if $P=Q$, it is easy to argue that the set $\Gamma_0(P_X,\alpha)$ takes the following expression:

$$\Gamma_0(P_X,\alpha)=\{P:\exists Q,Q'\in\mathcal{P}\text{ s.t. }P=P_X+\frac{\alpha}{(1-\alpha)}(Q-Q')\}.\tag{6.53}$$

Equation (6.53) can be rewritten by avoiding the reference to 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 vice versa). Since Q and Q' must be valid pmf's, we must 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:

$$\begin{aligned}\Gamma_0(P_X,\alpha)&=\left\{P:\sum_i[P(i)-P_X(i)]_+\leq\frac{\alpha}{(1-\alpha)}\right\}\\&=\left\{P:d_{L_1}(P,P_X)\leq\frac{2\alpha}{(1-\alpha)}\right\},\end{aligned}\tag{6.54}$$

where d_{L_1} denotes the L_1 distance. Equation (6.52) follows immediately from this way of writing $\Gamma_0(P_X,\alpha)$. □

According to Theorem 12, $\Gamma(P_X,\alpha,L)$ provides the *ultimate indistinguishability region* of the test, that is the set of all the pmf for which D will be defeated. Before going on, we discuss the geometrical meaning of the set $\Gamma_0(P_X,\alpha)$ in (6.53). To do so, we rewrite $\Gamma_0(P_X,\alpha)$ as follows:

$$\Gamma_0(P_X,\alpha)=\{P:\exists Q\in\mathcal{P}\text{ s.t. }P\in\Lambda^*_{\lambda\to 0}((1-\alpha)P_X+\alpha Q)\},\tag{6.55}$$

where $\Lambda^*_{\lambda\to 0}(P)$ plays the role of the *ultimate* acceptance region of the test and derives from the asymptotic version of the acceptance region $\Lambda^*(P)$ by letting λ go to 0:

$$\Lambda^*_{\lambda\to 0}(P)=\left\{P':\exists Q'\in\mathcal{P}\text{ s.t. }P'=\frac{P-\alpha Q'}{(1-\alpha)}\right\}.\tag{6.56}$$

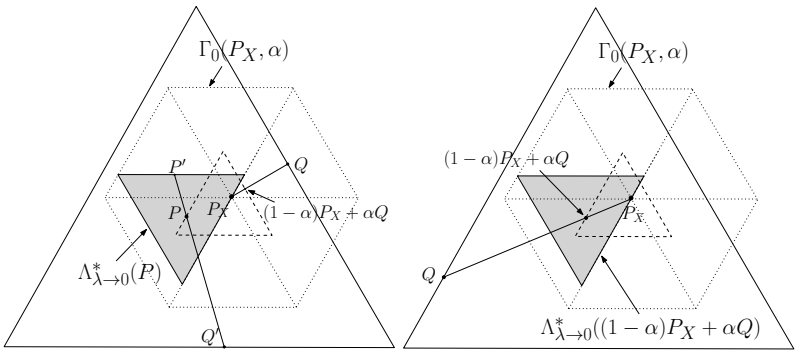


Figure 6.3: Geometrical interpretation of $\Lambda^*_{\lambda\to 0}(P)$ (left) and geometrical construction of $\Gamma_0(P_X,\alpha)$ (right). The size of the sets are exaggerated for graphical purposes.

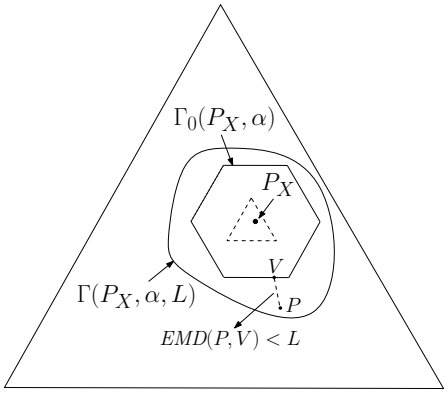


Figure 6.4: Geometrical interpretation of Theorem 12.

By referring to Figure 6.3 (left part) we can geometrically interpret $\Lambda^*_{\lambda\to 0}(P)$ as the set of the points P' such that P is a convex combination (with coefficient α) of P' with a point Q' of the probability simplex. Then, according to (6.55), $\Gamma_0(P_X,\alpha)$ is geometrically obtained as the union of the acceptance regions built starting from the points which can be written as a convex combination of P_X with some point Q in the simplex; this corresponds to an hexagonal region around P_X which, in the probability simplex, is equivalent to the set of the points whose L_1 distance from P_X is constrained to $2\alpha/(1-\alpha)$ (as stated in (6.52)). Obviously, only the points which lie inside the simplex are valid pmf's and then must be accounted for. The geometric interpretation of set $\Gamma_0(P_X,\alpha)$ in (6.55) is given in the right part of Figure 6.3. A pictorial representation of $\Gamma(P_X,\alpha,L)$ is given in Figure 6.4 for a smaller value of α.

6.3.2 Security margin and blinding corruption level (α_b)

By a closer inspection of the *ultimate indistinguishability region* $\Gamma(P_X, \alpha, L)$, we can derive some interesting parameters characterizing the distinguishability of two sources in adversarial setting (both with or without corrupted training, the latter case corresponding to $\alpha = 0$.). Let $X \sim P_X$ and $Y \sim P_Y$ be two sources. Let us focus first on the case in which the Attacker cannot modify the test sequence ($L = 0$). In this situation, the ultimate indistinguishability region boils down to $\Gamma_0(P_X, \alpha)$. We conclude that D can tell the two sources apart if $d_{L_1}(P_Y, P_X) > \frac{2\alpha}{(1-\alpha)}$. On the contrary, if $d_{L_1}(P_Y, P_X) \leq \frac{2\alpha}{(1-\alpha)}$, A is able to make the sources indistinguishable by corrupting the training sequence. Clearly, the larger the α the easier is for A to win the game. By adopting a different perspective, we can define the *blinding corruption level* α_b, that is the corruption percentage for which two sources X and Y cannot be distinguished. Specifically, we have:

$$\begin{aligned} \alpha_b(P_X, P_Y) &= \frac{\sum_i [P_Y(i) - P_X(i)]_+}{1 + \sum_i [P_Y(i) - P_X(i)]_+} \\ &= \frac{d_{L_1}(P_Y, P_X)}{2 + d_{L_1}(P_Y, P_X)}. \end{aligned} \tag{6.57}$$

Since $d_{L_1}(P, Q) \leq 2$ for any pair of pmf's (P, Q), from (6.57) it is easy to see that α_b is always lower that $1/2$. The limit situation $\alpha_b = 1/2$ corresponds to a case in which P_X and P_Y have completely disjoint supports (and hence, $d_{L_1}(P, Q) = 2$). As a result, for $\alpha \geq 1/2$, that is when A corrupts more than half of the training samples, there is always a choice of the pmf in $\mathcal{P}^{m_2}$ ($m \geq m/2$) for which no original sample remains in the training subsequence analyzed by D, hence making a reliable decision impossible (because of the worst case approach adopted by D over the strategies of corruption of A).

Let us now consider the more general case in which $L \neq 0$. For a given $\alpha < \alpha_b$, we look for the maximum attacking distortion for which it is possible to reliably distinguish between the two sources. By inspection of the ultimate indistinguishability region in (6.51), it is easy to argue that the Defender is able to distinguish X and Y, despite the attack, if $\min_{V:EMD(P_Y,V)\leq L} d_{L_1}(R, P_X) > \frac{2\alpha}{(1-\alpha)}$. This leads to the following definition, which extends the concept of Security Margin, introduced in Chapter 5, to the more general setup considered here.

Definition 11 (Security Margin in the $DT^{a}_{c\text{-}tr}$ 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 $DT^{a}_{c\text{-}tr}$ setup is called Security Margin and is given by*

$$\mathcal{SM}_\alpha(P_X, P_Y) = L^*_\alpha, \tag{6.58}$$

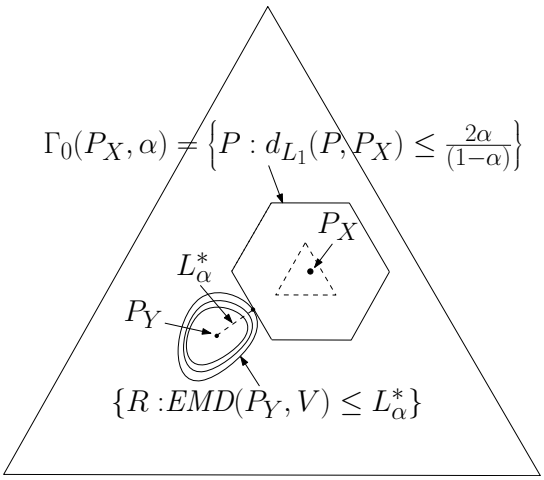


Figure 6.5: Geometrical interpretation of the Security Margin between two sources X and Y.

where $L^*_\alpha = 0$ *if* $P_Y \in \Gamma_0(P_X,\alpha)$[8]*, whereas, if* $P_Y \notin \Gamma_0(P_X,\alpha)$*,* L^*_α *is the quantity which satisfies*

$$\min_{V:\mathrm{EMD}(P_Y,V)\leq L^*_\alpha} d_{L_1}(V,P_X) = \frac{2\alpha}{(1-\alpha)}. \tag{6.59}$$

A geometric interpretation of the Security Margin is given in Figure 6.5.

By focusing on the case $P_Y \notin \Gamma_0(P_X,\alpha)$, since the left-hand side of (6.59) is a monotonic non-increasing function of L_α, the Security Margin $\mathcal{SM}_\alpha(P_X,P_Y)$ can be expressed in explicit form as

$$\arg\min_{L_\alpha} \min_{V:EMD(P_Y,V)\leq L_\alpha} \left| d_{L_1}(V,P_X) - \frac{2\alpha}{(1-\alpha)} \right|. \tag{6.60}$$

When $L > \mathcal{SM}_\alpha(P_X,P_Y)$, it is not possible for D to distinguish between the two sources with positive error exponents of the two kinds.

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 corrupted percentage of samples reaches the critical value, the sources cannot be distinguished even if the Attacker does not introduce any distortion. On the contrary, setting $\alpha = 0$ corresponds to study the distinguishability of the sources with uncorrupted training, in which case we have $\mathcal{SM}_0(P_X,P_Y) = EMD(P_X,P_Y)$, in agreement with (5.10).

[8]The corruption of the training with parameter α already makes the sources undistinguishable.

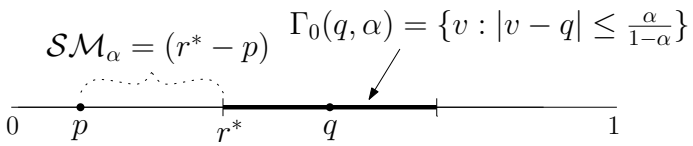


Figure 6.6: Geometrical interpretation of the Security Margin between X and Y. When $\alpha = 0$, $\Gamma_0(q,\alpha)$ boils down to point p and $\mathcal{SM} = (q-p)$ (see Section 5.1).

With reference to Figure 6.5, it is easy to see that when $\alpha = 0$ the hexagonal set (i.e., the indistinguishability region determined by the corruption of the training) boils down to the single point P_X and the Security Margin corresponds to the Earth Mover Distance between Y and X. Moreover, we notice that, for any $\alpha > 0$, the Security Margin in (6.60) is less than $EMD(P_X, P_Y)$. This is also an expected behavior since the general setting considered here is more favorable to the Attacker, with respect to the setting considered in the previous chapters.

Bernoulli sources

In order to get some insights about the practical meaning of the analysis carried out in the previous sections and the parameters 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) percentage of samples that A has to modify for inducing a decision error is, according to (6.57), $\alpha_b = \frac{|p-q|}{1+|p-q|}$. As suggested by intuition, when $|p-q| = 1$, in order for A to win the game, the number of fake samples should be equal to the number of samples of the correct training sequence (i.e. $\alpha = 0.5$). When some distortion is allowed ($L \neq 0$), we have

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p| - \frac{\alpha}{1-\alpha} & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b \end{cases} . \tag{6.61}$$

The geometrical meaning of (6.61) is illustrated in Figure 6.6 for two generic Bernoulli sources with $p > q$ (w.l.o.g.). If $\alpha = 0$, we get the same expression of the Security Margin for the uncorrupted training case, derived in Chapter 5, Section 5.1. Figure 6.7 depicts the behavior of the $\mathcal{SM}_\alpha(p,q)$ as a function of α when $p = 0.3$ and $q = 0.7$.

6.4 The $DT^a_{c\text{-}tr}$ game: an alternative perspective

In the adversarial setup considered in the previous section (and depicted in Figure 6.1), the Attacker adds a sequence of m_2 fake samples, τ^{m_2}, to an existing sequence

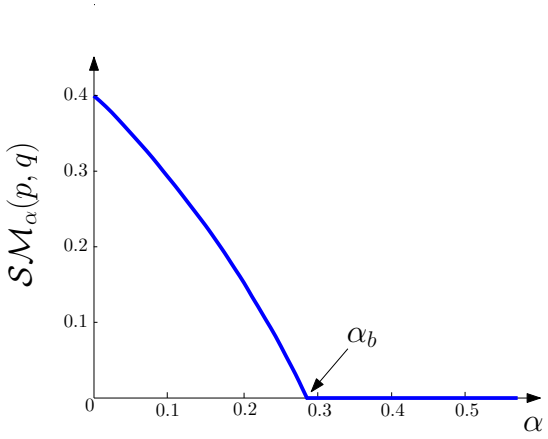


Figure 6.7: Security margin as a function of α for Bernoulli sources with parameters $p = 0.3$ and $q = 0.7$ ($\alpha_b = 0.286$).

of m_1 training sample, τ^{m_1}, and produce (after a random reordering σ) the corrupted training sequence t^m: formally, $t^m = \sigma(\tau^{m_1}||\tau^{m_2})$. It is worth noting that, due to the memoryless nature of the source X, such a scenario is equivalent to the following: the Attacker observes a training sequence τ^m and replaces a certain number m_2 of samples randomly chosen to produce the final corrupted sequence t^m. As before we assume that the Defender does not know the position of attacked samples

Let $\mathcal{M}$ denote the subset of m_2 indexes corresponding to the positions of the samples which the Attacker may corrupt. We indicate with $\tau_{\mathcal{M}}^{m_2}$ the subsequence formed by the samples indexed by $\mathcal{M}$. Hence, $\tau^m = \sigma^*(\tau_{\overline{\mathcal{M}}}^{m_1}||\tau_{\mathcal{M}}^{m_2})$ for some permutation σ^*, where $\overline{\mathcal{M}}$ indicates the complementary set of $\mathcal{M}$. Let ν^{m_2} the sequence of the corrupted samples which the Attacker replace to the original samples in the positions indicated by $\mathcal{M}$. Therefore, the corrupted training sequence observed by D is $t^m = \sigma^*(\tau_{\overline{\mathcal{M}}}^{m_1}||\nu^{m_2})$. This setup is represented by the general scheme illustrated in Figure 6.8.

It is straightforward to be convinced that, when the Attacker cannot choose the indexing set $\mathcal{M}$, the game with addition of fake samples and the one with replacement of random samples with fake ones are indeed equivalent. In fact, in both cases, the resulting sequence that the Defender observes is composed by m_1 original samples drawn from X and m_2 corrupted samples in unknown positions. Assuming that the Defender has no hint on how the Attacker replace the samples, it is easy to argue that the decision strategy does not change with respect to the previous case. On the other side, since the goal of the Attacker is to induce a decision error, it is reasonable to assume that $\nu^{m_2} = Q(\tau^m) = Q(\tau_{\overline{\mathcal{M}}}^{m_1})$, that is, the original value of the replaced

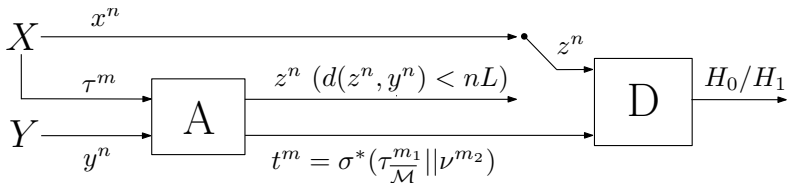


Figure 6.8: General block diagram of the adversarial setup considered in this chapter (targeted corruption case). In the $DT^r_{c\text{-}tr}$ game, addressed in Section 6.5, given the original training sequence τ^m, the adversary has the possibility of choosing which samples to replace with fake ones (i.e., A chooses set $\mathcal{M}$). Clearly, we have that $\sigma^*(\tau^{m_1}_{\overline{\mathcal{M}}}||\tau^{m_2}_{\mathcal{M}}) = \tau^m$.

samples does not matter. Therefore, the same analysis performed in Section 6.1, as well as the results we got, remain the same in the case of game with random replacement of the training samples. Such a view of the $DT^a_{c\text{-}tr}$ game opens the way to the definition of a more general adversarial setup, which is studied in Section 6.5.

6.5 Detection game with selective replacement of training samples

In this section we study a variant of the game with corrupted training, in which A observes the training sequence and can replace a certain fraction of samples. To be consistent with the notation introduced so far, we denote with τ^m the sequence drawn from X which A modifies by replacing a portion α of samples. With respect to the previous case, now the adversary can choose which samples to corrupt in the original training sequence, that is, he has, as an additional degree of freedom, the choice of the index set $\mathcal{M}$.

More formally, given an original training sequence τ^m, the training sequence observed by the Defender is $t^m = \sigma^*(\tau^{m_1}_{\overline{\mathcal{M}}}||\nu^{m_2})$, where $\mathcal{M}$ and ν^{m_2} are determined by the Attacker. Figure 6.8 illustrates the adversarial setup considered in this section. Arguably, this scenario is more favorable to the Attacker with respect to the previous one.

6.5.1 Formal definition of the $DT^r_{c\text{-}tr}$game

In the sequel, we formally define the source identification game with replacement of training samples, namely the $DT^r_{c\text{-}tr}$game.

Defender's strategies.

From the point of view of the Defender, the additional difficulty of this setup is that, even if the he knew the m_2 corrupted samples in the training sequence, simply throwing them away does not guarantee that the remaining part of the sequence follows the same statistics of X, since the Attacker may have deliberately changed it by selectively choosing the samples to replace. Similarly as before, 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

$$\mathcal{S}_D = \{\Lambda_r^{n\times m} \subset \mathcal{P}^n \times \mathcal{P}^m : \max_{P_X\in\mathcal{P}} \max_{s\in\mathcal{S}_{A,T}} P_{\text{FP}} \leq 2^{-\lambda n}\}, \tag{6.62}$$

where $\Lambda_r^{n\times m}$ denotes the acceptance region of the test in the $DT^r_{c\text{-}tr}$ setup and $\mathcal{S}_{A,T}$ indicates the set of corruption strategies.

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 $DT^a_{c\text{-}tr}$ case, while the corruption strategy of the training sequence must be redefined. To this purpose, we observe that the corrupted training sequence t^m may be any sequence 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 distributions of the corrupted sequence, searching for a sequence t^m s.t. $d_H(t^m, \tau^m) \leq \alpha m$ is equivalent to search 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 2 in Chapter 3).

Therefore, the set of strategies of the Attacker is defined by $\mathcal{S}_A = \mathcal{S}_{A,T} \times \mathcal{S}_{A,O}$, where

$$\mathcal{S}_{A,T} = \{Q(P_{\tau^m}, P_{y^n}) \in \mathcal{P}^m \text{ such that } d_{L_1}(Q(P_{\tau^m}, P_{y^n}), P_{\tau^m}) \leq 2\alpha\} \tag{6.63}$$

$$\mathcal{S}_{A,O} = \{S^n_{YZ}(P_{y^n}, P_{t^m}) \in \mathcal{A}^n(L, P_{y^n})\}. \tag{6.64}$$

Note that, in this case, the function $Q(\cdot,\cdot)$ gives the whole corrupted training sequence observed by D (not only the fake subpart, as it was in the $DT^a_{c\text{-}tr}$ game); that is, $t^m = Q(\tau^m, P_{y^n})$. Clearly, due to the specific corruption procedure, there will be m_1 samples in t^m which do not change with respect to the original training sequence.

In the following, we will find convenient to express the attacking strategies in $\mathcal{S}_{A,T}$ in an alternative way. Since the Attacker *replaces* the samples of a subpart of the training sequence (i.e., the amount of modification he can introduce in the

samples he corrupts is unconstrained), the corruption strategy is indeed equivalent to first removing a subpart of the training sequence $\tau_{\mathcal{M}}^{m_2}$ and then adding a synthetic subpart ν^{m_2} to the remaining m_1-length sequence. Then, the sequence is reordered according to σ^*. Hence, by focusing on the type of the observed training sequence, we can write:

$$P_{t^m} = P_{\tau^m} - \alpha(Q_R - Q_A). \tag{6.65}$$

where $Q_R(P_{\tau^m}, P_{y^n})$ and $Q_A(P_{\tau^m}, P_{y^n})$ are the types's of the removed and injected subsequences ($Q_R(\cdot,\cdot), Q_A(\cdot,\cdot) \in \mathcal{P}^{m_2}$). 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.

The set of corruption strategies can then be rewritten as

$$\mathcal{S}'_{A,T} = \{(Q_R(P_{\tau^m}, P_{y^n}), Q_A(P_{\tau^m}, P_{y^n})) \in \mathcal{P}^{m_2} \times \mathcal{P}^{m_2} \text{ s. t. } P_{\tau^m} - \alpha(Q_R - Q_A) \in \mathcal{P}^m \text{ and } P_{\tau^m} - \alpha Q_R \in \mathbb{R}_+^{|\mathcal{X}|}\}, \tag{6.66}$$

where $\mathbb{R}_+$ denotes the set of non-negative real numbers. To clarify the meaning of the constraint we put in (6.66), we observe that not all the pairs (Q_R, Q_A) which result in a valid pmf P_{t^m} (where $P_{t^m} = P_{\tau^m} - \alpha(Q_R - Q_A)$) are valid strategies for the removal and addition. In fact, given a difference $(Q_R - Q_A)$ producing a valid pmf in $\mathcal{P}^m$, in order to find the admissible pairs (Q_R, Q_A), we have to impose that after the removal, $P_{\tau^m} - \alpha Q_R > 0$ for all the alphabet symbols.

Choosing a pmf $Q(P_{\tau^m}, P_{y^n})$ in $\mathcal{S}_{A,T}$ is indeed equivalent to choose a pair of pmf's $(Q_R(P_{\tau^m}, P_{y^n}), Q_A(P_{\tau^m}, P_{y^n}))$ in $\mathcal{S}'_{A,T}$ and then consider the pmf $P_{t^m} = P_{\tau^m} - \alpha(Q_R - Q_A)$ (see also the proof of Theorem 12 in Section 6.3.1).

Payoff function.

As usual, the payoff function is defined as

$$u(\Lambda_r^{n\times m}, (Q(\cdot,\cdot), S_{YZ}^n(\cdot,\cdot))) = -P_{\text{FN}}. \tag{6.67}$$

6.5.2 Equilibrium point and payoff at the equilibrium

We need to make explicit the expression for the false positive probability by considering the possible strategies used by the Attacker to corrupt the training sequence. Similarly to the previous case, for a fixed attacking strategy $Q(P_{\tau^m}, P_{y^n})$ the Defender cannot do better than ignore the fake samples, since he has no hint about how these samples have been corrupted by the Attacker. Now, differently from the previous case, because of the memory introduced by the Attacker, a test sequence x^n (drawn from X) and the remaining part of the (original) training sequence $\tau_{\mathcal{M}}^{m_1}$ may possibly belong to different sources. Hence, to state whether X has been drawn by the same source that generated the uncorrupted training, the Defender has to reconsider the part of the training which has been destroyed by the Attacker.

In order to ensure that P_{FP} is always lower than $2^{-\lambda n}$, it is convenient to use the attack formulation given in (6.66). For a given P_X and $(Q_R(\cdot,\cdot), Q_A(\cdot,\cdot))$, P_{FP} is the probability that X generates 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 $\Lambda_r^{n\times m}$. Accordingly, the set of strategies available to D can be rewritten as:

$$\mathcal{S}_D = \Bigg\{ \Lambda_r^{n\times m} : \max_{P_X\in\mathcal{P}} \max_{(Q_R(\cdot,\cdot),Q_A(\cdot,\cdot))\in\mathcal{P}^{m_2}\times\mathcal{P}^{m_2}} \sum_{P_{y^n}\in\mathcal{P}^n} P_Y(\mathcal{T}(P_{y^n}))\cdot \tag{6.68}$$

$$\sum_{(P_{x^n},P_{t^m})\in\bar{\Lambda}_r^{n\times m}} P_X(\mathcal{T}(P_{x^n}))\cdot \sum_{\substack{P_{\tau^m}\in\mathcal{P}^m:\\ P_{\tau^m}-\alpha(Q_R(P_{\tau^m},P_{y^n})-Q_A(P_{\tau^m},P_{y^n}))=P_{t^m}}} P_X(\mathcal{T}(P_{\tau^m})) \le 2^{-\lambda n} \Bigg\}.$$

By following the same steps as in Section 6.2.1, it is easy to show that the asymptotically optimum strategy for the Defender corresponds to the following:

$$\Lambda_r^{n\times m,*} = \left\{ (P_{z^n}, P_{t^m}) : \min_{(Q_R,Q_A)\in\mathcal{P}^{m_2}\times\mathcal{P}^{m_2}} h\left(P_{z^n}, P_{t^m} + \alpha(Q_R - Q_A)\right) \le \lambda - \delta_n \right\}, \tag{6.69}$$

where δ_n tends to 0 as $n \to \infty$ and the minimization is limited to the pairs in $\mathcal{P}^{m_2}\times\mathcal{P}^{m_2}$ such that $P_{t^m} + \alpha(Q_R - Q_A)$ is a valid pmf (nonnegative and lower than, at most equal to, 1 for all the alphabet symbols).

Consequently, the optimum attacking strategy is given by:

$$(Q^*(P_{\tau^m}, P_{y^n}),\ S_{YZ}^{n,*}(P_{y^n}, P_{t^m})) = \arg\min_{\substack{P_{t^m}: d_{L_1}(P_{t^m},P_{\tau^m})\le 2\alpha \\ S_{YZ}^n\in\mathcal{A}^n(L,P_{y^n})}} \min_{Q_R,Q_A} h\left(P_{z^n}, P_{t^m} + \alpha(Q_R - Q_A)\right). \tag{6.70}$$

Then, the following theorem holds.

Theorem 13. *The $DT_{c\text{-}tr}^r$ game is a dominance solvable game, whose only rationalizable equilibrium corresponds to the profile $(\Lambda_r^{n\times m,*}, (Q^*(\cdot,\cdot),\ S_{YZ}^{n,*}(\cdot,\cdot)))$.*

For the case $L = 0$, we get the optimum strategy of corruption of the training, which is

$$Q^*(P_{\tau^m}, P_{y^n}) = \arg \min_{P_{t^m}\ \text{s.t.}\ d_{L_1}(P_{t^m},P_{\tau^m})\le 2\alpha} \min_{Q_R,Q_A} h\left(P_{y^n}, P_{t^m} + \alpha(Q_R - Q_A)\right). \tag{6.71}$$

Since the corruption of the test sequence works the same as in the $DT_{c\text{-}tr}^a$ case, in the sequel, we focus on the case $L = 0$ (corruption of the training only). Let us define

the set of pairs of types for which D will finally accept H_0 as a consequence of the attack:

$$\Gamma^n_{0,r}(\lambda,\alpha) = \{(P_{y^n},P_{\tau^m}) : \exists P_{t^m} \text{ s.t. } d_{L_1}(P_{t^m},P_{\tau^m}) \leq 2\alpha \text{ and } (P_{y^n},P_{t^m}) \in \Lambda_r^{n\times m,*}\}. \tag{6.72}$$

Given the type of the original training sequence, we define

$$\begin{aligned}\Gamma^n_{0,r}(P_{\tau^m},\lambda,\alpha) =& \{P_{y^n} : \exists P_{t^m} \text{ s.t. } d_{L_1}(P_{t^m},P_{\tau^m}) \leq 2\alpha \text{ and } P_{y^n} \in \Lambda_r^{n,*}(P_{t^m})\} \\ =& \{P_{y^n} : \exists Q,Q' \in \mathcal{P}^{m_2}, P_{t^m} \in \mathcal{P}^m \text{ s.t. } d_{L_1}(P_{t^m},P_{\tau^m}) \leq 2\alpha \\ & \text{and } h(P_{x^n},P_{t^m} - \alpha Q' + \alpha Q) \leq \lambda - \delta_n\big\}.\end{aligned} \tag{6.73}$$

The asymptotic counterpart of the above set, for a generic $R \in \mathcal{P}$, takes the following expression:

$$\begin{aligned}\Gamma_{0,r}(R,\lambda,\alpha) = \big\{P : \exists Q,Q',P' \in \mathcal{P} \text{ s.t. } & d_{L_1}(P',R) \leq 2\alpha \\ & \text{and } h_c(P,P' - \alpha C' + \alpha C) \leq \lambda\big\}.\end{aligned} \tag{6.74}$$

The set can be easily generalized to the case $L \neq 0$ as follows

$$\Gamma_r(R,\lambda,\alpha,L) = \{P : \exists V \in \Gamma_{0,r}(R,\lambda,\alpha) \text{s.t. } EMD(P,V) \leq L\}. \tag{6.75}$$

With the above definitions, it is straightforward to extend Theorem 11 to the $DT^r_{c\text{-}tr}$ case, thus proving that the set in (6.75), evaluated in $R = P_X$, corresponds to the indistinguishability region for the $DT^r_{c\text{-}tr}$ game.

6.5.3 Security margin and blinding percentage

As a last contribution, we are interested in studying the ultimate distinguishability of two sources X and Y in the $DT^r_{c\text{-}tr}$ setting and compare it with the corresponding results for the $DT^a_{c\text{-}tr}$ case. To achieve this goal, we consider the indistinguishability region for the game and study the behavior of such a region when λ tends to 0. We have:

$$\Gamma_r(P_X,\alpha,L) = \{P : \exists V \in \Gamma_{0,r}(P_X,\alpha) \text{ s.t. } EMD(P,V) \leq L\}, \tag{6.76}$$

where

$$\begin{aligned}\Gamma_{0,r}(P_X,\alpha) =& \big\{P : \exists Q,Q',P' \in \mathcal{P} \text{ s.t. } d_{L_1}(P',P_X) \leq 2\alpha \text{ and } P = P' + \alpha(Q - Q')\big\} \\ =& \big\{P : \exists P' \in \mathcal{P} \text{ s.t. } d_{L_1}(P',P_X) \leq 2\alpha \text{ and } d_{L_1}(P,P') \leq 2\alpha\big\}.\end{aligned} \tag{6.77}$$

It is easy to prove the following:

Theorem 14. *Set* $\Gamma_{0,r}(P_X, \alpha)$ *can be equivalently rewritten as*

$$\Gamma_{0,r}(P_X, \alpha) = \big\{P : d_{L_1}(P, P_X) \leq 4\alpha\big\}. \tag{6.78}$$

Proof. Let us show that the set (6.77) is contained in the set (6.78). From the triangular inequality we have that, for any $P' \in \mathcal{P}$, $d(P, P_X) \leq d_{L_1}(P, P') + d_{L_1}(P', P_X)$. Then, if P belongs to $\Gamma_{0,r}(P_X, \alpha)$ in (6.77), it also belongs to the set in (6.78). To see that the sets are indeed equivalent, it is sufficient to show that the reverse implication 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 both from P and P_X is less or at most equal to 2α. By letting $P^* = \frac{P+P_X}{2}$, we have

$$\begin{aligned} d_{L_1}(P, P^*) + d_{L_1}(P^*, P_X) &= \sum_i \left| \frac{P(i) - P_X(i)}{2} \right| + \sum_i \left| \frac{P_X(i) - P(i)}{2} \right| \\ &= d_{L_1}(P, P_X). \end{aligned} \tag{6.79}$$

Since P^* has the same L_1 distance from the pmf's P and P_X, we have that $d_{L_1}(P, P^*) = d_{L_1}(P, P_X)/2 \leq 2\alpha$, and the same holds for $d_{L_1}(P^*, P_X)$. This implies that any P inside the set in (6.78) is also within the set in (6.77). Then, the sets in (6.77) and (6.78) are indeed equivalent. □

Upon inspection of equation (6.78), we can deduce that, as expected, the indistinguishability region for $L = 0$ (and hence, also for the case $L \neq 0$) is larger than that of the $DT^a_{c\text{-}tr}$ game, in which case the L_1 distance was constrained to the value $2\alpha/(1-\alpha)$ (see equation (6.54)), thus confirming that the game with selective replacement of the samples is more favourable to the Attacker. In fact, the quantities 4α and $2\alpha/(1-\alpha)$ are, respectively, a linear and a convex function of α: they take the same value in $\alpha = 0$ and $\alpha = 1/2$ while, for any $\alpha \in (0, 1/2)$, $2\alpha/(1-\alpha) < 4\alpha$. A graphical comparison between the indistinguishability regions for the two setups is shown in Figure 6.9. The difference between the regions reduces as α gets close to the critical value $1/2$ and they coincide for $\alpha = 1/2$. In this case, in fact, the Attacker always wins, being able to bring any pmf inside the acceptance region regardless of the game version.

It is worth noting that the relation between $\mathrm{Vol}(\Gamma_0)$ and $\mathrm{Vol}(\Gamma_{0,r})$ depends on α; that is, the gain of choosing the samples to replace with respect to selecting them at random depends on α. For small α ($\alpha \approx 0$) and α close to the critical value $1/2$ we have that $\mathrm{Vol}(\Gamma_0)/\mathrm{Vol}(\Gamma_{0,r}) \approx 1$, set $\Gamma_{0,r}$ is much larger than Γ_0 for intermediate values of α (the maximum difference between the sets being achieved for $\alpha \approx 0.3$). In fact, when α is close to $1/2$, starting from any P_X, A is able to make impossible to distinguish most of the pmf's from P_X even by choosing the samples at

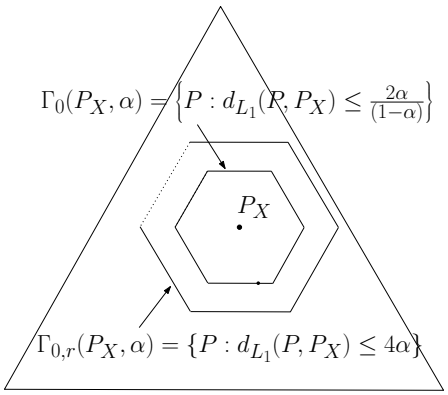


Figure 6.9: Comparison of the indistinguishability regions for the $DT^a_{c\text{-}tr}$ and $DT^r_{c\text{-}tr}$ game with $L=0$ for a generic α ($\alpha<1/2$).

random. Similarly, when $\alpha\approx 0$ very few pmf's can be made indistinguishable from P_X by corrupting the training samples and choosing the samples does not give any significant advantage. Arguably, intermediate values α are those for which the choice of the samples gives the maximum gain with respect to a random choice.

Given two sources X and Y, the blinding value takes the expression:

$$\alpha_b=\frac{d_{L_1}(P_Y,P_X)}{4}. \tag{6.80}$$

Since $d_{L_1}(P_Y,P_X)\leq 2$ for any pair (P_Y,P_X), the blinding value in the current setting is always lower than the blinding value for the previous one (the same value is reached when the two sources yield non-zero values over different symbols).

When the Attacker can also corrupt the test sequence, the *ultimate indistinguishability region* of the $DT^r_{c\text{-}tr}$game is the following

$$\Gamma_r(P_X,\alpha)=\Big\{P:\min_{V:EMD(P,V)\leq L} d_{L_1}(V,P_X)\leq 4\alpha\Big\}. \tag{6.81}$$

Starting from (6.81) we can defining the Security Margin in the $DT^r_{c\text{-}tr}$ setup.

Definition 12 (Security Margin in the $DT^r_{c\text{-}tr}$ 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 $DT^r_{c\text{-}tr}$ setup is given by*

$$\mathcal{SM}_\alpha(P_X,P_Y)=L^*_\alpha, \tag{6.82}$$

where L^*_α satisfies

$$\arg\min_{L_\alpha}\min_{V:\mathrm{EMD}(P_Y,V)\leq L_\alpha}|d_{L_1}(V,P_X)-4\alpha| \tag{6.83}$$

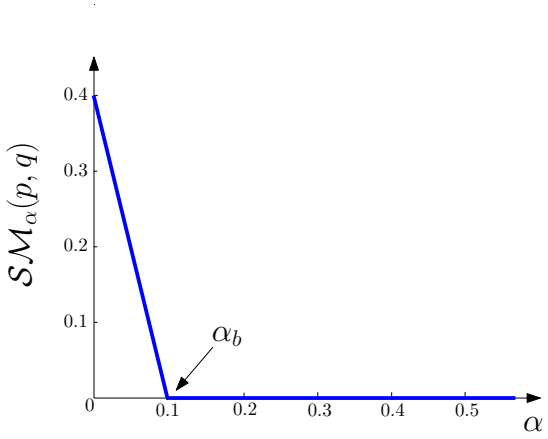


Figure 6.10: Security margin as a function of α for Bernoulli sources with parameters $p = 0.3$ and $q = 0.7$ ($\alpha_b = 0.1$).

if $P_Y \notin \Gamma_{0,r}(P_X, \alpha)$*, and it is 0 otherwise.*

Considering again the case of two Bernoulli sources and by adopting the same notation of Section 6.3.2, we have that $\alpha_b = |p - q|/4$, while the Security Margin takes the expression

$$\mathcal{SM}_\alpha(p,q) = \begin{cases} |q-p| - 2\alpha & \alpha < \alpha_b \\ 0 & \alpha \geq \alpha_b \end{cases} . \tag{6.84}$$

Figure 6.7 plots $\mathcal{SM}_\alpha$ as a function of α when $p = 0.3$ and $q = 0.7$.

Chapter 7

Multiple-Observations (Multivariate) Detection Games

In this chapter, we extend the framework introduced in Chapter 3 to deal with binary detection under multiple observations.

The scenario with multiple observations is relevant in various problems of data fusion and distributed detection [120] and in several applications, including sensor networks [121], cognitive radio networks [122] and multimedia forensics [123]. In all these cases, a Fusion Center (the Defender) has to take a decision about the status of a system (that can be in two states) by relying on a number of observations made available by different sensors (as in [120]) or a number of traces detected by different investigation tools (as in [123]). In many situations, it is possible, in fact probable, that an attacker, or more attackers, corrupts the observations or deliberately provide misleading data to induce a decision error at the fusion center.

In this chapter, we introduce a general information-theoretic framework to analyze the above situations and devise the optimal strategies for both the Defender and the Attacker in a game-theoretic sense, that is by determining the equilibrium point of the game. We will do so for several versions of the game, thus encompassing a large number of scenarios addressing many diverse applications.

The chapter is organized as follows. In Section 7.1, we introduce the general Multiple-Observation Hypothesis Testing setup. In Section 7.2, we adopt the point of view of the Defender and derive the optimum decision strategies in some different data fusion scenario. In Section 7.3, we consider the optimum attacking strategy under some different adversarial conditions. The results are summarized and discussed in Section 7.4.

7.1 Adversarial Multiple-Observation Decision

The Multiple-Observation Binary Decision (MO-DT) problem studied in this chapter is schematized in Figure 7.1. The status of a system is observed by k nodes which gather k observation sequences, $x_1^n, x_2^n \dots x_k^n$, each of which consists of n samples, i.e., $x_l^n = (x_{l,1}, x_{l,2} \dots x_{l,n}), l = 1 \dots k$. The nodes summarize their observa-

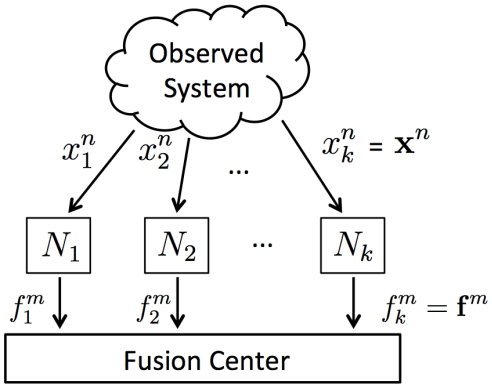


Figure 7.1: The multiple-observation decision scheme.

tions into k feature sequences of length m ($m \leq n$), $f_1^m, f_2^m \dots f_k^m$, with $f_l^m = (f_{l,1}, f_{l,2} \dots f_{l,m}), l = 1 \dots k$. The summaries are sent to a fusion center which has to either accept or reject the hypothesis that the state is in a safe or normal condition (H_0).

This is a very general setup that can be used to model a wide variety of situations. The most obvious application regards distributed hypothesis testing [120]. As an example, the nodes may be part of a sensor network and the observed sequences $x_1^n \dots x_k^n$ may describe the physical state of the system over time, e.g., the temperature, measured at different locations. In more complex situations, the observed sequences may correspond to complex signals like a video or an audio sequence. As to the summaries, in the simplest case they coincide with the observed sequences. More often, they are obtained by extracting a number of features from the observed sequences, or by taking a local decision on the system status. In the latter case, $m = 1$ and $f_l^m = 0$ or 1 depending on the local decision on the validity of hypothesis H_0 taken by node l.

A less obvious instantiation of the setup reported in Figure 7.1 regards the use of data fusion techniques for multimedia forensics. In this case, the observed system is a document, for instance an image or a video, which is analyzed by means of different tools (identified here by $N_1, N_2 \dots N_k$). Each tool analyzes a different aspect of the document. In the case of still images, for instance, the tools may analyze different color bands, or different frequency coefficients, in the case of video, the observables may refer to the audio and video tracks and so on. The tools extract a number of features and send them to a data fusion center, that is in charge of making the final decision on a certain aspect of the analyzed document (e.g., its origin). As in

the distributed hypothesis testing scenario, two extreme cases are obtained when the features correspond to the entire set of observables, and when each tool makes a local decision and fusion is carried out at the decision level.

When MO-DT is framed in an adversarial setting, we must take into account the possibility that an adversary corrupts part of the system so to induce a decision error. We consider two main possibilities. As a first case, we assume that the Attacker corrupts h out of k summaries. This is possible if the Attacker seizes h nodes or if he controls h links between the nodes and the fusion center (see for instance [124]). Two sub-cases are possible depending on whether the Attacker can choose which nodes he is going to attack or not. For the rest, we do not put any further limitation on the Attacker's actions. In the following, we will refer to this setting as *MO-DT with (chosen) corrupted nodes*[1]. In a second scenario, the nodes and the links between the nodes and the fusion center are under the full control of the analyst and hence the Attacker can only modify h out of k observed sequences. This is typically the case in applications wherein the system is analyzed from different points of view by using different analysis tools and the decision on system status is taken by fusing the output of the tools. As an example, we mention data fusion for multimedia forensics analysis, in which an analyst studies various aspects of the document at hand, and takes a decision on the provenance or integrity of the document by fusing the results of the different analyzes. The Attacker, on his side, modifies the document so to hide its true origin or its previous history. In these cases, it makes sense to require that the amount of modification the Attacker can introduce into the document is limited. In the following, we will refer to this scenario as *MO-DT with corrupted observations*. A graphical representation of the two kinds of attacks is given in Figure 7.2.

Several versions of the two general settings described above are obtained depending on the actions allowed to the Attacker and the analyst, their specific goals, the knowledge they have about the system, including its status and its statistical characterization, the knowledge that the Attacker has on the links and nodes that he does not control and so on. In the next sections, we will analyze some of these variants, by framing them into a rigorous game-theoretic setting. As we will see, game-theory provides a natural and flexible way to take into account all the above information and to study the optimal strategies of the two players in terms of game equilibrium and achievable payoff.

[1] In principle we should distinguish between an adversary that *takes full control of the nodes* and an adversary that *controls only the links* between the nodes and the fusion center, since in the former case the Attacker can observe the sequences x_l^n of the corrupted nodes, thus acquiring information about the system status. In this paper we consider an omniscient Attacker, hence making the distinction between the two cases irrelevant.

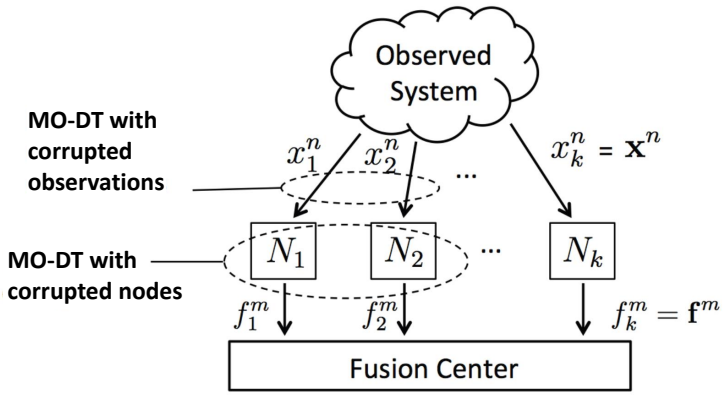


Figure 7.2: Multiple-observation decision under adversarial conditions.

7.1.1 Some notation

Before going on, we introduce further notation necessary to formalize the multiple observations scenario. In our framework the observed system is modeled by a vector of discrete random variables $\mathbf{X} = X_1, X_2 \dots X_k$ taking values in the same alphabet $\mathcal{X}$. Being related to the same system, the random variables are not independent and hence they are described by means of the joint probability mass function (pmf), say $P_{\mathbf{X}}(x_1, x_2 \dots x_k) = P_{\mathbf{X}}(\mathbf{x})$.[2] We indicate by $\mathbf{x}_i = (x_{1,i}, x_{2,i} \dots x_{k,i})$ the vector with the observations of all the nodes at the time instant i, and with $\mathbf{x}^n = \mathbf{x}_1, \mathbf{x}_2 \dots \mathbf{x_n}$ the sequence with all the observed vectors $\mathbf{x}_i$. Similarly to the single observation case, we use the notation $P_{\mathbf{x}^n}$ to indicate the empirical joint pmf (i.e., the type) induced by the sequence $\mathbf{x}^n$ and with $\mathcal{T}(P)$ the type class with all the vector sequences having the empirical pmf equal to P. Finally, we indicate with $\boldsymbol{\mathcal{P}}_n$ the set of all the types for vector sequences of size k and length n[3].

7.1.2 Formalization of the adversarial multiple-observations test

We adopt a Neyman-Pearson perspective according to which D is interested to accept or reject the hypothesis H_0 that the state is in a safe or normal condition characterized by a pmf $P_{\mathbf{X}}$. In doing so D must ensure that the false positive error probability (P_{FP}) of rejecting H_0 when H_0 holds stays below a threshold. On his side, the Attacker

[2] $P_{\mathbf{X}}(\mathbf{x})$ is the joint pmf of the vector $\mathbf{X}$.
[3] We remind that $\mathcal{P}_n$ denotes the types of sequences of length n for the scalar case ($k = 1$).

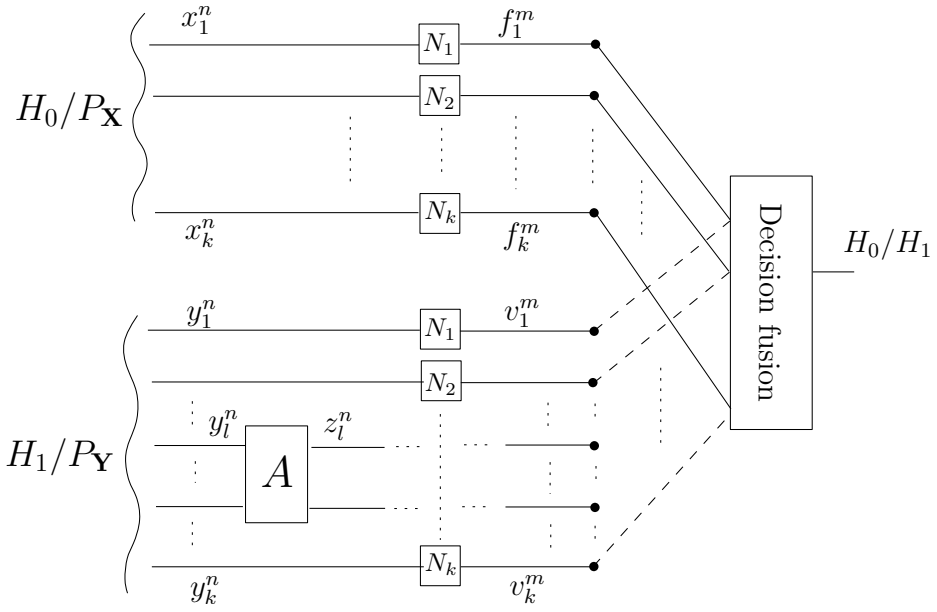


Figure 7.3: The adversarial multiple-observation hypothesis testing setup with corrupted observations considered in this chapter.

aims at inducing a Type II error, i.e., to hide the fact that the system exited its normal status.

To induce a false negative error, he corrupts either the observation sequences (*MO-DT with corrupted observations*), or the summaries sent by the nodes to the fusion center (*MO-DT with corrupted nodes*). In the former case, A must satisfy a distortion constraint specifying to which extent the sequences can be modified. In both cases, A may be allowed to attack all the sequences or only h of them. We denote by $P_{\mathbf{Y}}$ the pmf when H_0 does not hold (H_1 holds). In the following, we indicate with y_l^n the observed sequences when H_1 holds and with v_l^m the corresponding feature sequences. The action of the Attacker corresponds to applying a function $g(\cdot)$ either to y_l^n or v_l^m to produce k attacked sequences z_l^n (w_l^m in the case of corrupted nodes). The hypothesis testing setup for the case of multivariate detection with corrupted observations and corrupted nodes are depicted in Figure 7.3 and 7.4 respectively.

As in Chapter 3, we consider an asymptotic version of the problem (by letting n go to infinity) and require that P_{FP} decays exponentially fast with error exponent at least equal to λ. In addition, we force D to rely on first order statistics only, i.e. to neglect the possible dependence between consecutive observations (we know from Chapter 3 that this assumption is sometimes referred to as limited resources

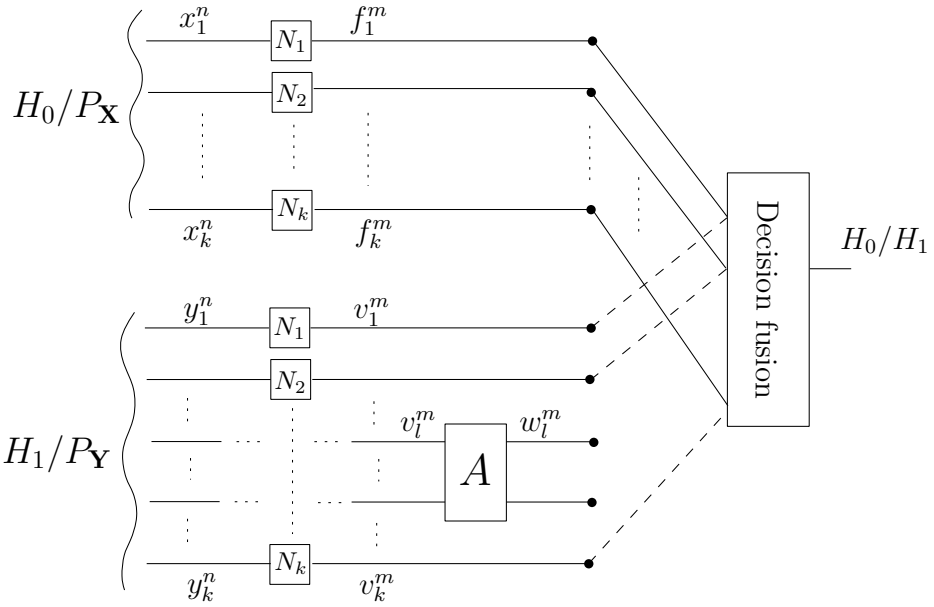


Figure 7.4: The adversarial multiple-observation hypothesis testing setup with corrupted nodes considered in this chapter.

assumption).

7.2 Dominant fusion strategies for the Defender

As anticipated, we use game-theory to give a formal definition of the MO-DT problems outlined in the previous section. In this section we adopt the perspective of the Defender, defining his goals, his possible actions and deriving the optimum fusion strategies under some general assumptions.

7.2.1 MO-DT with full knowledge

As a first scenario, we consider a simplified case in which the nodes take the observed sequences and pass them to the data fusion center as they are, i.e., $f_l^m = x_l^n, \forall l$ under H_0, and $v_l^n = y_l^n, \forall l$ under H_1 ($v_l^n = z_l^n$ for the corrupted sequences). Even if the above condition is rarely verified in practice, this scenario represents a kind of most favorable case for the Defender since he can base his decision on all the available information. In addition, the analysis is rather simple since it is a straightforward extension of the game considered in Chapter 3. In the following, we will refer to this

scenario as the MO-DT game with full knowledge. Let us, then, define the strategies and payoff of the Defender. By adopting the Neyman-Pearson criterion, the possible strategies for D are all the acceptance regions ensuring a given false positive error probability. In formulas:

$$\mathcal{S}_D = \{\Lambda^n \in 2^{\boldsymbol{\mathcal{P}}_n} \text{ s.t. } P_{\text{FP}} \leq 2^{-\lambda n}\}, \tag{7.1}$$

where Λ^n is seen as a union of types (a subset of the power set of $\boldsymbol{\mathcal{P}}_n$) due to the limited resources assumption. Thanks to this assumption, in fact, if an observed vector sequence stays in Λ^n, all the other sequences in the same type class must belong to Λ^n, hence permitting to define Λ^n as a union of type classes and hence a union of types.

As to the payoff, the Defender wishes to minimize the Type II error probability, which, when no Attacker is present under H_1, takes the expression:

$$u_D = -P_{\text{FN}} = -\sum_{\mathbf{y}^n : P_{\mathbf{y}^n} \in \Lambda^n} P_{\mathbf{Y}}(\mathbf{y}^n), \tag{7.2}$$

where with a light abuse of notation $P_{\mathbf{Y}}(\mathbf{y}^n)$ indicates the probability that $\mathbf{Y}$ emits the vector sequence $\mathbf{y}^n$. When an Attacker is present, we should consider $P_{g(\mathbf{y}^n)}$ in place of $P_{\mathbf{y}^n}$ (similarly, for the more general case in which A takes control of only a subset h of the links). It is worth observing that, in the MO-DT setup with full knowledge considered in this section, distinguishing between corruption of the observations or of the summaries is unnecessary (the only difference being the distortion constraint the Attacker is subject to in the former case). Our main result regarding the MO-DT game with perfect knowledge is the following.

Theorem 15. *The strategy*

$$\Lambda^{n,*} = \left\{ P \in \boldsymbol{\mathcal{P}}_n : \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n} \right\} \tag{7.3}$$

is a dominant strategy for D.

Proof. The proof is identical to the proof of Lemma 1 in Chapter 3 and then is omitted. □

In practice the fusion center gathers all the observations and verifies if their joint empirical pmf is in accordance with the expected statistics of $\mathbf{X}$ when H_0 holds.

7.2.2 Marginal-based MO-DT

As a second scenario we consider a situation in which the nodes summarize their observations by passing to the fusion center the first order statistics of the observed

sequences. In other words, we assume that $m = |\mathcal{X}|$ and $f_l^{|\mathcal{X}|} = P_{x_l^n}$, under H_0, and $v_l^{|\mathcal{X}|} = P_{y_l^n}$ (or $v_l^{|\mathcal{X}|} = P_{z_l^n}$ for the case with corrupted observations) under H_1. As an example in which such a scenario applies, we may consider the case of a sensor network in which the nodes observe the system but their link to the fusion center has a very low transmission rate (hypothetically tending to 0). The nodes, then, transmit only the empirical pmf of the observed sequences, i.e., the number of times that each symbol of $\mathcal{X}$ appears in the sequence x_l^n (res. y_l^n, or z_l^n, under H_1 in the case with corruption of the observations). The number of necessary bits to transmit such an information is upper bounded by $|\mathcal{X}| \times \log_2 n$, since each symbol may appear in the sequence at most n times. The rate necessary to code this information is hence $\frac{|\mathcal{X}| \times \log_2 n}{n}$, which tends to 0 when $n \to \infty$. Another possible justification for this scenario is the practical difficulty of getting a reliable estimate of the empirical joint pmf. It makes sense, then, for the Defender to rely only on the empirical marginal pmf's, but still exploit the knowledge he has on the joint pmf of $\mathbf{X}$.

Given that decision fusion is carried out by considering only the empirical marginal distribution of $\mathbf{x}^n$, the Defender is forced to choose a region for H_0 which is a subset of the Cartesian product among the marginal types, i.e. $\mathcal{P}_n^k = \mathcal{P}_n \times \mathcal{P}_n \dots \mathcal{P}_n$. More precisely we have:

$$\mathcal{S}_D = \{\Lambda^n \in 2^{\mathcal{P}_n^k} \text{ s.t. } P_{\text{FP}} \leq 2^{-\lambda n}\}. \tag{7.4}$$

As to the payoff, it is easy to deduce that D still aims at minimizing the same term P_{FN} in (7.2). When an Attacker is present, depending on the adversarial scenario (corrupted observations or nodes), he may corrupt the vector of observations $\mathbf{y}^n$ or directly the node summaries $v_l^{|\mathcal{X}|}$, i.e., the probability distributions $P_{y_l^n}$. However, due to the limit resources assumption, the optimum strategy for D will be the same regardless the attacking behavior, so we do not need to explicitly focus on a specific attacking scenario.

Finding the optimal acceptance region requires that we compute the probability that a source with a joint pmf $P_{\mathbf{X}}$ emits a sequence having certain marginals. This can be done by considering the probability, under $P_{\mathbf{X}}$, of all the joint type classes having the desired marginals. To elaborate, let us indicate by $\mathcal{M}_n(P_1, P_2 \dots P_k)$ the set with all joint types with marginals $P_1, P_2 \dots P_k$, that is:

$$\mathcal{M}_n(P_1 \dots P_k) = \{P \in \boldsymbol{\mathcal{P}}_n : \sum_{-i} P(x_1 \dots x_k) = P_i \ \forall i\}, \tag{7.5}$$

where $\sum_{-i}$ indicates summation over all variables x_j but x_i. Given that the probability of a generic type class Q under $P_{\mathbf{X}}$ decays exponentially fast with exponent $\mathcal{D}(Q||P_{\mathbf{X}})$ and given that the number of types increases polynomially with n, we can proceed as in Lemma 1 in Chapter 3 to prove the following theorem.

Theorem 16. *The strategy*

$$\Lambda^{n,*} = \Bigg\{ (P_1 \dots P_k) \in \mathcal{P}_n^k : \min_{P \in \mathcal{M}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n} \Bigg\} \tag{7.6}$$

is a dominant strategy for D.

Proof. The proof follows by proceeding as in the proof of Lemma 1 in Chapter 3. □

One may wonder how the above result changes when the Defender does not know $P_{\mathbf{X}}$ but only its marginals. This is the case, for instance, of JPEG forensic tools that analyze separately the DCT coefficients of an image without considering the dependencies between them. In this case it makes sense to adopt a worse case perspective and require that $P_{\text{FP}} \leq 2^{-\lambda n}$ for all joint pmf's with assigned marginals. The dominant strategy then includes a double minimization as follows:

$$\Lambda^{n,*} = \Bigg\{ (P_1 \dots P_k) \in \mathcal{P}_n^k : \min_{P_{\mathbf{X}} \in \mathcal{M}(P_{X_1} \dots P_{X_k})} \min_{P \in \mathcal{M}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}}) < \lambda - |\mathcal{X}|^k \frac{\log(n+1)}{n} \Bigg\}. \tag{7.7}$$

Therefore, given the set of source marginals $(P_{X_1} \dots P_{X_k})$, for any set of observed marginals $(P_1 \dots P_k)$, the Defender considers the closest pair of joint pmfs (in divergence terms) in order to decide if accepting or not H_0.

7.2.3 MO-DT based on local decisions

The last scenario we are going to consider assumes that the nodes can send to the fusion center only one bit of information: formally, $m = 1$ and $f_l^1 \in \{0, 1\}$ under H_0 and $v_l^1 \in \{0, 1\}$ under H_1. This is a common situation, occurring, for instance but not only, when the nodes make their own decision about the state of the system and data fusion is carried out at the decision level. This scenario also models a multimedia forensic analysis in which the analyst applies several tools each of which provides a binary output regarding the origin or the authenticity of the analyzed document. It is the task of the fusion center to make a final decision by considering the output of all the tools. In principle we would like to derive the optimal decision strategies at the nodes and the optimal fusion strategy. This is a complex task, so we make the simplifying assumption that D adopts an AND fusion strategy, that is H_0 is accepted only if all the nodes accept it. Assuming an AND-based decision rule is equivalent to imposing that the the overall acceptance region is the Cartesian

product of the acceptance regions adopted by the nodes, i.e., $\Lambda^n = \Lambda_1^n \times \Lambda_2^n \dots \Lambda_k^n$. As in the previous sections, we assume that the nodes can rely only on the first order statistics of the observed sequences.

According to the above scenario, the space of strategies of the Defender consists of all k-uple of local acceptance regions, that is:

$$\mathcal{S}_D = \{(\Lambda_1^n \dots \Lambda_k^n) : \Lambda_i^n \in 2^{\mathcal{P}_n} \text{ and } P_{\text{FP}} \leq 2^{-\lambda n}\}. \tag{7.8}$$

The payoff function is again the false negative error probability. We now prove the following theorem.

Theorem 17. *The strategy*

$$\Lambda_i^{n,*} = \left\{ P_i \in \mathcal{P}_n : \mathcal{D}(P_i||P_{X_i}) < \lambda - |\mathcal{X}|\frac{\log(n+1)}{n} \right\} \ \forall i \tag{7.9}$$

is a dominant strategy for D.

Proof. The proof consists of two steps. First, we prove that the acceptance region $\Lambda^{n,*}$ resulting from the local decision rules defined in (7.9) is an asymptotically admissible choice for D (i.e. it satisfies the constraint on Type I error probability). Then we show that, under the assumption that D adopts an AND fusion rule, the local acceptance regions in (7.9) minimize the overall Type II error probability. Let $\bar{\Lambda}_i^{*,n}$ be the rejection region of H_0 at node i. We have:

$$\begin{aligned} P_{\text{FP}} &= P_{\mathbf{X}}(\mathbf{x}^n \in \bar{\Lambda}^{n,*}) \\ &= P_{\mathbf{X}}(x_1^n \in \bar{\Lambda}_1^{n,*} \text{ OR } x_2^n \in \bar{\Lambda}_2^{n,*} \text{ OR } \dots \text{ OR } x_k^n \in \bar{\Lambda}_k^{n,*}) \\ &\leq \sum_{i=1}^{k} P_{X_i}(x_i^n \in \bar{\Lambda}_i^{n,*}). \end{aligned} \tag{7.10}$$

Due to the first-order assumption, the acceptance region at each node is a union of type classes (or equivalently a union of types with denominator n), hence we can write:

$$\begin{aligned} P_{\text{FP}} &\leq \sum_{i=1}^{k} \sum_{P \in \bar{\Lambda}_i^{n,*}} P_{X_i}(\mathcal{T}(P)) \\ &\overset{a}{\leq} \sum_{i=1}^{k} (n+1)^{|\mathcal{X}|} \max_{P \in \bar{\Lambda}_i^{n,*}} P_{X_i}(\mathcal{T}(P)) \\ &\overset{b}{\leq} \sum_{i=1}^{k} (n+1)^{|\mathcal{X}|} 2^{-n \min_{P \in \Lambda_i^{\bar{n},*}} \mathcal{D}(P||P_{X_i})} \\ &\overset{c}{\leq} k(n+1)^{|\mathcal{X}|} 2^{-n\left(\lambda - |\mathcal{X}|\frac{\log(n+1)}{n}\right)} \end{aligned} \tag{7.11}$$

where a and b derive from known upper bound on the number of types with denominator n and on the probability of a type class under a probability measure P_{X_i} [90], and c is a consequence of (7.9). We have thus shown that $P_{\text{FP}} \leq 2^{-n(\lambda-\delta_n)}$ with $\delta_n \to 0$ for $n \to \infty$, and hence $\Lambda^{n,*}$ asymptotically satisfies the constraint on P_{FP}.

We now pass to the second part of the proof to show that the strategy in (7.9) is indeed optimal. Let Λ^n be an AND-based acceptance region resulting from any other set of local regions Λ_i^n satisfying the constraint on false positive error probability. Finally, let $\mathbf{x}^{n,*}$ belong to $\bar{\Lambda}^n$. This means that $x_i^{n,*} \in \bar{\Lambda}_i^n$ for at least one i, say j. We have:

$$\begin{aligned} 2^{-n\lambda} &\geq P_{\mathbf{X}}(x_i^n \in \bar{\Lambda}_i^n, \text{ for some } i) \\ &\overset{a}{\geq} P_{X_j}(x_j^n \in \bar{\Lambda}_j^n) \\ &= \sum_{P \in \bar{\Lambda}_j^n} P_{X_j}(\mathcal{T}(P)) \\ &\overset{b}{\geq} P_{X_j}(\mathcal{T}(P_{x_j^{n,*}})) \overset{c}{\geq} \frac{1}{(n+1)^{|\mathcal{X}|}} 2^{-n\mathcal{D}(P_{x_j^{n,*}} || P_{X_j})}, \end{aligned} \tag{7.12}$$

where a is obtained by observing that the probability of a union of events is always larger than the probability of one such event, b holds since we have assumed that $\bar{\Lambda}_j^n$ contains at least $x_j^{n,*}$ (and the corresponding type class), and c derives from the usual lower bound on the probability of a type class [90]. By considering the first and the last term in (7.12), we see that $\mathbf{x}^{n,*} \in \bar{\Lambda}^{n,*}$ and hence $\Lambda^{n,*} \subseteq \Lambda^n$. This shows that any other acceptance region Λ^n satisfying the false positive constraint results in a higher false negative probability, thus proving the optimality of $\Lambda^{n,*}$. □

In practice, according to Theorem 17, H_0 is accepted only if the empirical marginals of the sequences observed by the nodes are in accordance with the system model under H_0. Moreover, somewhat expectedly, D does not exploit the knowledge of the joint pmf $P_{\mathbf{X}}$, the optimum decision rule depending only on P_{X_i}.

A unifying, and very important, characteristic of all the scenarios considered in this section, is that the requirement that P_{FP} tends to zero exponentially fast with decay exponent λ and the adoption of a decision rule based on first order statistics already define the optimum Defender's strategy regardless of the strategy chosen by Attacker, thus resulting in the existence of a dominant strategy for D. Moreover, the dominant strategy does not depend on $P_{\mathbf{Y}}$, that is the statistical characterization of the system when H_0 does not hold, making such a knowledge un-necessary.

7.3 Optimal Attacker's strategies

Having derived the optimal strategies for the Defender, we now adopt the perspective of the Attacker. As for the *DT* game in Chapter 3, the existence of a dominant strategy for D makes it possible to study the optimal Attacker's strategy by knowing that the acceptance region adopted by D is equal to $\Lambda^{n,*}$. Together with $\Lambda^{n,*}$, A's optimum strategy defines the equilibrium point of the game, which, because of the dominance of D's strategy, is the only rationalizable equilibrium of the game.

7.3.1 Strategy space of the Attacker

As a first step, we must define the space of strategies of A and the information he has access to. As detailed in Section 7.1, A acts only when H_1 holds (H_0 does not hold) with the aim of inducing a Type II error. In order to do so, he corrupts either the observation sequences (MO-DT with corrupted observations), or the summaries sent by the nodes to the fusion center (MO-DT with corrupted nodes), as depicted in Figure 7.3 and 7.4 respectively.

$\mathcal{S}_A$ for MO-DT with corrupted observations

The set of strategies available to A for the MO-DT game with corrupted observations is given by:

$$\mathcal{S}_A = \{g(\cdot) : d(\mathbf{z}^n, \mathbf{y}^n) \leq nL\}, \tag{7.13}$$

where L is the maximum allowed average distortion for the vector of observations at each time instant. Alternatively, we can impose independent constraints on the distortion introduced in each of the observed sequences:

$$\mathcal{S}_A^{'} = \{g(\cdot) : d(z_j^n, y_j^n) \leq nL_j \ \forall j\}, \tag{7.14}$$

where L_l is the maximum allowed average per letter distortion at node l.

Similar definitions hold when A can corrupt up to h sequences.

With the exception of the case of MO-DT with full knowledge, in which the fusion center receives the whole sequence of observed vectors $\mathbf{x}^n$, res. $\mathbf{y}^n$ (or $\mathbf{z}^n$), and then looks at the joint pmf of the observations $\mathbf{x}_i$, res. $\mathbf{y}_i$ (or $\mathbf{z}_i$), in the case of MO-DT based on summaries (marginals or local decisions), the information on the joint relations between the observations is lost in the data received by the fusion center. For these setups, we can (equivalently) adopt the 'transportation' perspective and rephrase the attacking strategy as a *vector of transportation maps* $S^n_{YZ,j}(\cdot,\cdot; y_j^n)$, $j = 1, .., k$ [4]. Accordingly, the set of attacking strategies in (7.15) can be rewritten

[4] We are (reasonably) assuming that the distance d is an additive pair-wise distance, that is $d(\mathbf{z}^n, \mathbf{y}^n) = \sum_j d(z_j^n, y_j^n) = \sum_j \sum_i d(z_{j,i}, y_{j,i})$.

as

$$\mathcal{S}_A = \{(S^n_{YZ,1}(\cdot,\cdot),....,S^n_{YZ,k}(\cdot,\cdot)),\ \text{s.t. } S^n_{YZ,j}(\cdot,\cdot) \in \mathcal{A}(P_{y_l^n}, L_j),\ j=1,...,k \\ \forall (L_1,..L_k) \text{ s.t. } \sum_j L_j = L\}, \tag{7.15}$$

In the case of independent distortion constraints, the set in (7.16) rephrased in terms of transportation maps takes the following expression:

$$\mathcal{S}_A^{'} = \{(S^n_{YZ,1}(\cdot,\cdot),....,S^n_{YZ,k}(\cdot,\cdot)),\ \text{s.t. } S^n_{YZ,j}(\cdot,\cdot) \in \mathcal{A}(P_{y_j^n}, L_j),\ \forall j\}. \tag{7.16}$$

$\mathcal{S}_A$ for MO-DT with corrupted nodes

In the case of corrupted nodes the Attacker has much more freedom, since in this case he can work directly on the feature sequences v_l^m. All the more that, due to the absence of the distortion constraint, he can replace the feature sequences of the attacked nodes at will. The only applicable constraint is that he can substitute up to h sequences. In the case of chosen corrupted nodes, the space of strategies includes also the choice of the to-be attacked nodes.

Having defined $\mathcal{S}_A$, we must specify the information available to A. To do so, we adopt a worst case assumption and consider an omniscient Attacker, who knows the system status (this is implicit in the Neyman-Pearson setup) and can observe all observation and feature sequences, even those that he is not allowed to modify.

As to the payoff, arguably, the Attacker's goal is to maximize the Type II error probability, that is $u_A = -u_D = P_{\text{FN}}$, thus leading to zero-sum games.

7.3.2 Optimum attack for MO-DT with full knowledge

Let us consider the case of corrupted observations first. Given the optimal Defender's strategy in (7.3), it is easy to realize that the optimum strategy for A is to modify the observed sequences so that the divergence between their empirical joint pmf and $P_{\mathbf{X}}$ is as small as possible while satisfying the distortion constraint, that is:[5]

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nL} \mathcal{D}(P_{\mathbf{z}^n} || P_{\mathbf{X}}). \tag{7.17}$$

This result is analogous to Theorem 1 in Chapter 3 (see equation (16)), the only difference being that vector sources are involved instead of scalar ones. We point

[5] Only the case of global distortion constraint is considered, the extension to the case of independent constraints being straightforward.

out that, in principle, A could reach the same goal by using a lower $\mathcal{D}$, stopping as soon as the pmf gets inside the acceptance region. Given our definition of the game, however, such a situation would not result in a higher payoff. This is the way to save as much distortion as possible which however, in our case, is unnecessary. A similar result holds when the distortion constraint applies to each observed sequence separately. Note that, even if theoretically simple, solving the minimization in (7.17) may be computationally very expensive, as already pointed out in Chapter 3 for the scalar case.

In the case of MO-DT with corrupted nodes, the situation is by far more favorable to the Attacker, since he has to solve the minimization problem without any constraint. It is obvious, then, that A can pass to the fusion center completely fake sequences for which the divergence between the empirical joint pmf and $P_{\mathbf{X}}$ is arbitrarily small. Such sequences will pass the test in (7.3), thus always resulting in a false negative error.

The situation is different when A can attack only h out of k nodes. Even in the most favorable case of corrupted nodes, A can not control the empirical marginals of the non-attacked nodes and the joint pmf between them. If such marginals, or joint pmf, under H_1 are different from those under H_0, it may still be possible for the Defender to reliably distinguish between the two hypothesis (though with a higher P_{FN}). It is also evident that, in the case of chosen corrupted nodes, A will attack the nodes for which the pmf's of the observations under H_0 and H_1 differ most in terms of divergence.

7.3.3 Optimum attack for Marginal-based MO-DT

Even in this case the optimal attacking strategy follows directly from the knowledge of D's dominant strategy. In fact, for the case of corrupted observations, from equation (7.6), it follows that the optimum attacking strategy is given by

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nL} \min_{P \in \mathcal{M}_n(P_{z_1^n} \dots P_{z_k^n})} \mathcal{D}(P||P_{\mathbf{X}}). \tag{7.18}$$

In this case of MO-DT based on marginals, we can adopt the alternative view of the attacking strategy in terms of vector of transportation maps, and rephrase the optimal attack as

$$(S_{YZ,j}^{n,*}(\cdot,\cdot))_{j=1}^{k} = \arg \min_{(S_{YZ,1}^{n,*}(\cdot,\cdot),\dots,S_{YZ,k}^{n,*}(\cdot,\cdot)) \in \mathcal{S}_A} \min_{P \in \mathcal{M}_n(P_{z_1^n} \dots P_{z_k^n})} \mathcal{D}(P||P_{\mathbf{X}}). \tag{7.19}$$

A similar result holds when equation (7.7) is applied instead of (7.6).

The situation is more favorable when the Attacker can corrupt the output of the nodes, since in this case he has no distortion constraint to fulfill and then he can

choose directly the pmf's $P_1 \dots P_k$ that minimize $\min_{P \in \mathcal{M}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}})$. In fact, by letting $w_i^{|\mathcal{X}|,*} = P_i = P_{X_i}$ for all i, we have a perfect attack, since in this case $\min_{P \in \mathcal{M}_n(P_1 \dots P_k)} \mathcal{D}(P||P_{\mathbf{X}})$ is equal to 0. Of course, this is not possible when the Attacker controls only h nodes, in which case the optimum attack boils down to the following minimization (w.l.o.g. we assume that A attacks the first h nodes):

$$P^* = \arg \min_{P \in \mathcal{M}_n(\dots, P_{y_{h+i}^n} \dots P_{y_k^n})} \mathcal{D}(P||P_{\mathbf{X}}), \tag{7.20}$$

where $\mathcal{M}_n(\dots, P_{y_{h+i}^n} \dots P_{y_k^n})$ denotes the set with all joint pmf's with only the last $n-h$ marginals fixed. Once the minimization is solved, A sets $w_i^{|\mathcal{X}|,*} = P_i^*, \forall i = 1...h$.

Finally, when the Attacker chooses which nodes to attack, a further minimization is required to minimize (7.20) over all possible subsets of attacked nodes.

7.3.4 Optimum attack for MO-DT based on local decisions

Once again, the optimum Attacker's strategy follows directly from the knowledge of the dominant strategy of the Defender. By considering Theorem 17, in fact, it is easy to conclude that the optimum strategy for A in the case of corrupted observations is:

$$g^*(\mathbf{y}^n) = \arg \min_{\mathbf{z}^n : d(\mathbf{z}^n, \mathbf{y}^n) \leq nL} \max_i \mathcal{D}(P_{z_i^n}||P_{X_i}). \tag{7.21}$$

or, equivalently, in terms of transportation maps

$$(S_{YZ,j}^{n,*}(\cdot,\cdot))_{j=1}^k = \arg \min_{(S_{YZ,1}^{n,*}(\cdot,\cdot),\dots,S_{YZ,k}^{n,*}(\cdot,\cdot)) \in \mathcal{S}_A} \max_i \mathcal{D}(P_{z_i^n}||P_{X_i}). \tag{7.22}$$

As before, the derivation of the optimum attack may be computationally expensive due to the presence of the distance constraint. If the squared Euclidean distance is adopted, a kind of *waterfilling* approach [125] can be applied. The Attacker, in fact, can operate as follows: choose i such that $\mathcal{D}(P_{y_i^n}||P_{X_i})$ is maximum, and compute z_i^n such that $\mathcal{D}(P_{z_i^n}||P_{X_i}) = \lambda - |\mathcal{X}| \log(n+1)/n - \varepsilon$ (with ε arbitrarily small), and the squared Euclidean distance between z_i^n and y_i^n is minimum. If the distortion is lower than nL, go on with the next i such that $\mathcal{D}(P_{y_i^n}||P_{X_i})$ is maximum, and iterate the above procedure until all $\mathcal{D}(P_{y_i^n}||P_{X_i})$ are lower than the decision threshold or when the maximum distortion is reached.

A considerably simpler situation is obtained when separate distortion constraints apply to the different sequences. In this case in fact, the Attacker has to solve at most k independent scalar minimizations.

To conclude, we consider the case of corrupted nodes. In this case the optimum attack is trivial, since the Attacker needs only to set the output of all the nodes

under his control to 0, namely $w_i^{1,*} = 0, \forall i = 1 \dots h$. Note however that, if A does not control all the nodes, this may not be enough to make the final decision fail, since the fusion center accepts H_0 only if all the nodes accept it.

In the case of chosen attacked nodes, A will attack the nodes for which the marginals under H_1 differ most (in terms of divergence) from those under H_0.

We point out that this scenario is somewhat different from the usual case of decision fusion in the presence of Byzantines [124]. In that case, in fact, the Byzantine nodes do not have a full knowledge of system status (which they know only through the observation of $\mathbf{x}^n$) and flip the output of the local decisions with a certain probability. In addition they usually act both when H_0 holds and when it doesn't.

7.4 Discussion and conclusions

Having derived the equilibrium point of several versions of the MO-DT game, we now draw some conclusions and summarize the main lessons that we learnt from the analysis developed in this chapter.

Inspired by the theoretical analysis of Chapter 3, we devised a theoretical framework for the problem of multiple observation binary detection in presence of adversaries with the taxonomy of several kinds of scenarios referring to different practical applications. With regard to the specific results we have proven, the most interesting one regards the existence of a dominant strategy for the Defender. Accordingly, the Defender may choose its strategy without caring about the Attacker: for instance, he would get no advantage from the knowledge of the attacked nodes, let alone from any attempt to discover them. This marks an important difference with respect to previous works in the field in which the Defender tries to distinguish between honest and malicious nodes (for some examples of such an approach see [126, 127]). In hindsight, the reason for such behavior, is again (as it was for the case of the *DT* game studied in the previous chapter) the adoption of a Neyman-Pearson setup wherein the Attacker acts only when H_0 does not hold, while the Defender is asked to satisfy a requirement on P_{FP}. Coupled with the adoption of an asymptotic setup, this results in the existence of a dominant strategy for D that does not need to know whether a node (or an observation) is controlled by the adversary or not. It goes without saying that in some applications the assumptions we made may not be reasonable, thus opening the way to different formulations of the MO-DT game.

Having determined the equilibrium point of the various games, the next step would be to evaluate the payoff at the equilibrium so to know who is going to *win* the game. Given the pmf's under H_0 and H_1 (res. $P_{\mathbf{X}}$ and $P_{\mathbf{Y}}$), and a distortion constraint L (in the corrupted observations setup), this corresponds to determine whether the probability of a Type II error ultimately tends to 0 or 1 when $n \to \infty$, in a similar

way to the case of DT game in Chapter 3. Doing so for $\lambda \to 0$ would finally permit us to decide whether the two hypothesis H_0 and H_1 are ultimately distinguishable or not, when the Attacker is allowed to attack h observation sequences (or nodes) with a maximum per letter distortion L.

Chapter 8

Detection Games in a Two-Side Attack Scenario

In the previous chapters we studied many variants of the attack-detection game under the one-side attack scenario. There are many situations in which it is reasonable to assume that the Attacker is active under both hypotheses with the goal of confusing the Defender and inducing a wrong decision, that is, causing both false positive and false negative decision errors. For instance, in applications of camera fingerprint detection, an adversary might be interested to remove the fingerprint from a given image so that the generating camera would not be identified and, at the same time, to modify the specific fingerprint to frame an innocent victim, [98, 128].

In this chapter, we focus on a scenario in which the Attacker acts under both hypotheses, namely the *two-side* attack scenario, and address both the case in which the underlying hypothesis is known to the Attacker and the case in which it is not. We define and solve two versions of the game, corresponding to two different decision setups: in the former, we assume that the Defender bases its decision on an adversary-aware N-P test; in the latter, a Bayesian approach is adopted, where the role of the two error probabilities is symmetrized, and the decision is based on the minimization of a Bayesian risk function.

To be able to study the Defender-Attacker interaction in the two-side attack scenario, we focus on an asymptotic version of the games for which we prove the existence of an attacking strategy which is both *dominant* (i.e., optimal no matter what the defence strategy is) and *universal* (i.e., independent of the underlying sources).

This also marks a significant difference with respect to the previous analyses, where, by focusing on finite (non-asymptotic) setups, the existence of a dominant strategy was proven only with reference to the Defender.

The chapter is subdivided into two main sections: in the first one (Section 8.1) we generalize the analysis developed in Chapter 3 for the one-side attack by considering the possibility for the players to randomize the decision strategies and solving the asymptotic version of the game. Then, in the second section (Section 8.2) we define and study the binary detection game under two-side attack.

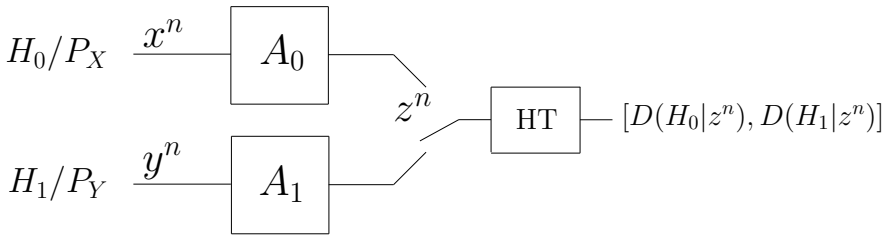


Figure 8.1: Schematic representation of the general adversarial setup with two-sided attack considered in this chapter. In the case of one-sided attack, channel A_0 corresponds to the identity channel, i.e. $A_0 = I$.

8.1 Randomized detection games with one-side attack

Here we generalize the analysis of the detection game with known sources studied in Chapter 3, where the Attacker is active under H_1 only, by considering *randomized detection* and *attack* strategies and focusing on the asymptotic version of the game. Such analysis introduces new interesting results with respect to Chapter 3; moreover, it represents the basis for studying the version of the game with two-side attack, which is the purpose of this chapter. We refer to the generalized version of the detection game with known sources and randomization of the players' strategies simply as detection game with one-side attack, to distinguish it from the case of two-side attack. As to notation, we use the acronym *A-DT* to denote such game, where the letter *A* stands for 'asymmetric' (namely, one-side), as opposed to the symmetric (i.e. two-side) version defined in the sequel[1].

Figure 8.1 illustrates the general two-side attack framework considered in this chapter. The sequences emitted by discrete memoryless sources P_X and P_Y pass through two attack channels defined by conditional probability distributions A_0 and A_1 respectively.

For the one-side scenario, given a sequence z^n observed by the Defender, we have that $z^n = x^n$ under H_0 (no attack occurs), whereas under H_1, z^n is obtained as the output of an attack channel defined by a conditional probability distribution $A_1(z^n|y^n)$. We denote by $Q_i(\cdot)$ the probability distribution of z^n under hypothesis H_i; then, we have $Q_X(z^n) = P_X(z^n)$ and $Q_Y(z^n) = \sum_{y^n} P_Y(y^n)A_1(z^n|y^n)$.

With regard to the Defender, we assume a possibly randomized decision strategy, where $D(H_i|z^n)$ designates the probability of deciding in favor of H_i, $i = 0, 1$, given

[1]We leave implicit in the notation the fact that the sources are fully known, as the case with training data is not considered in this chapter.

the observed sequence z^n [2].

The adoption of randomized strategies calls for a redefinition of the expression of the error probabilities. In particular, the probability of a false positive decision error is given by

$$P_{\text{FP}}(D) = \sum_{z^n} P_X(z^n) D(H_1|z^n), \tag{8.1}$$

whereas the false negative error probability assumes the form:

$$P_{\text{FN}}(D, A_1) = \sum_{y^n, z^n} P_Y(y^n) A_1(z^n|y^n) D(H_0|z^n). \tag{8.2}$$

According to the analysis in Chapter 3, due to the limited resources assumption, the Defender makes a decision based on first order empirical statistics of z^n, which implies that $D(\cdot|z^n)$ depends on z^n only via its type class $\mathcal{T}(P_{z^n})$. Concerning the attack, in order to limit the amount of distortion introduced, as in the previously studied games, we consider a distortion constraint. Specifically, for some chosen permutation-invariant distortion function $d(\cdot,\cdot)$ and maximum per-symbol distortion L, we define the class of admissible channels $\mathcal{C}$ as the class of channels A that assign zero probability to output sequences such that the distance from the input is larger than the prescribed maximum value; i.e., $A(w^n|v^n) = 0\ \forall w^n \in \mathcal{X}^n$ s.t. $d(v^n, w^n) > nL$, $\forall v^n \in \mathcal{X}^n$. Hence, we require that the attack channel A_1 belongs to $\mathcal{C}$.

8.1.1 Definition of the A-DT game

We now define the generalized detection game with *one-side attack*.

Definition 13. *The* A-DT$(\mathcal{S}_D, \mathcal{S}_A, u)$ *game is a zero-sum, strategic game, played by a Defender and an Attacker, defined as follows:*

- *The set of strategies of the Defender is the class $\mathcal{S}_D$ of randomized decision rules $D(\cdot|\cdot)$ that satisfy the following properties:*
  - **(i)** *$D(H_0|z^n) = D(H_0|z^{n\prime})$ whenever $z^{n\prime} \in \mathcal{T}(z^n)$, i.e. $z^{n\prime}$ is a permutation of z^n* [3].
  - **(ii)** *$P_{FP}(D) \leq 2^{-\lambda n}$ for a given prescribed $\lambda > 0$.*
- *The set of strategies for the Attacker is the class $\mathcal{S}_A$ of attack channels A_1 with the property that $d(y^n, z^n) > nL$ implies $A_1(z^n|y^n) = 0$; that is $\mathcal{S}_A \equiv \mathcal{C}$.*

[2] With a slight abuse of notation, we use letter D, already adopted for indicating the Defender, to denote the probabilistic decision; however, the meaning is always recoverable form the context.

[3] Limiting the decision to the first order statistics corresponds to assume that $D(\cdot|z^n)$ is invariant to permutations of z^n.

- *The payoff function:* $u(D, A) = P_{FN}(D, A)$*, where the Attacker's perspective is adopted (the Attacker is in the quest for maximizing* $u(D, A)$ *while the Defender wishes to minimize* $u(D, A)$*).*

Discussion

We point out again that the *A-DT* game is an extension of the DT_{ks} game, since in the *A-DT* both players are allowed to employ randomized strategies, while in the DT_{ks} only deterministic strategies were considered. Specifically, in the definition of the DT_{ks} game, the Defender's strategies were confined to deterministic decision rules whereas for the attack we considered deterministic functions of the to-be-attacked sequence (or, equivalently, deterministic transportation maps). As already pointed in Chapter 3, considering only deterministic strategies for the Attacker is not a limitation. In fact, because of the existence of a dominant strategy for the Defender (which then corresponds to a deterministic test function, see Lemma 3.12), even by allowing the Attacker to play randomized attacking strategies, the optimum strategy for the Attacker would be the same. Then, the randomization of the defence strategy is the only real difference between the DT_{ks} and the *A-DT* setup.

8.1.2 Asymptotic solution of the *A-DT* game

Studying the *A-DT* game is a cumbersome task; so, we focus on the asymptotic behavior of the *A-DT* game, that is, the behavior when the length of the sequence n tends to infinity.

Regarding the notation, for two positive sequences $\{a_n\}$ and $\{b_n\}$, we use the compact notation $a_n \doteq b_n$ to indicate that $\lim_{n\to\infty} 1/n \log(a_n/b_n) = 0$. Similarly, $a_n \dot{\leq} b_n$ designates that $\limsup_{n\to\infty} 1/n \log(a_n/b_n) \leq 0$.

We start by asserting the following lemma:[4]

Lemma 7. *The strategy*

$$D^*(H_1|z^n) \stackrel{\triangle}{=} 2^{-n[\lambda - \mathcal{D}(P_{z^n}\|P_X)]_+}, \tag{8.3}$$

is an asymptotically dominant strategy for the Defender.

Proof. The asymptotic optimality of $D^*(\cdot|z^n)$ follows directly from the false positive

[4] We say that a strategy is asymptotically optimum (or dominant) strategy if the strategy is optimum (dominant) with respect to the exponent of the payoff, that is, the false negative error exponent.

constraint:

$$e^{-\lambda n} \geq \sum_{z^{n\prime}} P_X(z^{n\prime}) D(H_1|z^{n\prime}) \geq |\mathcal{T}(z^n)| \cdot P_X(z^n) D(H_1|z^n)$$
$$\dot{\geq} 2^{-nD(P_{z^n}\|P_X)} D(H_1|z^n), \quad \forall z^n, \tag{8.4}$$

where, in the second inequality, we have exploited the permutation–invariance of $D(H_1|z^n)$ and the memoryless nature of P_X, which implies $P_X(z^n) = P_X(z^{n\prime})$ whenever $z^{n\prime}$ is a permuted version of z^n. It follows that

$$D(H_1|z^n) \dot{\leq} \min\{1, 2^{-n[\lambda - D(P_{z^n}\|P_X)]}\} = D^*(H_1|z^n).$$

By using the method of types [89], it is easy to see that D^* satisfies the false positive constraint within a polynomial factor. Since $D^*(H_1|z^n) \dot{\geq} D(H_1|z^n)$, obviously, $D^*(H_0|z^n) \dot{\leq} D(H_0|z^n)$, and so, $P_{\text{FN}}(D^*, A_1) \dot{\leq} P_{\text{FN}}(D, A_1)$ for every attack channel A_1. □

According to Lemma 7, the best strategy for D is *dominant*, and then it is the optimum strategy regardless of the attacking channel. Furthermore, we observe that the optimum decision function asymptotically tends to a deterministic function which essentially corresponds to the Hoeffding test [92], in line with the results obtained in Chapter 3, where the analysis is confined to deterministic decision rules. As the optimum strategy D^* depends only on P_X, but not on P_Y (this was also the case for the optimum acceptance region in the DT_{ks} setup), it is said to be *semi–universal*.

We now move on to the analysis of the attack. One of the most interesting results of this chapter is stated by the following theorem.

Theorem 18. *For any sequence y^n, let $c_n(y^n)$ denote the reciprocal of the total number of conditional type classes $\mathcal{T}(z^n|y^n)$ that satisfy the constraint $d(y^n, z^n) \leq nL$, namely, admissible conditional type classes*[5]*. The attack channel*

$$A^*(z^n|y^n) = \begin{cases} \frac{c_n(y^n)}{|\mathcal{T}(z^n|y^n)|} & d(y^n, z^n) \leq nL \\ 0 & \textit{elsewhere} \end{cases}, \tag{8.5}$$

is an asymptotically dominant strategy for the Attacker.

Proof. Let us take an arbitrary admissible channel A. For a fixed y^n, let us consider the probability that the channel assign to the sequences z^n in $\mathcal{T}(z^n|y^n)$, that is

[5] From the method of the types, $1 \geq c_n(y^n) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)}$ for any y^n (the total number of type classes is polynomial in n) [90].

$A\{z^n \in \mathcal{T}(z^n|y^n)|y^n\}$. We can define a channel $\bar{A}$ which spreads out uniformly this quantity among all the $z^n \in \mathcal{T}(z^n|y^n)$:

$$\bar{A}(z^n|y^n) = \frac{A\{z^{n\prime} \in \mathcal{T}(z^n|y^n)|y^n\}}{|\mathcal{T}(z^n|y^n)|}. \tag{8.6}$$

Then,

$$\begin{aligned}
P_{\text{FN}}(D,A) &= \sum_{y^n} P_Y(y^n) \sum_{z^n} D(H_0|z^n) A(z^n|y^n) \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{\mathcal{T}(z^n|y^n)} \sum_{z^{n\prime}\in\mathcal{T}(z^n|y^n)} D(H_0|z^{n\prime}) A(z^{n\prime}|y^n) \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{\mathcal{T}(z^n|y^n)} D(H_0|z^n) \sum_{z^{n\prime}\in\mathcal{T}(z^n|y^n)} A(z^{n\prime}|y^n) \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{\mathcal{T}(z^n|y^n)} D(H_0|z^n) A\{z^n \in \mathcal{T}(z^n|y^n)|y^n\} \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{\mathcal{T}(z^n|y^n)} D(H_0|z^n) |\mathcal{T}(z^n|y^n)| \cdot \bar{A}(z^n|y^n) \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{\mathcal{T}(z^n|y^n)} D(H_0|z^n) \sum_{z^{n\prime}\in\mathcal{T}(z^n|y^n)} \bar{A}(z^{n\prime}|y^n) \\
&= \sum_{y^n,z^n} P_Y(y^n) \sum_{z^n} D(H_0|z^n) \bar{A}(z^n|y^n) \\
&= P_{\text{FN}}(D,\bar{A}).
\end{aligned} \tag{8.7}$$

Then, for any probability density $A(z^n|y^n)$, the flattened density $\bar{A}(z^n|y^n)$ achieves the same P_{FN}.

From (8.6) we argue that, for every admissible $\mathcal{T}(z^n|y^n)$

$$\bar{A}(z^n|y^n) \leq \frac{1}{|\mathcal{T}(z^n|y^n)|} = A^*(z^n|y^n)/c_n(y^n), \tag{8.8}$$

which implies that, for every permutation–invariant strategy D, $P_{\text{FN}}(D,A) \leq (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D,A^*)$, or equivalently

$$P_{\text{FN}}(D,A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D,A). \tag{8.9}$$

We conclude that A^* minimizes the error exponent of $P_{\text{FN}}(D,A)$ among all the channels $A \in \mathcal{S}_A$ and for every $D \in \mathcal{S}_D$. □

According to the theorem, given a sequence y^n, in order to generate an attacked sequence z^n which undermines the detection (with the prescribed maximum allowed

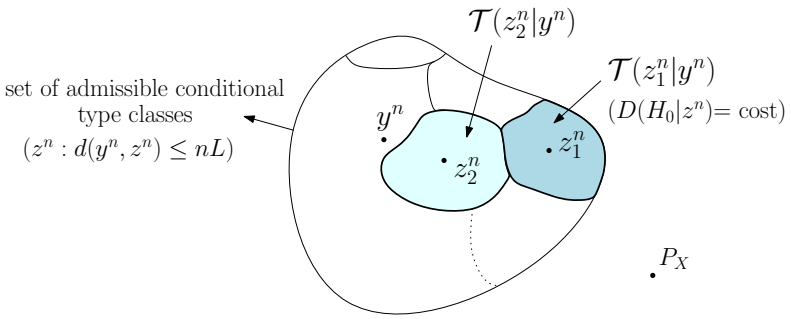


Figure 8.2: Graphical interpretation of the idea behind Theorem 18.

distortion), the best way is to choose an admissible conditional type class according to the uniform distribution (i.e., at random) and then select at random a sequence z^n within this class.

Figure 8.2 should help to get the intuition behind the optimum attack behavior: since the number of conditional type classes is only polynomial in n, the random choice of the conditional type class does not affect the exponent of the error probabilities; besides, since the decision is the same for all sequences within a conditional type class (because of the limited resources assumption), the choice of the sequence inside this set is also immaterial.

It is worth stressing that, according to Theorem 18, strategy A^* is *dominant* for the Attacker, and so the optimum attacking channel does not depend on the decision strategy $D(\cdot|z^n)$. As a further result, Theorem 18 states that the optimum attacking strategy is *universal*, i.e., it depends neither on P_X nor on P_Y. The existence of dominant strategies for both players is a strong result which directly leads to the following result.

Theorem 19. *The profile (D^*, A^*) is an asymptotically dominant equilibrium for the A-DT game.*

We observe that the price to pay for the generalization with respect to the DT_{ks} setup (Theorem 1 and Corollary 1) is that the optimality of the strategies and then the equilibrium point holds only asymptotically.

Finally, we remark that the attack channel in (8.5) is an asymptotically optimum channel even in the setup with deterministic decision considered in Chapter 3. However, in that case, the optimum attack can be found for finite n.

8.2 Detection games with two-side attack

We now consider the detection game when the Attacker is active under both hypotheses. This is the case when the goal of the Attacker is to distort the given sequence, no matter whether it has emerged from P_Y or not, in order to induce a decision error.

In principle, we must distinguish between two cases: in the first one, the Attacker is aware of the underlying hypothesis (*hypothesis-aware* attacker), whereas in the second case, he is not (*hypothesis-unaware* attacker).

In the hypothesis-aware case, the attack strategy is defined by two attack channels: A_0 (carried out when H_0 holds) and A_1 (carried out under H_1), whereas, in the hypothesis-unaware case, the attack strategy consists of only an attack channel A, which is 'blindly' played under both hypothesis.

By focusing for the moment on the hypothesis-aware case, the attack induces the following distributions on the observed sequence z^n: $Q_X(z^n) = \sum_{x^n} P_X(x^n)A_0(z^n|x^n)$ and $Q_Y(z^n) = \sum_{y^n} P_Y(y^n)A_1(z^n|y^n)$. The false positive probability becomes:

$$P_{\text{FP}}(D, A_0) = \sum_{x^n, z^n} P_X(x^n)A_0(z^n|x^n)D(H_1|z^n), \tag{8.10}$$

while for the false negative probability, equation (8.2) continues to hold.

The schematic representation of the adversarial binary detection with two-side (or symmetric) attack is given in Figure 8.1. Clearly, the one-side (or asymmetric) case is a degenerate case of the two-side one (where A_0 is the identity channel).

By reasoning as in the proof of Theorem 18, we now show that the asymptotically optimum attack strategy is independent on the underlying hypothesis. As a consequence, the best attack under the fully active regime is to apply the same A^* regardless of which hypothesis holds. Due to this property, *it becomes immaterial whether the Attacker is aware or unaware of the true hypothesis.*

To be more specific, let u denote a generic payoff function of the form

$$u = \gamma P_{\text{FN}}(D, A_1) + \beta P_{\text{FP}}(D, A_0), \tag{8.11}$$

where β and γ are given positive constants, possibly dependent on n. The following theorem asserts the asymptotic dominance of the channel A^* w.r.t. the payoff function u for every choice of β and γ.

Theorem 20. *Let A^* denote the attack channel in* (8.5)*. Among all pairs of channels in $\mathcal{C}$, the pair (A_0^*, A_1^*) with $A_0^* = A_1^* = A^*$ minimizes the asymptotic exponent of u for any $\gamma, \beta \geq 0$ and any permutation–invariant decision rule $D(\cdot|\cdot)$.*

Proof. Due to the memorylessness of P_Y and the permutation-invariance of $D(H_0|\cdot)$, and by reasoning as we did in Theorem 18, we know that, for every $A_1 \in \mathcal{C}$, we have:

$$P_{\text{FN}}(D, A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FN}}(D, A_1), \tag{8.12}$$

and then A^* minimizes the error exponent of $P_{\text{FN}}(D, A_1)$.

A similar argument can be applied to the FP probability; that is, from the memorylessness of P_0 and the permutation–invariance of $D(H_1|\cdot)$, we have:

$$P_{\text{FP}}(D, A^*) \geq (n+1)^{-|\mathcal{A}|\cdot(|\mathcal{A}|-1)} P_{\text{FP}}(D, A_0), \tag{8.13}$$

for every $A_0 \in \mathcal{C}$. Accordingly, A^* minimizes the asymptotic exponent of $P_{\text{FP}}(D, A_0)$ as well. We then have:

$$\begin{aligned} &\gamma P_{\text{FN}}(D, A_1) + \beta P_{\text{FP}}(D, A_0) \\ &\quad \leq (n+1)^{|\mathcal{A}|\cdot(|\mathcal{A}|-1)} (\gamma P_{\text{FN}}(D, A^*) + \beta P_{\text{FP}}(D, A^*)) \\ &\quad \doteq \gamma P_{\text{FN}}(D, A^*) + \beta P_{\text{FP}}(D, A^*), \end{aligned} \tag{8.14}$$

for every $A_0 \in \mathcal{C}$ and $A_1 \in \mathcal{C}$. Notice that, since the asymptotic equality is defined in logarithmical scale, relation (8.14) holds whichever is the dependence of γ and β on n. Hence, $A_0 = A_1 = A^*$ minimizes the asymptotic exponent of u for any permutation–invariant decision rule $D(H_0|\cdot)$ and for any $\gamma, \beta > 0$. □

We point out that, whenever γ (res. β) is equal to 0, all the attacking strategies A_1 (res. A_0) are equivalent, in the sense that all the pairs (A^*, A_1) for every A_1 (res. (A_0, A^*), for every A_0) lead to the same asymptotic payoff.

From Theorem 20 we deduce that, whenever an adversary aims at maximizing a payoff function of the form (8.11), and as long as the Defender's strategy is confined to the analysis of the first order statistics, A^* is the asymptotically optimal attack under either hypothesis. As a consequence, as anticipated, we do not need to distinguish between hypothesis-aware and unaware attackers. In the sequel, without any loss of generality, we confine our analysis to the case of hypothesis-aware Attacker.

As a final notice, it is worth pointing that the result stated by Theorem 20 generalizes the one stated by Theorem 18 in the previous section, where the optimality of A^* was proved for the case $u = P_{\text{FN}}$ (which correspond to definition 8.11 with $\gamma = 1$ and $\beta = 0$).

Turning the attention to the Defender, the main difficulty of the two-side attacking scenario is that, not only P_{FN}, but also P_{FP} depends on the attack, thus forcing us to reconsider the constraint on P_{FP}.

In the sequel, we consider two different decision setups which lead to different formulations of the detection game with two-sided attack, namely the *S-DT* game.

8.2.1 The $S1$-DT game: Neyman-Pearson approach

As first case, we consider the Neyman-Pearson test. To define the S-DT game in this setup, we assume that the Defender adopts a conservative approach by imposing a false positive constraint pertaining to the worst case attack under H_0. Specifically, we define the game as follows.

Definition 14. *The* S1-DT$(\mathcal{S}_D, \mathcal{S}_A, u)$ *game is a zero-sum, strategic game defined as follows*

- *The set of strategies for the Defender is the the class $\mathcal{S}_D$ of randomized decision rules that satisfy*
  - **(i)** $D(H_0|z^n) = D(H_0|z^{n\prime})$ *whenever* $z^{n\prime} \in \mathcal{T}(z^n)$.
  - **(ii)** $\max_{A_0 \in \mathcal{C}} P_{FP}(D, A_0) \leq 2^{-n\lambda}$ *for a prescribed* $\lambda > 0$.
- *The set of strategies for the Attacker is the class $\mathcal{S}_A$ of the pairs of attack channels (A_0, A_1) such that $A_0, A_1 \in \mathcal{C}$.*
- *The payoff function is $u(D, A) = P_{FN}(D, A_1)$.*

Having already determined the best attacking strategy, we focus on the best Defender's strategy. We can prove the following lemma:

Lemma 8. *The strategy*

$$D^*(H_1|z^n) \stackrel{\triangle}{=} 2^{-n\left[\lambda - \min_{x^n : d(x^n, z^n) \leq nL} \mathcal{D}(P_{x^n}||P_X) - |\mathcal{A}|^2 \frac{\log(n+1)}{n}\right]_+} \tag{8.15}$$

is asymptotically dominant for the Defender.

Proof. Before providing the (very technical) full proof, we first explain the intuition behind the lemma. We know from Lemma 7 that for the case of no attack under H_0, the asymptotically optimal detection rule is based on $\mathcal{D}(P_{x^n} \| P_X)$. In the setup of the A-DT game, where the Attacker is active also under H_0, the Defender is subject to a constraint on the maximum false positive probability over $\mathcal{S}_A$. We know from Theorem 20 that, in the asymptotic exponent sense, this maximum value is achieved when $A_0 = A^*$. From (8.5), we see that A^* assigns a probability which is the reciprocal of a polynomial term at each conditional type class that satisfies the distortion constraint (admissible conditional type class). Then, in order to be compliant with the constraint, for a given sequence z^n, the Defender has to consider the minimum of $\mathcal{D}(P_{x^n} \| P_X)$ over all the type classes $\mathcal{T}(x^n|z^n)$ which satisfy the distortion constraint, or equivalently, all the sequences x^n such that $d(x^n, z^n) \leq nL$.

The proof goes along the following lines: we first show that $P_{\text{FN}}(D^*, A_1) \stackrel{\cdot}{\leq} P_{\text{FN}}(D, A_1)$ for every $D \in \mathcal{S}_D$; then, by proving that $\max_A P_{\text{FP}}(D^*, A)$ fulfills the

false positive constraint, it follows that $D^*(\cdot|z^n)$ is the optimum defence strategy (asymptotically). By exploiting the memorylessness of P_X and the permutation invariance of $D(H_1|z^n)$, we can write:

$$
\begin{aligned}
2^{-\lambda n} \geq & \max_{A} \sum_{x^n, z^n} P_X(x^n) A(z^n|x^n) D(H_1|z^n) \\
\geq & \sum_{z^n} \left(\sum_{x^n} P_X(x^n) A^*(z^n|x^n) \right) D(H_1|z^n) \\
= & \sum_{z^n} \left(\sum_{x^n : d(x^n, z^n) \leq nL} P_X(x^n) \cdot \frac{c_n(x^n)}{|\mathcal{T}(z^n|x^n)|} \right) D(H_1|z^n) \\
\geq & (n+1)^{-|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \sum_{z^n} \left(\sum_{x^n : d(x^n, z^n) \leq nL} \cdot \frac{P_X(x^n)}{|\mathcal{T}(z^n|x^n)|} \right) D(H_1|z^n) \\
\overset{(a)}{\geq} & (n+1)^{-|\mathcal{X}|\cdot(|\mathcal{X}|-1)} |\mathcal{T}(z^{n\prime})| \left(\max_{x^n : d(x^n, z^{n\prime}) \leq nL} |\mathcal{T}(x^n|z^{n\prime})| \cdot \frac{P_X(x^n)}{|\mathcal{T}(z^{n\prime}|x^n)|} \right) D(H_1|z^{n\prime}) \\
= & (n+1)^{-|\mathcal{X}|\cdot(|\mathcal{X}|-1)} D(H_1|z^{n\prime}) \max_{x^n : d(x^n, z^{n\prime}) \leq nL} P_X(x^n) \cdot |\mathcal{T}(x^n)| \\
\geq & D(H_1|z^{n\prime}) \max_{x^n : d(x^n, z^{n\prime}) \leq nL} \frac{1}{(n+1)^{|\mathcal{X}|^2\cdot(|\mathcal{X}|-1)}} 2^{-n\mathcal{D}(P_{x^n}||P_X)} \\
= & \frac{D(H_1|z^{n\prime})}{(n+1)^{|\mathcal{X}|^2\cdot(|\mathcal{X}|-1)}} 2^{-n \min_{x^n : d(x^n, z^{n\prime}) \leq nL} \mathcal{D}(P_{x^n}||P_X)}, \qquad (8.16)
\end{aligned}
$$

where in (a) we exploited the permutation invariance of the distance function d. Since the inequality holds for any $z^{n\prime}$, we argue that

$$
D(H_1|z^n) \leq 2^{-n\left[\lambda - \min_{x^n : d(x^n, z^n) \leq nL} \mathcal{D}(P_{x^n}||P_X)\right]} \qquad (8.17)
$$

and then

$$
D(H_1|z^n) \dot{\leq} \min\left\{1, 2^{-n\left(\lambda - \min_{x^n : d(x^n, z^n) \leq nL} \mathcal{D}(P_{x^n}||P_X)\right)}\right\} = D^*(H_1|z^n). \qquad (8.18)
$$

Consequently, $D^*(H_0|z^n) \dot{\leq} D(H_0|z^n)$ for every z^n, and so, $P_{\text{FN}}(D^*, A_1) \dot{\leq} P_{\text{FN}}(D, A_1)$ for every A_1. For convenience, let us name $k_n(z^n)$ the expression at the exponent in (8.18); so, $D^*(H_1|z^n) = \min\{1, 2^{-n\cdot k_n(z^n)}\}$. Below we show that $D^*(H_1|z^n)$ satisfies the constraint, up to a polynomial term in n, i.e., it satisfies the constraint

asymptotically.

$$
\begin{aligned}
&\max_{A} P_{\mathrm{FP}}(D^*, A) \\
&\quad \le (n+1)^{|\mathcal{X}|\cdot(|\mathcal{X}|-1)} P_{\mathrm{FP}}(D^*, A^*) \\
&\quad = (n+1)^{|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \sum_{x^n, z^n} P_X(x^n) A^*(z^n|x^n) D^*(H_1|z^n) \\
&\quad = (n+1)^{|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \sum_{(x^n,z^n):d(x^n,z^n)\le nL} P_X(x^n) \cdot \frac{c_n(x^n)}{|\mathcal{T}(z^n|x^n)|} \cdot D^*(H_1|z^n) \\
&\quad \le (n+1)^{|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \sum_{(x^n,z^n):d(x^n,z^n)\le nL} \frac{P_X(x^n)}{|\mathcal{T}(z^n|x^n)|} \cdot D^*(H_1|z^n) \\
&\quad \le (n+1)^{2|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \sum_{z^n} \left(\max_{x^n:d(x^n,z^n)\le nL} |\mathcal{T}(x^n|z^n)| \cdot \frac{P_X(x^n)}{|\mathcal{T}(z^n|x^n)|} \right) D^*(H_1|z^n) \\
&\quad = (n+1)^{2|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \left(\sum_{P_{z^n}: k_n(z^n)\ge 0} 2^{-n\cdot k_n(z^n)} \left(\max_{x^n:d(x^n,z^n)\le nL} |\mathcal{T}(x^n)| \cdot P_X(x^n) \right) \right. \\
&\qquad\qquad \left. + \sum_{P_{z^n}: k_n(z^n)<0} \left(\max_{x^n:d(x^n,z^n)\le nL} |\mathcal{T}(x^n)| \cdot P_X(x^n) \right) \right) \\
&\quad \le (n+1)^{2|\mathcal{X}|\cdot(|\mathcal{X}|-1)} \left(\sum_{P_{z^n}: k_n(z^n)\ge 0} 2^{-n\lambda} + \right. \\
&\qquad\qquad \left. + \sum_{P_{z^n}: k_n(z^n)<0} 2^{-n \min_{x^n:d(x^n,z^n)\le nL} \mathcal{D}(P_{x^n}||P_X)} \right) \\
&\quad \le (n+1)^{(|\mathcal{X}|^2+2|\mathcal{X}|)\cdot(|\mathcal{X}|-1)+|\mathcal{X}|} 2^{-n\lambda}. \qquad (8.19)
\end{aligned}
$$

□

Lemma 8 asserts the dominance and the semi-universality of the defence strategy, which depends only on the source P_X.

With regard to the attack, since the payoff of the game is a special case of (8.11) with $\gamma = 1$ and $\beta = 0$, the optimum pair of attacking channels is given by Theorem 20 and is (A^*, A^*). We point out that, as a consequence of Theorem 20, the optimum attacking strategy is *fully universal*: the Attacker does not need to know either sources (P_X and P_Y) or the underlying hypothesis.

We observe that, since the Defender adopted a conservative approach to ensure the constraint on the false positive, the pairs (A_0, A^*), for every $A_0 \in \mathcal{S}_A$, are all equivalent, that is, they lead to the same payoff, and then the Attacker does not even

need to carry out the attack under the null hypothesis. Therefore, if the Attacker is aware of the true hypothesis, then he could play any channel under H_0. [6] In the N-P decision setup, the sole fact that the Attacker is allowed to attack under H_0 forces the Defender to take countermeasures that ultimately make the attack under H_0 useless.

Due to the existence of dominant strategies for both players, we can immediately state the following theorem:

Theorem 21. *Profile* $(D^*, (A^*, A^*))$ *is an asymptotically dominant equilibrium for the* S1-DT *game.*

8.2.2 The $S2$-DT game: Bayesian approach

In this section, we study another version of the *S-DT* game.

Specifically, we assume that the Defender follows a less conservative approach and consider a Bayesian decision setup. This is a quite natural approach to follow for dealing with the symmetric attack scenario[7]. Accordingly, the Defender tries to minimize a particular Bayes risk function. The resulting game is defined as follows:

Definition 15. *The* S2-DT$(\mathcal{S}_D, \mathcal{S}_A, u)$ *game is a zero-sum, strategic game defined by*

- *The set of strategies for the Defender is the class* $\mathcal{S}_D$ *of randomized decision rules that satisfy* $D(H_0|z^n) = D(H_0|z^{n\prime})$ *whenever* $z^{n\prime} \in \mathcal{T}(z^n)$.
- *The set of strategies for the Attacker is the same set as before;*
- *The payoff function:*

$$u = P_{FN}(D, A_1) + 2^{an} P_{FP}(D, A_0), \tag{8.20}$$

for some positive a.

We observe that, in the definition of the payoff, the parameter a controls the tradeoff between the two error exponents; we anticipate that the optimum strategy D will be the one making the difference between the two error exponents exactly equal to a. Notice also that, with definition (8.20), we are implicitly considering for the Defender only the strategies $D(\cdot|z^n)$ such that $P_{\text{FP}}(D, A_0) \dot{\leq} 2^{-an}$. Indeed, any $D(\cdot|z^n)$ which does not satisfy this constraint cannot be the optimum strategy,

[6] We remind that the set of strategies available to the Defender is defined by considering the worst case on the attack channel A_0.

[7] It is also worth observing that, for the games with one-side attack studied in the previous chapters, the Neyman-Pearson setup was a quite natural choice.

yielding a payoff $u > 1$ which can be improved by always deciding in favor of H_0 ($u = 1$).

Let us define:

$$\tilde{\mathcal{D}}(P_{z^n}, P_X) \stackrel{\triangle}{=} \min_{\{P_{x^n|z^n}: E_{x^n y^n}(d(X,Y)) \leq L\}} \mathcal{D}(P_{x^n} \| P_X), \tag{8.21}$$

where $E_{x^n y^n}(\cdot)$ defines the empirical expectation and the minimization is carried out for a given empirical distribution of z^n, P_{z^n}. A similar definition can be given for $\tilde{\mathcal{D}}(P_{z^n}, P_Y)$.

Our solution for the $S2$-DT game is given by the following theorem.

Theorem 22. *Let*

$$D^{\#,1}(H_1|z^n) = U\left(\frac{1}{n}\log\frac{Q_Y(z^n)}{Q_X(z^n)} - a\right), \tag{8.22}$$

where $U(\cdot)$ denotes the Heaviside step function,[8] *and let A^* be defined as usual. Strategy $D^{\#,1}$ is an optimum strategy for the Defender.*

If, in addition, the distortion measure is additive the strategy

$$D^{\#,2}(H_1|z^n) = U\left(\tilde{\mathcal{D}}(P_{z^n}, P_X) - \tilde{\mathcal{D}}(P_{z^n}, P_Y) - a\right) \tag{8.23}$$

is asymptotically optimum for the Defender.

Proof. Since (8.20) is a special case of (8.11) (with $\gamma = 1$ and $\beta = 2^{\alpha n}$), for any defence strategy $D(H_0|\cdot) \in \mathcal{S}_D$, the asymptotically optimum attack channel under both hypotheses is the same and corresponds to the channel A^* defined in (8.5), see Theorem 20. Then, we can determine the best defence strategy by assuming that the Attacker will play (A^*, A^*) and evaluating the best response of the Defender. Given the probability distributions $Q_X(z^n)$ and $Q_Y(z^n)$ induced by A^*, the optimum decision rule is deterministic and is given by the likelihood ratio test (LRT):

$$\frac{1}{n}\log\frac{Q_Y(z^n)}{Q_X(z^n)} \underset{H_0}{\overset{H_1}{\gtrless}} a, \tag{8.24}$$

which proves the optimality of the decision rule in (8.22). In fact, let P_X and P_Y be two any probability distributions of the test and let R be the the Bayes risk with general costs function C_{10} and C_{01}, i.e.,

$$R = C_{10}P_{\text{FP}} + C_{01}P_{\text{FN}}. \tag{8.25}$$

[8] The Heaviside step function or unit step function $U(x)$ is equal to 1 for $x \geq 0$, 0 otherwise.

Given a test sequence z^n, the optimum decision is the one which minimizes (8.25) (optimality criterion), i.e.

$$D^*(\cdot|z^n) = \arg \min_{D(\cdot|z^n)} R. \tag{8.26}$$

We have

$$\begin{aligned} R =& C_{10} \sum_{z^n} P_X(z^n) D(H_1|z^n) + C_{01} \sum_{z^n} P_Y(z^n) D(H_0|z^n), \\ =& C_{01} \sum_{z^n} P_Y(z^n) + \sum_{z^n} D(H_1|z^n) \left(C_{10} P_X(z^n) - C_{01} P_Y(z^n) \right). \end{aligned} \tag{8.27}$$

It is easy to deduce that the decision rule which minimizes R is the one for which $D(H_1|z^n)$ is equal to 1 for all the sequences z^n such that $C_{10}P_X(z^n) < C_{01}P_Y(z^n)$ and 0 for the sequences z^n such that $C_{10}P_X(z^n) \geq C_{01}P_Y(z^n)$. Therefore, the optimum detector for random decision rules is deterministic and works according to the following rule:

$$\frac{P_Z(z^n)}{P_V(z^n)} \underset{H_0}{\overset{H_1}{\gtrless}} \frac{C_{10}}{C_{01}}. \tag{8.28}$$

In our case, $C_{10} = 2^{an}$ and $C_{01} = 1$; then, by considering the log, we get exactly the ratio test in (8.24).

Let us now pass to the second part. To prove the asymptotic optimality of the decision rule in (8.23) for the case of additive distortion measure, we approximate $Q_X(z^n)$ and $Q_Y(z^n)$ using the method of types as follows:

$$\begin{aligned} Q_X(z^n) &= \sum_{x^n} P_X(x^n) A^*(z^n|x^n) \\ &\doteq \sum_{x^n:\ d(x^n,z^n) \leq nL} 2^{-n[H_{x^n}(X) + \mathcal{D}(P_{x^n}\|P_X)]} \cdot 2^{-nH_{x^n z^n}(Z|X)} \\ &\doteq \max_{x^n:\ d(x^n,z^n) \leq nL} 2^{nH_{x^n z^n}(X|Z)} \cdot \left(2^{-n[H_{x^n}(X) + \mathcal{D}(P_{x^n}\|P_X)]} \cdot 2^{-nH_{x^n z^n}(Z|X)} \right) \\ &= \max_{x^n:\ d(x^n,z^n) \leq nL} 2^{-n[H_{z^n}(Z) + \mathcal{D}(P_{x^n}\|P_X)]} \\ &\overset{(a)}{\doteq} 2^{-n\left[H_{z^n}(Z) + \min_{\{P_{x^n|z^n}:\, E_{x^n z^n}(d(X,Z)) \leq L\}} \mathcal{D}(P_X\|P_0)\right]} \\ &= 2^{-n[H_{z^n}(Z) + \tilde{\mathcal{D}}(P_{z^n}, P_X)]}, \end{aligned} \tag{8.29}$$

where in (a) we exploited the additivity of the distortion function d. Similarly,

$$Q_Y(\boldsymbol{y}) \doteq 2^{-n[H_{z^n}(Z) + \tilde{\mathcal{D}}(P_{z^n}, P_Y)]}. \tag{8.30}$$

Thus, we have the following asymptotic approximation to the LRT:

$$\tilde{\mathcal{D}}(P_{z^n}, P_X) - \tilde{\mathcal{D}}(P_{z^n}, P_Y) \underset{H_0}{\overset{H_1}{\gtrless}} a, \tag{8.31}$$

which corresponds to the expression in (8.23), thus concluding the proof of the second part of the theorem. □

The reason why it is meaningful to provide also the asymptotical optimum strategy, is the following: although strategy $D^{\#,1}$ is preferable for the Defender in a game theoretical sense (being optimal for finite n), it requires the non trivial computation of the two probabilities $Q_Y(z^n)$ and $Q_X(z^n)$. Strategy $D^{\#,2}$, instead, is easier to implement because of its single-letter form, and leads to the same payoff asymptotically. Given the above, we can state the following:

Theorem 23. *The profile* $(D^{\#,1}, (A^*, A^*))$ *and* $(D^{\#,2}, (A^*, A^*))$ *are asymptotic rationalizable equilibria for the* S2-DT *game.*

As final remark, we observe that the analysis in this section can be easily generalized to any payoff function defined as in (8.11), i.e., for any $\gamma, \beta \geq 0$. It is straightforward to argue that in the general case the best defence strategy corresponds to (8.22) with $\log \beta / \log \gamma$ in place of a.

Part II

Adversarial Detection Games in practice: an application to Image Forensics

Abstract

The authenticity and the integrity of multimedia documents can be investigated through the tools provided by Multimedia Forensics. However, each improvement in forensic methods is followed by an opposite effort to devise more powerful techniques that leave less and less evidence into the forged documents in the attempt to impair the detection. This leads to the so called 'cat & mouse' loop. In this part of the thesis, we apply our theoretical findings to the multimedia forensic scenario. The theoretical analysis developed in the first part of the thesis, in fact, provides a sound framework to study the interplay between forensic analyst and counterfeiter (adversary). Within this framework, in which the analyst is limited to a first order analysis, we are able to investigate the ultimate *limits of forensic (and counter-forensic) analysis, thus definitely interrupting the loop.*

Chapter 9

Our Take on the Forensic (and Counter-Forensic) Problem

Nowadays, with the widespread diffusion of powerful, user-friendly software, editing digital media (image, video, audio) does not longer require professional skills. Typically, media are edited in order to improve their quality, e.g. by enhancing image contrast, denoising an audio track or re-encoding a video to reduce its size. However, altering a digital media can serve less 'innocent' purposes, such as to remove or implant evidence or to distribute fake content. Since creating a deceiving forgery is now a matter of few clicks, the truthfulness of the message conveyed by media contents must be questioned. Therefore, "seeing is not believing" [129] and a photographic image can not be considered anymore a strong evidence supporting a fact. *Multimedia Forensics* is a relatively new research field whose mission is to tackle with this growing problem. Multimedia Forensics is based on the observation that any processing tends to leave traces (some very thin, others quite evident) that can be exploited to expose the occurrence of manipulations. By looking for such footprints, it is possible to gather information on the life cycle of a digital media and to determine for example, which device acquired it, whether it is a authentic or what kind of processing did it undergo. The final, ambitious aim, is to restore the credibility of digital image.

In the last years, however, several counter-measures have been devised to bypass the forensic analysis by hiding or falsifying traces of illegal processing so that the counterfeited content is deemed authentic. All the solutions in this sense fall into the so called *Counter-forensics* discipline. So far, any attempt to improve the reliability of the forensic analysis has been followed by the dual effort to devise more powerful counter-forensic techniques that leave less and less evidence into the forged documents. Even though this unavoidable and possibly virtuous iterations (the so called 'cat & mouse' game) has led to improved forensic and counter-forensic tools, it is clearly necessary to investigate the ultimate limits of forensic (and counter forensic) analysis, thus definitely breaking the loop.

By adapting the adversarial detection game in order to cope with the multimedia forensic scenario, in this chapter we present the *multimedia forensic game*. Casting

the interaction between the Forensic Analyst and the Adversary (Counterfeiter) into a rigorous theoretical framework, we are able to study their interplay and determine their optimum strategies, thus avoiding the 'cat & mouse' loop.

This chapter is organized as follows: in Section 9.1 we briefly review the basics of Multimedia Forensics; then, in Section 9.2 we introduce the Counter-forensics, and we explain its importance and the challenges it poses. Finally, in Section 9.3, we present our view of the forensic and counter-forensic problem as a competitive game that can then be investigated through the theoretical tools derived in the first part of the thesis (as long as theoretical assumptions are met in practice).

9.1 A brief introduction to Multimedia Forensics

Multimedia Forensics (MMF) emerged as a discipline whose aim is to retrieve information on the history of multimedia documents by searching for the subtle traces left by processing operators. The investigation is performed using a *blind* approach, meaning that no other information is available, apart for the content itself. Unlike *active* techniques like digital watermarking [130], Multimedia Forensics does not assume that the content is generated or controlled by the subject that will have to ensure its authenticity. The idea at the basis of MMF is that almost every step of the digital life cycle typically undergone by a digital content, e.g. acquisition, coding, editing and, more in general, any application of processing operators, leaves a number of traces in the media. By leveraging these traces, several methods have been proposed in the literature to reach some conclusions on the past history of the object under analysis. Different kinds of investigation are possible; among those which received the greatest attention we mention: source identification, whose goal is to determine which kind of device generated the content (i.e., whether an image comes from a camera, a scanner, or it is computer-made) [131] or to identify the specific device used to acquire the content (e.g., [132]); techniques for integrity verification, to understand whether the content has been manipulated in order to alter its semantic content (e.g. image slicing, 'cut & paste', cloning, or 'copy-move' [133]); reverse engineering of processing operators, to detect if a specific chain of processing operators has been applied.

9.2 What is Counter-Forensics?

We now first briefly review the state-of-the-art on Counter-forensics and introduce the 'cat & mouse' paradigm.

9.2.1 A brief overview

The origins of Counter-forensics (CF) trace back to a seminal work by Kirchner and Böhme [134] where the concept of *fighting* against Image Forensics was introduced. Besides, a simple yet important taxonomy was also introduced in [134] (and later on in [44]) that distinguishes between *post-processing* and *integrated* techniques, and between *targeted* and *universal* ones. In a nutshell, a counter-forensic technique which belongs to the post-processing class consists of two steps: first the Attacker performs the tampering, thus obtaining a desired modified content, then he processes the content so to conceal the traces left during the first step. In doing so, the Attacker must satisfy some distortion constraint so that the perceptual quality of the content is preserved. On the contrary, an integrated counter-forensic technique modifies the image so that by construction it does not introduce detectable traces. It is easy to guess that developing integrated methods is much harder in most cases. The second distinction focuses on the target of the counter forensic method: if it aims at removing the trace searched for by a specific detector, then it belongs to the family of targeted attacks. A universal method, instead, attempts to maintain as many statistical properties as possible, so to make the processed image hard to detect also with tools unknown to the adversary.

An example of targeted technique is the one proposed in [135] to hide the traces left in the image histogram by contrast enhancement, so to deceive the detector developed in [136]. Since the method in [135] introduces a local random dithering in the enhancement step, it can be classified as an integrated attack. Nevertheless, the authors also mention the possibility of turning this attack into a post-processing one. Targeted approaches were also proposed to delete the traces left by the acquisition devices [137]. Stamm et al. proposed several post-processing CF techniques to hide both traces of JPEG compression [138, 139], and some kinds of tampering that are revealed thanks to JPEG compression side effects [140]. The basic idea underlying these works is to remove an important trace left by JPEG compression into the image, namely the quantization of DCT coefficients. Since the goal is pursued by introducing additive noise to remove discontinuities in DCT coefficients, these methods can be thought of as post-processing CF attacks. Counter-forensics has also been applied to video: [141] proposed a targeted method that allows to remove/add frames from a MPEG video without introducing statistical artifacts in the prediction error, which are traces exploited in the detector introduced by Wang and Farid to detect video doctoring [142].

So far, the majority of the proposed methods have adopted a targeted approach, whose idea is to exploit the knowledge of the forensic algorithm and of its weaknesses to erase the traces it looks for, while limiting the impact of the modifications on

the perceptual quality of the forgery. However, the imperfections of the counter-forensic techniques introduce new artifacts detectable by developing new detectors or by improving existing ones. For instance, in an attempt to re-establish the validity of forensic analysis, researchers has started building new tools to detect the traces left by anti-forensic algorithms, as in [103], where a so called triangle-test is introduced to prevent the possibility of transplanting the acquisition traces left by a photocamera into an image taken by a different source. In [143], Valenzise et al. presented a detector which is able to detect the anti-forensic technique in [139, 140] by measuring the noisiness after recompression with different quality factors. Another targeted detectors capable to counter the anti-forensic algorithm by Stamm et al. [139] by exploring the features of the high-frequency AC coefficients were proposed in [144]. However, in turn, the refined detectors can be defeated by improved counter-forensic algorithms [144]. This leads to the series of iterations of the forensic and counter-forensic moves researchers usually refer to as 'cat & mouse' game. While this iterative loop will finally lead to powerful forensics and anti-forensics tools, the need to investigate the ultimate limits of forensics (and anti-forensics) techniques clearly exists. In this respect, it would be interesting, instead, to devise universal counter-forensic methods that could guarantee to the Attacker the *undetectability* of the processing he carries out, by means of any forensic tool, at least under some (reasonable) assumptions. To this aim, when designing a counter-forensic method, it is always necessary to simultaneously consider the presence of a Forensic Analyst which is able to react to the Attacker's attempts.

9.2.2 The anti-counter forensic problem as a game theory problem

In the attempt to develop universal counter-forensic tools, research has started moving towards more theoretical approaches; in [44], Böhme and Kirchner cast the forensic problem in a hypothesis testing framework. Several versions of the problem are defined according to the particular hypothesis being tested, including distinction between natural and computer-generated images, tampering detection and source identification. Counter-forensics is then defined as a way to degrade the performance of the hypothesis test envisaged by the analyst. By relying on arguments similar to those used in steganography and steganalysis [9], Böhme and Kirchner argue that a proper way to measure the effectiveness of the attack does not depend on the particular investigation technique adopted by the analyst. Even if Böhme and Kirchner do not explicitly use a game-theoretic formulation, their attempt to decouple the counter-forensic attack from a specific forensic strategy can be seen as a first step towards the definition of the equilibrium point of a general multimedia forensic game.

A work where the game-theoretic framework is explicitly introduced, to evaluate the effectiveness of a given attacking strategy and derive the optimum countermeasures, is [145]. However, in such a work, the Attacker's strategy is fixed and the game-theoretic framework is used only to determine the optimum parameters of the forensic analysis and the attack, thus failing to provide a complete characterization of the game between the Attacker and the Analyst. A first attempt to lay the basis for the construction of a game theoretical framework where casting forensic and anti-forensic technique was made in [146], were a rigorous framework is proposed to model the source identification problem.

9.3 The multimedia forensic game

In order to try to stop the multimedia forensic 'cat & mouse' loop, we exploit the theoretical analysis of the adversarial detection problem developed in the first part of this thesis, and then study the interplay between the Forensic Analyst (FA) and the Adversary (AD), i.e., the counterfeiter, as a zero-sum game: on one hand, the task of the FA is to perform an hypothesis test on a certain document; on the other hand, the AD wants to carry out the attack in such a way to deceive the FA.

With specific reference to the image forensic scenario, the goal of the FA is to tell apart untouched images from those that have undergone some (usually very specific) processing. In a realistic scenario, it is reasonable to assume that the FA has limited resources for performing measurements over the signal. To meet the theoretical assumptions, we focus on the case where the FA considers only the first order statistics of the observed signal, such as the image histogram in pixel domain or the histograms of the DCT coefficients in the frequency domain (the limitation provided by such an assumption in real forensic scenarios is discussed in Section 9.3.1). On the opposite side, the AD will first produce a tampered image having some desired characteristics and then modify the image in such a way to bring it as close as possible to an original, unprocessed image, while respecting some distortion constraints to preserve the visual quality.

Within this framework, the theoretical analysis developed in Part I allows us to derive optimal strategies for both the FA and the AD. Thanks to the adopted game theoretical approach, the counter-forensic strategy derived in this way is 'universal', in that the AD does not need to know anything about the FA detection algorithms (apart from the fact that they are based on first-order statistics), and 'post-processing', since the AD can use the proposed technique as is to hide the traces introduced by any kind of processing tool.

We now discuss how the theoretical analysis can be applied to the multimedia

forensic case, and then exploited to develop a universal attack.

First of all, it is reasonable to assume that, in devising a test to distinguish between untouched and processed images, the FA can only rely on the knowledge of 'examples' coming from both classes. In fact, when a statistical model for the two classes does not exists, as it is often the case in image forensic applications (and more in general in all those applications involving multimedia signals), the best solution for the Forensic Analyst is to adopt a data-driven approach, usually based on machine learning techniques, wherein the characteristics of the image classes are derived from a number of examples (the training set).

Let, then, $\mathcal{C}$ be a class of images (e.g., the class of the never processed images). Given a test image I, the goal of the Forensic Analyst is to accept or reject the hypothesis that I belongs to $\mathcal{C}$. To make his decision, the FA can rely on a set of sample images belonging to $\mathcal{C}$, let us call it $\mathcal{S}$. Moreover, by assuming that the Defender relies only on the first order statistics of I, that is the image histogram, the goal of the AD is to take an image J belonging to another class $\mathcal{C}'$ and modify it in such a way that the Defender classifies it as belonging to $\mathcal{C}$.

Since the FA knows the training sequences but not the 'real' probability distribution for class $\mathcal{C}$, we argue that the FA-AD interplay can be modeled as a detection game with training data (DT_{tr} game), studied in Chapter 4. The equilibrium point for such a game is given by Theorem 3: while the optimum test function for the analyst is the h function, computed between the attacked sequence and the training sequence (see equation (4.17)), the optimum strategy for the AD is to minimize such a function.

In contrast to the scenario where the decision of the Defender is based on the observation of a single training sequence, here the Forensic Analyst relies on a set of samples images to make a decision. Nevertheless, substantially there are no differences: in fact, it is easy to argue that, for an analyst who relies on more than one training sequence (image), the optimum log-likelihood function is the *minimum* of the h function over the entire training set.

Therefore, by taking the role of the Adversary, we can implement the optimum counter-forensic strategy against first-order based forensics, which is the purpose of the next chapter. It is worth stressing that the optimality holds in a game theoretical sense, that is, with respect to the best possible first order detector. Clearly, an AD with deeper knowledge of the forensic tools used by the analyst could resort to more powerful attacks. However, such attacks would be 'targeted' to the specific features the forensic detector looks for, thus risking to fall again into the 'cat & mouse' loop. Indeed, it is not surprising that the gain in generality and applicability of universal CF tools comes at the price of reduced performance with respect to tailored schemes.

9.3.1 Impact of theoretical assumptions on practical setups

The two main assumptions behind the theoretical analysis developed in the first part of the thesis are that the sources are memoryless, and that the Defender relies only on first order statistics to make his decision. We fall into this category whenever the Defender relies on statistics that can be derived from the analysis of the relative occurrences of the symbols within the observed sequence, including higher order moments like, for instance, the empirical skewness and the kurtosis of the sequence. On the other hand, the joint statistics among samples, like transition probabilities and co-occurrence matrices [147] are not included in this category.

From a practical point of view, the main problem with the memoryless assumption is that it may not be met in real-world applications. Real signals, such as images, for instance, cannot be assimilated to memoryless sources and consequently, the Defender could decide to go beyond first order statistics to make his decision. In some cases, the memoryless assumption can be justified because the Defender operates in a transformed domain, e.g., the DCT domain, or in a random projections domain [148]. However, since even in the case of sources with memory, by the law of large numbers, the sources will end up generating sequences with a type arbitrarily close to the marginal pmf, we conjecture that our analysis remains valid for sources with memory, as long as the FA decides to rely *only* on the empirical marginal distributions for his analysis. Clearly, such an assumption can be very restrictive (and then make less sense) when dealing with the memory case. In any event, the use of first order detectors is quite common in practical applications even when dealing with correlated sources (often due to the complexity of higher-order statistical analysis...). In the case of Image Forensics, for instance, several techniques rely only on the analysis of the image histogram or a subset of features derived from it. As an example, this is the case of the detection of contrast enhancement operations [136, 149] or the detection of cut and paste based on image noisiness [150, 151]. Double JPEG forensics is another example where detection is often accomplished by looking only at the histograms of block-DCT coefficients [152], or first digits occurrences in block-DCT coefficients [153], which again is an information that depends only on DCT histograms. A similar analysis based on first digits occurrences has also been adopted for the detection of single [154] and multiple [155] JPEG compression (we refer to Section 11.1 for a state-of-the-art on related JPEG forensic methods).

Another assumption underlying the theoretical analysis which may not be valid in practice is that X and Y are stationary sources. Time varying sources are encountered in many practical applications. In PRNU-based camera identification [98], for instance, images produced by a specific camera are detected due to the presence of a distinctive time varying signature, the Photo Response Non-Uniformity noise, intro-

duced by the camera during image acquisition. Other examples can be drawn from biometric recognition, where the biometric templates used for identity verification can not be assimilated to stationary signals [11] and steganalysis, where the residual noise of the cover image is often modelled as a sequence of independent Gaussian variable with different variance [156, 36, 35, 37]. In principle, the analysis of these situations requires a different theoretical analysis with respect to the one derived in Part I of the thesis. Yet, even when dealing with time varying signals, the use of first order statistics obtained by a global analysis of the analysed signal is sometime common practice. Related to this, it is interesting observing that, first order statistics are sometimes used instead of more powerful joint statistics in biometrics, e.g., in [157], where the adoption of the arbitrarily varying sources (AVS) model [158] permits to account for a (slightly) time variant behavior of the sources and still resort to a memoryless formulation.

Chapter 10

Universal Attacks in the Pixel Domain

In this chapter, we exploit the results of the theoretical analysis of the decision game with training sequence studied in Chapter 4, to derive a universal image counter-forensic scheme that is able to counter *any* detector based on the analysis of the image histogram. Being *universal*, the scheme does not require the knowledge of the specific detection algorithms used by the Forensic Analyst, and can be used to conceal the traces left in the histogram of the image by *any processing tool*.

From the theoretical analysis we know that, for the general case of multi-valued sources, the analytic computation of the optimum adversarial strategy is not possible and then we need to resort to numerical analysis. Then, before delving into the analysis of the counter-forensic algorithm, in Section 10.1 we describe the *constrained optimization problem* that the Attacker has to solve to determine its optimum strategy for both the DS_{ks} and DT_{tr} games and discuss the resolution methodologies. An extensive description of our universal counter-forensic algorithm in all its steps is provided in Section 10.2. Finally, in Section 10.3 the validity of the scheme is assessed through experimental validation by focusing on a specific forensic applications, namely the contrast-enhancement detection.

10.1 Numerical evaluation of the optimum attack for the DT_{ks} (and DT_{tr}) game

Here we describe the numerical analysis for deriving the optimum adversarial strategy for the detection game with known sources (DT_{ks} game) studied in Chapter 3. The same arguments holds for the case of detection game with training data (DT_{tr} game)[1].

While the formula defining the optimum acceptance region in (3.23) 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 (3.24)[2]. Such a minimization

[1] We remind that the main difference between the DT_{tr} and DT_{ks} game relies on the adoption of the h function in place of the $\mathcal{D}$ function as log-likelihood function for the test.

[2] The formulation in terms of transportation map is more convenient for the numerical analysis than the one in (3.17).

resembles some instances of the *optimal transport problem* [93, 105], however here we are interested in minimizing the divergence between a source pmf and a target one, subject to a distortion constraint, whereas, classically, Optimal Transport faces with the somewhat-dual problem of minimizing the distortion needed to make the two pmf's equal. This is exactly the case with the analysis of the limiting performance developed in Chapter 5, where the minimum transportation cost, i.e., the Earth Mover Distance (EMD) to move a pmf into an other is evaluated, as a measure of the Security Margin between two sources.

Let y^n be the to-be-attacked sequence. We introduce the *displacement map* $N = \{n(i,j)\}_{i\in\mathcal{X},j\in\mathcal{X}}$, whose (i,j)-th element tells how many elements should be moved from the i-th to the j-th bin. Accordingly, $N = n \cdot S^n_{YZ}$, where S^n_{YZ} is the transportation map defined in (3.4). By expressing the divergence term $\mathcal{D}(P_{z^n}||P_X)$ as a function of the displacement map N, the minimization problem in (3.24) can be formulated in terms of the $n(i,j)$ variables as follows:

$$\min_{n(i,j)} \sum_{j=1}^{|\mathcal{X}|} \frac{\sum_k n(k,j)}{n} \cdot \log\left(\frac{\sum_k n(k,j)}{nP_X(j)}\right) \tag{10.1}$$

subject to the constraints (i.e., the admissibility set $\mathcal{A}(L, P_{y^n})$):

$$\begin{cases} \sum_j n(i,j) = nP_{y^n}(i) \quad \forall i \\ \sum_{i,j} n(i,j)d(i,j) \le nL \\ n(i,j) \ge 0 \\ n(i,j) \in \mathbb{N}, \end{cases} \tag{10.2}$$

where we considered a generic additive distortion function with per-letter distortion $d(i,j)$.

The optimization problem in (10.1)–(10.2) belongs to the MINLP (Mixed integer nonlinear problems) class [159]. Besides, the objective function is convex in the optimization variables $n(i,j)$, and then in N (see Appendix D for the proof). Since the constraint functions defining the feasible set are also convex in the $n(i,j)$ variables and upper bounded, the problem is actually a *convex* MINLP, for which a global optimum solution exists. To be more specific, the feasible region of the problem is described by linear functions, that is, the constraint matrix is linear, then the set of the possible solutions, namely, the admissibility set, is a limited polyhedron, i.e., a polytope.

For convex MINLPs there are several efficient solvers yielding the optimum solution [160]. Among the most common algorithms for solving convex MINLPs, a remarkable candidate is the branch-and-bound method, according to which we solve the NLP (nonlinear programming) relaxation of the problem obtained by removing

the constraint that the $n(i,j)$ variables must assume integer values [161]. Given the convexity of the objective function, the relaxed problem can be solved efficiently by resorting to steepest gradient method [162]. In our applications, we used the BONMIN[3] solver in the BB mode [160] which implements the NLP-based branch and bound algorithm. By default, it resorts to the software package IPOPT [163] to solve the NLP relaxation.

As to the computational complexity, we notice that the number of optimization variables is quadratic in $|\mathcal{X}|$. It is proper to remark that, in practical applications, such number can be quite large: e.g., for imaging applications in the pixel domain, we have $|\mathcal{X}| = 256$ (i.e., the possible values assumed by a pixel in the image), whereas for applications in the frequency domain, $\mathcal{X}$ is given by all the possible values of the DCT coefficients, which, especially at low frequencies, fall in a very large range.

By considering the L^∞ distortion measure in the constraint matrix (10.2) in place of L_p^p, we can drastically reduce the number of optimization variables and make the optimization viable also for very large values of $|\mathcal{X}|$ (see Section 10.2.2).

As a final observation, we point out that, in principle, it makes sense to consider only solutions for which one between $n(i,j)$ and $n(j,i)$ is equal to 0. However, it is not necessary to explicitly express this constraint, since the solutions for which this condition does not hold can be easily pruned after the optimization problem is solved.[4]

Before concluding this section, we notice that in the case of the DT_{tr} game the optimization problem the Attacker must solve is the same 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 can be derived by the same arguments used for proving the convexity of $\mathcal{D}$.

10.2 A universal counter forensic algorithm

By expliciting the arguments in Section 9.3, we derive a universal (and post processing) counter-forensic algorithm that works against *any* forensic detector based on the analysis of the image histogram, whatever is the trace it looks for.

As regards the notation, from now on, all images will be denoted with the underline notation, e.g., $\underline{x}$, where $\underline{x}(i) \in \mathcal{I}$ is the value of the i-th pixel of the image. Accordingly, $\underline{x}$ is a vector of size $r \times c$, where r and c denote, respectively, the number

[3] Basic Open-source Nonlinear Mixed Integer programming.

[4] A transportation map with $n(i,j) = a$ and $n(j,i) = b$, where $a \geq b$ (w.l.o.g.), corresponds to the same solution of the map where $n(i,j) = a - b$ and $n(j,i) = 0$, which has a lower contribution to the overall distortion.

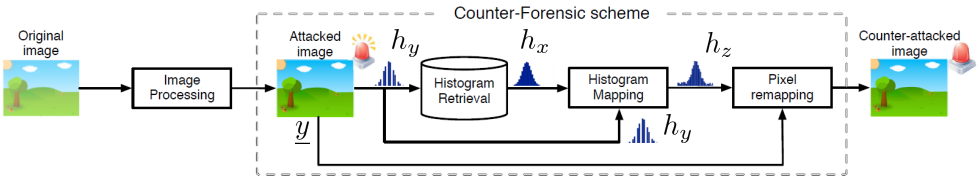


Figure 10.1: A schematic representation of the proposed universal counter forensic approach. Notice that, being a post-processing approach, we are not interested about the specific processing carried out by AD.

of rows and columns of the image[5]. We denote with $\mathcal{I}$ the set of possible values for the pixel, i.e., $\mathcal{I} = \{0, 1, 2, ..., 255\}$. Finally, we use h_x to indicate the histogram of $\underline{x}$[6].

A sketch of the proposed counter-forensic scheme is depicted in Figure 10.1. To begin with, let us assume that the AD has already created the processed image $\underline{y}$, and that he has access to a set $\mathcal{S}$ of histograms of non-processed images. Then, the AD proceeds as follows:

1. *Histogram retrieval* (Section 10.2.1): among all histograms in $\mathcal{S}$, find the one that is most similar to h_y, denote it with h_x.

2. *Histogram mapping* (Section 10.2.2): find the best way to modify h_y so to bring it as close as possible to h_x, while satisfying some constraints on the maximum distortion incurred by $\underline{y}$;

3. *Implementation of the mapping* (Section 10.2.3): change pixels in the image according to the histogram mapping, keeping the perceptual distortion as low as possible.

10.2.1 Histogram retrieval phase

The goal of this phase is the following: given a processed image $\underline{y}$ with histogram h_y find the most "similar" histogram h_x among in set $\mathcal{S}$. We use the h function as similarity measure between a processed histogram and a target one. Hence, the search for the target histogram can be carried out by performing the following minimization:

$$\min_{h_x \in \mathcal{S}} h(\nu_x, \nu_y), \tag{10.3}$$

[5]Due to the first order assumption, we do not need to consider the matrix notation for the image

[6]With a slight abuse of notation, letter h, without the pedex, denotes the generalized log-likelyhood function.

where ν_x and ν_y denote the normalized versions of the histograms h_x and h_y, i.e., the empirical pmf of images $\underline{x}$ and $\underline{y}$.

We observe that the histogram resulting from (10.3) could not be the one that minimizes the h function *after* the histogram mapping phase is performed, which is the final goal of the AD. This is in contrast with the fact that the optimum attack has to guarantee the achievement of the minimum possible value for h among all the possible training sequences and mappings. To face this problem, we retrieve the best K matching histograms from the database, and run the histogram mapping on all of them. Among these K candidates, the one resulting in the best mapping (based on the value assumed by the objective function) will actually be used. Strictly speaking, the number of K which ensures to obtain the minimum at the end of the mapping stage is not known a priori and, above all, it usually depends on the shape of the modified histogram h_y. From a theoretical point of view, for ensuring the optimality of the procedure, we should consider all the histograms in the database. Although this is possible, it would be computationally too expensive, so we retain the first K histograms for the mapping phase[7].

Notice that, the use of the h function allows to retrieve, in this phase, a histogram h_x that is near to h_y even from the "shape" point of view. By looking at the behavior of the h function, which is similar to that of the divergence, histograms having many bins with considerably different occurrences lead to large values for both terms in (4.8) and then they will not be chosen among the best K target histograms. We stress that the use of other frequently used histogram distance functions, like the Chi-Square distance, or the use of cross-bin histogram distances, like the Quadratic-Chi [164], designed to improve the search results for retrieval applications, is also possible. However, according to the theoretical results, in the absence of other knowledge, the use of the h function corresponds to the best choice.

As a last consideration about histogram retrieval, we point out two facts. The first is that the search is carried out directly on histograms, and not on images. This considerably reduces the size of the dataset (10.000 histograms can be represented with less than 10MB) and the complexity of the search routine, since only the histogram of the processed image must be computed on-line. The second observation is that the goal of this phase has nothing to do with content-based retrieval: since the FA relies on first order statistics only (and then he does not care of the image content), the AD simply wants to know if an original image exists (no matter what its content is) whose histogram is not far from that of the processed one, but he is not interested in what is actually represented in the image.

[7] In practice, it turned out in our experiments that taking K equal to 10 is usually sufficient.

10.2.2 Histogram mapping phase

Given the processed image $\underline{y}$ and an original histogram h_x belonging to the reference histogram database, the AD aims at creating an attacked image $\underline{z}$ that is similar to $\underline{y}$ but has an histogram which is as close as possible to h_x.

For sake of generality, we assume that the image of the database from which the histogram h_x has been drawn, let us name it $\underline{x}$, has a different number of pixels than that of the processed image $\underline{y}$ (for any image of the database the number of pixels of the image is preserved, by storing the histogram instead of its normalized version). Let m be the number of pixels of image $\underline{x}$, and let n indicate the number of pixels of the processed image $\underline{y}$, which reasonably will be the same of the attacked one $\underline{z}$.

The goal of the AD is to find the displacement matrix $N^* = \{n^*(i,j)\}_{i=0\ldots255, j=0\ldots255}$ that minimizes the h function between the normalized versions of the histograms, namely ν_z and ν_x while satisfying some distortion constraint between $\underline{z}$ and $\underline{y}$. According to (4.8), the h function is defined as

$$\begin{aligned} h(\nu_z, \nu_x) =& \mathcal{D}(\nu_z||\nu_r) + \frac{m}{n}\mathcal{D}(\nu_x||\nu_r) \\ =& \sum_{i=1}^{|\mathcal{I}|} \nu_z(i) \log \frac{\nu_z(i)}{\nu_r(i)} + \frac{m}{n} \sum_{i=1}^{|\mathcal{I}|} \nu_x(i) \log \frac{\nu_x(i)}{\nu_r(i)}, \end{aligned} \tag{10.4}$$

where $\nu_r(i) = \frac{n}{n+m}\nu_z(i) + \frac{m}{n+m}\nu_x(i)\ \forall i$. To simplify the notation, we define $c = \frac{n}{n+m}$ and $d = \frac{m}{n+m}$ where $c + d = 1$, and rewrite explicitly (10.5) as follows

$$\begin{aligned} h(\nu_z, \nu_x) =& \sum_{i=1}^{|\mathcal{I}|} \nu_z(i) \log \frac{\nu_z(i)}{c\nu_z(i) + d\nu_x(i)} \\ &+ \frac{d}{c} \sum_{i=1}^{|\mathcal{I}|} \nu_x(i) \log \frac{\nu_x(i)}{c\nu_z(i) + d\nu_x(i)}. \end{aligned} \tag{10.5}$$

Since, reasonably, the distortion should measure the perceptual similarity between the images, then the L^∞ distance is used. Specifically, we impose a maximum value L for the pixel absolute distortion (large pixel changes would almost surely lead to annoying artifacts): [8]

$$\max_i |\underline{y}(i) - \underline{z}(i)| \leq L. \tag{10.6}$$

By expressing constraint (10.6) as a function of the $n(i,j)$ variables, the optimization

[8] We stress that, in the case of the infinity distance, the maximum value for the distortion L is not an average (as it is with the L_p^p distortion constraint), but it is defined on a per-pixel basis (point-wise).

problem the Attacker has to solve is the following:

$$\min_{n(i,j)} \sum_{i=1}^{|\mathcal{I}|} \frac{(\sum_k n(k,i)}{n} \cdot \log \frac{(\sum_k n(k,i)/n}{c(\sum_k n(k,i)/n + d\nu_x(i)} + \frac{d}{c} \sum_{i=1}^{|\mathcal{I}|} \nu_x(i) \cdot \log \frac{\nu_x(i)}{c(\sum_k n(k,i)/n + d\nu_x(i)} \tag{10.7}$$

subject to

$$\begin{cases} \sum_j n(i,j) = h_y(i) \ \forall i \\ n(i,j) = 0, \ \forall (i,j) \in \mathcal{I} \times \mathcal{I} : |i-j| > L \\ n(i,j) \geq 0 \quad \forall i,j \\ n(i,j) \ \in \mathbb{N}. \end{cases} \tag{10.8}$$

Some further considerations can be done regarding the formulation (10.7)–(10.8). First of all, we can remove the second constraint in (10.8) by properly restricting the sums in the objective function and in the first constraint. For notational simplicity let us define $\forall i \in \mathcal{I}$ the set $\mathcal{A}(i,L) = \{k \in \mathcal{I} : |k-i| \leq L\}$. Accordingly, we can rephrase the optimization problem in the following equivalent form:

$$\min_{n(i,j)} \sum_{i=1}^{|\mathcal{I}|} \frac{\left(\sum_{k \in \mathcal{A}(i,L)} n(k,i)\right)}{n} \cdot \log \frac{\frac{\left(\sum_{k \in \mathcal{A}(i,L)} n(k,i)\right)}{n}}{\frac{c\left(\sum_{k \in \mathcal{A}(i,L)} n(k,i)\right)}{n} + d\nu_x(i)} + \frac{d}{c} \sum_{i=1}^{|\mathcal{I}|} \nu_x(i) \cdot \log \frac{\nu_x(i)}{\frac{c\left(\sum_{k \in \mathcal{A}(i,L)} n(k,i)\right)}{n} + d\nu_x(i)}. \tag{10.9}$$

subject to

$$\begin{cases} \sum_{j \in \mathcal{A}(i,L)} n(i,j) = h_y(i) \ \forall i \\ n(i,j) \geq 0 \quad \forall i,j \\ n(i,j) \ \in \mathbb{N} \end{cases} \tag{10.10}$$

obtaining a slightly simplified set of constraints. Furthermore, by looking at the first constraint in (10.8) (and (10.10)), we notice that all the optimization variables $n(i,j)$ describing displacements from empty bins to any other bin will have a zero value, that is $h_y(i) = 0$ implies $n(i,j) = 0$ for all j. Let $\mathcal{E}$ be the set of the empty bins, with $\mathcal{E} \subset \mathcal{I}$. It is easy to argue that the actual complexity of the problem is $2L \cdot (|\mathcal{I}| - |\mathcal{E}|)$ which is often much less than $|\mathcal{I}|^2$. By referring to the problem rewritten as in (10.9)–(10.10), the optimization is very fast and, on average, the time taken by the solver to find the optimum mapping is less than one second[9].

[9]Tests have been performed on a computer equipped with a Intel i7 CPU, 8GB RAM, under the Windows 7 operating system.

10.2.3 Pixel remapping phase

After the target histogram h_z has been obtained, the AD needs to actually modify $\underline{y}$ into $\underline{z}$. All the operations performed in this phase will not affect the result of FA's forensic tools, since we assumed that they only consider the histogram of the image. Nevertheless, the AD is not interested in obtaining an attacked image $\underline{z}$ that is perceptually distant from the processed one $\underline{y}$. In this section we describe an approach that allows the AD to implement the pixel mapping defined by the displacement matrix N^* in a perceptually convenient way. Others, even more sophisticated approaches to implement the modifications (mapping) into the pixel domain could be used. It is worth noting that this phase does not have any impact on the result of the forensic analysis.

We begin by recalling that the human visual system (HVS) is known to be less sensitive to noise when this affects highly textured regions. On the contrary, noise in uniform regions, like the sky or a flat wall, is usually much more evident to the observer [165]. Therefore, the first intuition is that, whenever a choice is possible, regions of the image having high variance should be modified first.[10] Furthermore it is useful to iteratively determine which parts of the image are more insensitive to noise through all the computation, using a kind of similarity map between the image at the current iteration and $\underline{y}$. To compute this map, we adopt the Structural Similarity (SSIM) metric introduced by Wang et al. in [165]. This metric quantifies and localizes the structural similarity between two images, and provides a similarity value for each pixel; to determine this value, the system considers the contrast, brightness and other perceptually relevant information in the region surrounding the pixel. Since the image changes during pixel mapping, the map is evaluated several times in order to allow a better (i.e., less perceptible) distribution of noise throughout the image.

Then, we propose the following scheme:

1. Set all pixels as admissible
2. Compute a map of local variance[11] of $\underline{y}$;
3. For each couple (i, j):
   (a) find admissible pixels location having value i;
   (b) scan them selecting the first $n(i, j)$ with highest values in the map;

[10] Notice that this is similar to how embedding changes are made in content adaptive steganography in order to increase the steganographic security [166]. Here, instead, the modifications are made so to minimize the visual distortion.

[11] SSIM cannot be evaluated before the first modification (see comments).

(c) substitute them with j;

(d) remove selected pixels from the admissible ones[12];

(e) if no more pixels of value i must be remapped, compute the SSIM map between the current image and $\underline{y}$;

One first comment regards multiple computations of the similarity map: there is a clear tradeoff between computational complexity and perceptual fidelity. If we compute the map only once, then we do not take into account the distortion that is progressively introduced, experimental results show that this can lead to annoying false-contouring artifacts. On the other hand, computing the SSIM after each single pixel substitution is clearly prohibitive (and useless). A good tradeoff is obtained by computing the map $|\mathcal{I}|$ times, specifically when no more pixels from the i-th level are left to move. Notice that for the first iteration we cannot resort to SSIM (which is a full-reference metric) to get a similarity map, because no changes have been performed yet. Considering the HVS properties introduced before, we simply compute a map of the local variance of the image (working block-wise, with block size 5×5) and use it just for the first step.

While postponing a rigorous experimental validation to Section 10.3, Figure 10.2 shows an example of output image for each of the steps described so far, while Figure 10.3 reports the histograms for the same example: the histogram of a contrast-enhanced image (notice the peak-and-gap artifacts) is fed to the histogram retrieval module, which returns the K histograms yielding the lowest h distance in the dataset. After pixel remapping ($L = 4$), the histogram of the attacked image is close to that of the original one, and the perceptual similarity between the processed and attacked images is satisfactory.

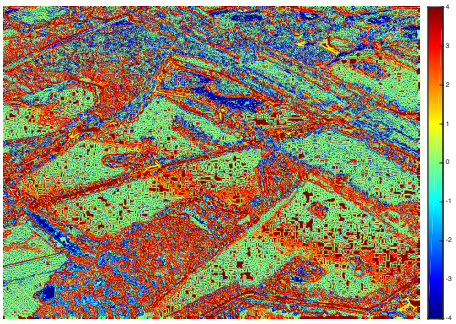
[12]This avoids multiple substitutions of the same pixel.

(a) Original image

(b) Processed image

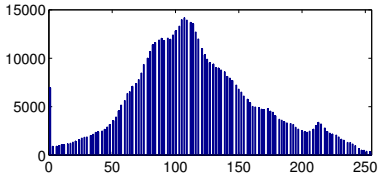
(c) Remapped image

(d) SSIM map between (b) and (c)

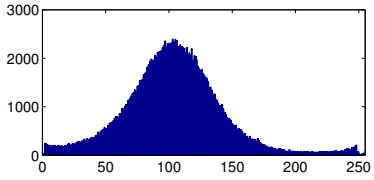


(e) Enhanced difference between (b) and (c)

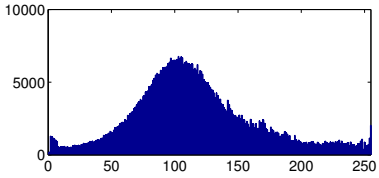
Figure 10.2: An example of application of the universal counter-forensic algorithm: (a) an original non-processed image; (b) its processed (contrast stretched) version, and (c) the image resulting from the proposed C-F technique; (d) structural similarity (SSIM) map obtained at the end of the application of the C-F algorithm; (e) difference between processed and remapped image.



(a) Histogram of the processed image



(b) Best matching histogram retrieved from the database



(c) Histogram of the remapped image

Figure 10.3: Histograms for the example in Figure 10.2: (a) histogram of the processed image, which is compared to those in the database to find the best match (b). The histogram mapping problem is solved yielding (c); notice that the peak-and-gap artifacts in the left histogram have been removed.

10.3 Experimental results

In this section we evaluate the performance of the proposed counter forensic technique against *contrast enhancement detection*, where the goal of the FA is to discover whether an image has been globally enhanced with some processing operator or not. By playing the role of the AD, we apply a histogram-based enhancement to a given image and then use the proposed technique to remove the traces left within the processed histogram.

In order to test the effectiveness of the proposed scheme, we implemented a state-of-the-art algorithm for the detection of histogram based image enhancement [136]. This tool exploits the fact that most histogram-enhancement techniques leave a characteristic fingerprint in the image histogram, namely the peak-and-gaps artifact. This effect is easily exposed in the frequency domain, where the mentioned behavior results in an anomalous amount of high-frequency components. Therefore, by investigating the Fourier transform of the image histogram, the authors devised a very reliable detector. Since this detector considers only the histogram of the image, the proposed universal counter-forensic scheme should be able to defeat it. To check if this is the case, we used that algorithm for distinguishing processed and attacked images from untouched ones. Performances are measured in terms of the Area Under Curve

(AUC) of the detector before and after the attack, while the quality of the attacked images is evaluated using PSNR and Structural Similarity (SSIM).

To generate the enhanced images, we employed two different techniques: one based on γ-correction and one based on histogram stretching. γ-correction enhancement is very simple, being fully described by the following equation:

$$\underline{y}(i) = 255 \times \left(\frac{\underline{x}(i)}{255} \right)^{\gamma} \tag{10.11}$$

where $\underline{y}$ denotes the enhanced image and $\underline{x}$ denotes the original one. Since values of γ very near to 1 would not result in a sensible modification, in our experiments γ is randomly chosen from the set $[0.5; 0.8] \cup [1.2; 2]$.

To formally define the histogram stretching operation, let us denote with l_{min} the gray level at the 1st percentile of the histogram and with l_{max} the gray level at the 99th percentile: then, we perform histogram stretching as:

$$\underline{y}(i) = 255 \times \frac{\underline{x}(i) - l_{min}}{l_{max} - l_{min}}. \tag{10.12}$$

Comparing Figure 10.2 (a) and (b), the effect of histogram stretching in improving image quality is evident. Since, while performing the CF algorithm, the AD wants to preserve the benefits obtained by processing the image $\underline{y}$, in the search phase, he would like to prevent the selection of target histograms having lower contrast than the one obtained with processing. However, such a filtering is implicitly done by searching for the histograms with the minimum value of the h function, as target histograms having different ranges with respect to h_y are discarded.

We conducted our experiments by using images from the UCID dataset [167], which is made of 1338 images of size 512 $\times$ 384. We also used another independent dataset, MIRFLICKR [168], composed of 25.000 images of size 330 $\times$ 500, to prepare the database of non-processed histograms $\mathcal{S}$. Throughout the experiments, all color images are converted to grayscale. The only parameters the Attacker has to choose are the number of candidates for which the optimization problem is solved (we used $K = 10$) and the maximum per-pixel distortion; of course, allowing a higher distortion will yield a more precise mapping of the attacked histogram into the one selected from the database, but will also result in a lower quality of the attacked image. We repeated the experiments with $L = 2$, 4 and 6 in order to investigate the relationship between distortion and effectiveness of the approach.

We performed, separately, contrast enhancement and histogram stretching over all pictures in the UCID dataset and run Stamm's detector on the resulting images; then, we applied the proposed counter-forensic scheme on each processed image, for

various L, and run again the detector. Figures 10.4 and 10.5 show, respectively, the results obtained by hiding traces of contrast-enhancement and histogram stretching operations with the proposed scheme. In both figures, ROC curves obtained for different values of maximum per-pixel distortion are plotted: we can state that the forensic detector no longer distinguishes untouched images from attacked ones even for $L = 2$. Experiments also confirm that, by allowing higher distortion, the AD can further hinder the performances of the detector.

The sole fact that the proposed method successfully deceives a specific detector does not prove its universality. In order to better highlight the fidelity of the remapped histogram h_z to the histogram coming from the database h_x, we calculated for each experiment the χ^2 distance between their normalized versions, defined as:

$$\chi^2(\nu_x, \nu_z) = \frac{1}{2} \sum_{i=0}^{255} \frac{(\nu_x(i) - \nu_z(i))^2}{\nu_x(i) + \nu_z(i)},$$

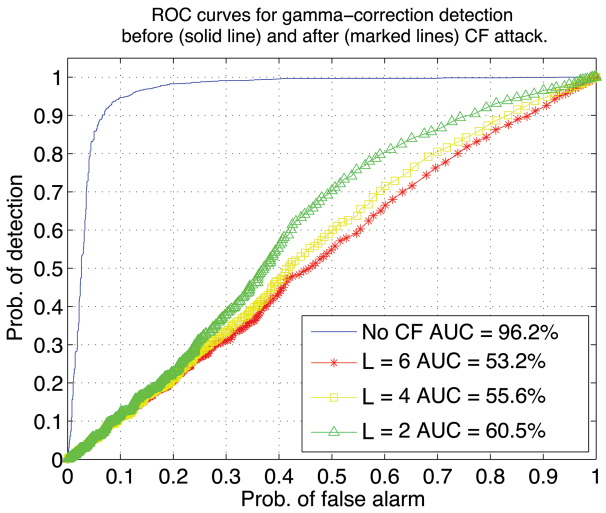
and reported its average and worst case value in Table 10.3. We chose the χ^2 distance to show that the remapped histogram is similar to the untouched one also according to measures that were not directly considered in our scheme. We see that, on average, the χ^2 distance between the histograms takes values in the order of 10^{-2}, which can be considered definitely small. This fact strongly supports the universality claim, because devising an histogram-based forensic detector capable of discriminating between such similar histograms would be extremely difficult.

Of course, the above performance measures would be meaningless if we do not investigate the fidelity of the attacked images to the processed ones: this information is reported in Tables 10.4 and 10.5 for contrast enhanced and histogram stretched images respectively. Notice that PSNR is sufficiently high even for $L = 6$, and the SSIM index confirms an extremely low perceptual distortion. This confirms that the counter-forensic attack does not produce annoying artifacts, nor it removes the benefits introduced by the processing carried by the AD.

L	**Average** χ^2	**95thperc** χ^2	**Average** χ^2	**95thperc** χ^2
2	0.092	0.27	0.060	0.16
4	0.058	0.19	0.032	0.11
6	0.040	0.15	0.019	0.07

Table 10.1: Average and maximum χ^2 distance between the remapped histogram and the one coming from the database for γ-correction (left) and histogram stretching (right) counter-forensics.

In the next chapter, we will investigate how the proposed method can be extended to remove traces left in the histograms of DCT coefficients, thus widening the ap-



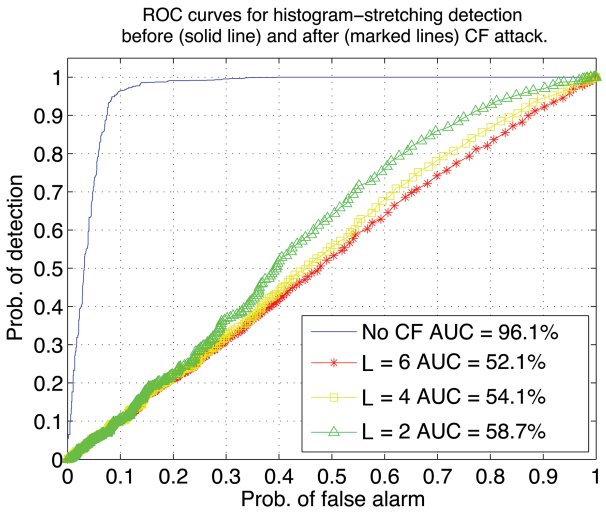
(a)

	PSNR			SSIM			AUC
L	mean	min	95th perc	mean	min	95th perc	
2	44.8	43.0	45.9	0.993	0.951	0.997	0.605
4	39.2	36.7	40.8	0.979	0.917	0.992	0.556
6	36.2	33.4	38.0	0.962	0.881	0.987	0.532

(b)

Figure 10.4: Results for γ-correction counter-forensics. (a): ROC curves for Contrast Enhancement Detector running on γ-corrected images (solid line) and on attacked ones (marked lines); (b): mean value, worst case value and 95th percentile for PSNR and SSIM between processed and attacked images, along with the Area Under Curve obtained by the forensic detector. The values are computed on the 1338 images of the UCID dataset.

plicability of the approach to a broader set of forensics tasks, and in particular to JPEG forensics.



(a)

	PSNR			**SSIM**			**AUC**
L	mean	min	95th perc	mean	min	95th perc	
2	44.8	43.3	45.6	0.994	0.977	0.998	0.587
4	39.2	37.3	40.4	0.981	0.938	0.993	0.541
6	36.1	34.1	37.6	0.964	0.908	0.989	0.521

(b)

Figure 10.5: Results for histogram stretching counter-forensics. (a): ROC curves for Contrast Enhancement Detector running on histogram-stretched images (solid line) and on attacked ones (marked lines); (b): mean value, worst case value and 95th percentile for PSNR and SSIM between processed and remapped images, along with the Area Under Curve obtained by the forensic detector. The values are computed on the 1338 images of the UCID dataset.

Chapter 11

Universal Attacks in the DCT Domain

Detection of multiple JPEG compression of digital images has attracted an increasing interest in the field of multimedia forensics. At the same time, techniques to conceal the traces of multiple compression are being proposed as well. Motivated by the quest for universal approaches, in this chapter we extend the universal counter forensic scheme developed in the previous chapter to the frequency (DCT) domain and devise a counter-forensic technique that aims at making multiple compression undetectable for any forensic detector based on the analysis of the histograms of quantized DCT coefficients.

From a forensic point of view, JPEG compression is one of the most important stages in the processing chain of a digital image, because it leaves peculiar statistical footprints that can be used as a telltale of tampering. In particular, traces left by *multiple* JPEG compressions are usually a powerful tool in analyzing the authenticity of an image. Most of the state-of-the-art multiple JPEG forensic detectors rely on the analysis of features extracted from the histograms of the DCT coefficients. Leveraging on the theoretical results of the first part of this thesis, the universal attack strategy proposed here is optimum against any (forensic) detector within the class of first order-based detectors i.e., detectors based on the analysis of the DCT histograms. In this way, the proposed method establishes a new challenge for future forensic detectors.

The chapter is organized as follows: Section 11.1 briefly reviews the state-of-the-art on JPEG Forensics and Counter-Forensics of multiple JPEG compression. A summary of the effects of multiple JPEG compressions in the frequency domain is given in Section 11.2. In Section 11.3 we discuss the theoretical framework behind the proposed universal technique, while Section 11.4 is devoted to a detailed description of all the phases of the algorithm. Experimental validation is finally reported in Section 11.5 where we show the effectiveness of our approach in removing the artifacts of double and also triple compression, while maintaining a good quality of the image.

11.1 Related works on Forensics and Counter-Forensics of multiple JPEG compression

The interest of forensic researchers in the detection of multiple compressions is motivated by the fact that when JPEG images are manipulated by a photo-editing software and later re-saved in JPEG format, artifacts are introduced in the image. Popescu et al. [169] showed that double quantization entailed by double JPEG compression leaves peculiar artifacts in the histograms of DCT coefficients, especially at low and medium frequencies. Inspired by this fact, the most frequent approach for detecting double JPEG compression consists in analyzing the histograms of the block-DCT coefficients. In this way, the possible correlation between DCT coefficients at different frequencies is discarded thus meeting the assumption that the Defender relies only on first order statistics (in the DCT domain). In [169], Popescu et al. proposed a technique examining the Fourier transform of the histograms of the DCT coefficients. A method for estimation of primary quantization matrix from a double compressed JPEG image that can be used to detection purposes is proposed in [170]. The paper discusses three different approaches: two of them are based on matching the histograms of individual DCT coefficients of the inspected image with the histograms calculated from estimates obtained by calibration [171, 172, 173] followed by simulated double-compression. In [152], Pevný et al. proposed a detector based on Support Vector Machine (SVM) classifiers with feature vectors formed by histograms of low-frequency DCT coefficients. Another method based on histograms of DCT coefficient and SVM is proposed in [174].

Many, recently proposed, forensic techniques rely on the analysis of the first significant digits (FSD) of the DCT coefficients, i.e., a statistic derived from the DCT histograms. Specifically, in [175] the authors found that the distribution of the first significant digit of DCT coefficients in single-compressed images follows a generalized Benford's law. Specifically, the distribution of the FSDs in the frequency domain is investigated in order to tell apart single compressed images from double compressed [153] and, more in general, multiple compressed ones [155]. On the other hand, counter-forensic schemes have been developed in order to remove or disguise the artifacts of multiple compression in the FSD distributions, like in [176] and [177]. A unifying characteristic of these anti-forensic methods is that they are targeted to deceive a specific forensic detector. As such, they do not guarantee that a possible different detector, even based on the analysis of the same class of statistics, would be defeated in turn: the analyst may develop a modified version of the detector that is robust to the counter-forensic approach, thus pushing forward the cat&mouse game. Recently, it is shown that FSD restoration has a strong impact on the distribution of the second significant digits (SSDs) which may be exploited to detect FSD restoration

[178]. To overcome this limitation, that is inherent in the use of targeted counter-forensic techniques, we should turn to *universal approaches*, for which the optimality at least for a certain class of detection methods is guaranteed.

It is proper to say that, many powerful techniques for detecting JPEG recompression go beyond the first order analysis of DCT coefficients. Among them, we mention the work by Chen [179], where the transition probability matrices derived from the differential JPEG 2-D array along various direction are used to reveal the presence of JPEG recompression. Lai and Böhme [180] studied the properties of block convergence during the repeated JPEG compressions with quality factor 100 that can be exploited to estimate the number of recompressions and also uncover local tampering. Based on higher order analysis of DCT coefficients, methods have been derived for detecting aligned double JPEG compression with the same quantization matrix [181, 182].

In the next section, we give some details about JPEG compression to better understand the reason for the artifacts and how they can be used to detect manipulations.

11.2 Basics of JPEG compression

The JPEG standard is today the most widely used method for storing digital images [183]. Despite its lossy nature, JPEG compression is designed not to introduce annoying artifacts in the compressed image, at least for reasonable compression ratios. On the other hand, appreciable artifacts are introduced in the Discrete Cosine Transform (DCT) domain. This fact fostered the development of a whole branch of image forensic techniques. For this reason, we find it worthy to introduce the basic concepts of JPEG coding and briefly describe how JPEG-based forensic algorithms work.

To begin with, we revisit the procedure of compression of a gray-scale image according to the JPEG standard. As regards the notation, the capital letter X is used to denote image $\underline{x}$ in the transformed domain and X_q to indicate the quantized version. In addition, $X(i,j)$, res. $X_q(i,j)$, indicates the transformed coefficient, res. its quantized version, at frequency (i,j) of a generic block; when a particular block k is addressed we denote it by $X_q(i,j;k)$.

As a first operation, the input is divided into blocks of 8×8 pixels each. For each block, the two dimensional DCT is computed. Let $X(i,j)$, $1 \leq i,j \leq 8$, denote the DCT coefficient at frequency (i,j) of the block. The DCT coefficients are then quantized into integer-valued levels $X_q(i,j)$ as follows:

$$X_q(i,j) = \text{sign}(X(i,j))\text{round}\left(\frac{|X(i,j)|}{q(i,j)}\right), \tag{11.1}$$

where the quantization steps $q(i,j)$ are given by a predetermined quantization matrix $Q = \{q(i,j)\}_{i,j=1}^{8}$. After quantization, the values $X_q(i,j)$ of the block are ordered by zig-zag scanning and finally compressed by a lossless encoder.[1] Viceversa, in the decompression procedure, the bit stream is first decoded, and the integer coefficients $X_q(i,j)$ are rearranged back into blocks. Then, the de-quantized DCT coefficients are recovered by multiplying the coefficients with the corresponding entry of the quantization matrix, i.e., $X_q(i,j) \cdot q(i,j)$. Due to quantization, the compression procedure is not invertible and the dequantized coefficients assume only values which are integer multiples of the corresponding quantization step. Finally, the inverse DCT of each block is computed and the result is rounded and truncated so that the pixel values assume integer values in the range $[0, 255]$. The quantization factor is the parameter which determines the amount of approximation introduced by the compression, thus affecting both the compression ratio and the quality of the reconstructed image. Typically, the quantization matrix is fixed by selecting a quality factor (QF), in $[0, 100]$; a high quality factor corresponds to a high quality of the reconstructed image, which also means lower values of the quantization step.

Now let us suppose that an image is compressed twice. Let $X_{q_1}(i,j)$ denote the quantized value at frequency (i,j) after the first encoding with quantization step $q_1(i,j)$. When the image goes through a second compression stage, the resulting quantization level is:

$$X_{q_2}(i,j) = \text{sign}(X_{q_1}(i,j))\text{round}\left(\frac{|X_{q_1}(i,j) \cdot q_1(i,j)|}{q_2(i,j)}\right), \tag{11.2}$$

where $q_2(i,j)$ is the quantization step of the second encoding. Popescu et al. [169] observed that double quantization, and more in general consecutive quantizations, introduce periodic artifacts in the histogram of DCT coefficients. Such a periodic pattern depends on the ratio between the quantization steps, that is, on the ratio between the quality factor of the first and second compressions. More specifically, when the step size decreases (i.e., QF increases) some bins in the histograms are empty, whereas when it increases (i.e., QF decreases) some bins contain a large number of samples and some other bins only a few. It is worth observing that forensic analysers have usually to deal with the first kind of artifacts, since in many applications the goal of the Attacker is to pass off a lower quality image as an image of higher quality. For this reason, we consider the case of multiple compression with increasing quality factors; however it is proper to stress that, being universal, our technique can equivalently be applied in the other case.

Below, we give a brief description of the Watson's model that we will use to characterize the distortion constraint of the DCT coefficients, that is for the estimation

[1] Strictly speaking, DC and AC's coefficients are treated separately by the JPEG standard.

of the Just Noticeable Difference (JND) of the block-based DCT coefficients [184].

Watson's DCT-based visual model

This model establishes a link between modifications in the unquantized DCT domain and their impact in the pixel domain. To account for the sensitivity of the Human Visual System (HVS) to different frequencies, the model defines a *sensitivity table*, which is an 8×8 matrix W whose element $W(i,j)$ gives the amount of modification for coefficient (i,j) that produces a JND in the pixel domain, i.e., the maximum modification of the DCT coefficient (i,j) which is visually undetectable. Lower values in the matrix correspond to higher sensibility for the HVS to that frequency. For our experimental evaluations, we use the matrix of standard values provided in [185]. The sensitivity table is the simplest estimation of the JND, as it does not take into account the local properties of the image. To obtain a more accurate evaluation of the JND for a DCT coefficient we need to consider two additional effects: *luminance masking* and *contrast masking*.

Luminance masking is due to the fact that, according to the HVS, a bright background hides more noise that a dark background [186]. To account for such an effect, Watson's model modifies the matrix for each block of the image on the basis of the value of the DC coefficient (mean luminance intensity of the block). The refined threshold for the (i,j) DCT coefficient of the k-th block is given by

$$T_L(i,j;k) = W(i,j) \cdot \left[\frac{C(1,1;k)}{\overline{C}}\right]^{\alpha}, \tag{11.3}$$

where $C(1,1;k)$ is the DC value of the k-th block, $\overline{C}$ is the mean intensity of the image, and α is a constant. The value suggested by Watson is $\alpha = 0.649$.
Watson's model further refines the estimation of the JND by considering also the contrast masking effect. This is done by evaluating the influence that the AC energy has in the DCT coefficients. The threshold for the DCT coefficient (i,j) of the k-th block is then given by:

$$T(i,j;k) = \max\{T_L(i,j;k), |C(i,j;k)|^{\eta} \cdot T_L(i,j;k)^{1-\eta}\}, \tag{11.4}$$

where η is a constant between 0 and 1 (Watson suggests $\eta = 0.7$).

11.3 The multiple JPEG compression game

Before introducing the proposed scheme, we need to specify the theoretical framework behind it outlining the differences with respect to the case in which the forensic analysis takes place in the spatial domain (see Chapter 10).

Although the problem here is similar to the one addressed in the spatial domain, working in the DCT domain poses several new challenges that need to be solved.

With a reference to the JPEG forensic problem, the FA/AD interplay can be described as follows: on the one hand, the FA wants to tell apart single compressed from multiple compressed images while, on the other hand, AD aims at hiding the effect of multiple compressions so that the image looks like a single compressed one. We assume that, as an extension of the previous case, the Forensic Analyst bases its decision on the analysis of the histograms of the DCT coefficients; as discussed in Section 11.1, this hypothesis actually holds for many existing forensic tools aimed at detecting double (multiple) JPEG compression. The main difference with respect to the previous case is that now the forensic analyst has to combine the information conveyed by 64 histograms, one for each DCT frequency (i, j). At the same time, the AD has 64 histograms to act upon in order to fool the detection, while preserving the constraint on the visual distortion of the image in the spatial domain.

It should now be evident that, although similar to the analogous problem in the pixel domain, the DCT case cannot be treated with the theoretical tools derived in Chapter 4. Instead, the detection of JPEG multiple compression in the frequency domain finds an appropriate background in the analysis of Chapter 7, where the case of *multiple observations* is considered, and the Defender bases the decision on a number of features (or summaries) each one extracted from an observed sequence describing the status of the system. This is exactly the case with JPEG forensic methods that separately analyse coefficients belonging to different DCT frequencies (see, for instance, [152, 174, 153, 155]). Let X be a reference single compressed image in the DCT domain; we denote by $h_{X_{ij}}$ the histogram of the quantized DCT coefficients at frequency (i, j) and with $v_{X_{ij}}$ the normalized versions, where each value of the histogram is divided by the total number of blocks in the image. Moreover, we indicate with v_{ij}, for $1 \leq i, j \leq 8$, the normalized DCT histograms of the image under analysis. Because of the decorrelation property of the DCT transform, the dependence among DCT coefficients in different subbands is low (intrablock dependence), and then we can approximately assume them to be independent.

With reference to the analysis in Chapter (7), it is interesting to observe that, if we consider independent sources, the expression for the optimum strategy of the Defender in the case of *marginal-based* detection given in (7.6) (Theorem 16) can be noticeably simplified. In fact, in the case of independent sources, i.e., when $P_{\mathbf{X}} = \prod_i^s P_{X_i}$ (with s, number of observations), given the set of marginal distributions estimated on the observed sequences $(\hat{P}_1, ..., \hat{P}_s)$, the optimum log-likelihood function

of the Neyman-Pearson test performed by D becomes:

$$\begin{aligned}
&\min_{P\in\mathcal{A}(\hat{P}_1,\ldots,\hat{P}_s)} \mathcal{D}(P||P_{X_1}\cdot P_{X_2}\cdot\ldots\cdot P_{X_s}) \\
&= \min_{P\in\mathcal{A}(\hat{P}_1,\ldots,\hat{P}_s)} \mathcal{D}(P||\hat{P}_1\cdot\hat{P}_2\cdot\ldots\cdot\hat{P}_K) + \mathcal{D}(\hat{P}_1\cdot\hat{P}_2\cdot\ldots\cdot\hat{P}_s||P_{X_1}\cdot P_{X_2}\cdot\ldots\cdot P_{X_s}) \\
&= \sum_{i=1}^{s} \mathcal{D}(\hat{P}_i||P_{X_i}),
\end{aligned} \tag{11.5}$$

where $\mathcal{A}$ is the set of all the joint distributions with marginals $\hat{P}_i$.

Then, in case of approximately independent observations, as the DCT coefficients in different sub-bands are, the optimum log-likelihood test function for the FA can be approximated by the sum of the divergence functions between the normalized DCT histograms, that is:[2]

$$\sum_{i,j=1}^{8} \mathcal{D}(v_{ij}||v_{X_{ij}}), \tag{11.6}$$

From the game theoretical analysis (see Chapter 7), we know that expression (11.6) is the optimum objective function that the AD must minimize in producing the forgery.

Similarly to the previous case, due to the lack of a proper statistical model for multiple JPEG compression, the forensic analysis may follow a *data-driven* approach (see discussion in 9.3). Then, the overall scheme of the attack proposed in 10.2 is preserved and the universal JPEG counter-forensic method works as follows: starting from the multiple compressed image Y, AD produces the attacked image Z in three steps: *retrieval* of a target histogram from a database of untouched single compressed histograms, *computation of the optimum mapping* and *application of the mapping* to the image.

11.4 A universal JPEG counter-forensic algorithm

In this section we describe in detail each phase of the proposed universal counter-forensic algorithm. The Attacker has an image which has been compressed two or more times with increasing quality factor, i.e., with $QF_k > QF_{k-1}$, where k denotes the number of times that the image has been compressed. In order to pass off the image as a single compressed image, the Attacker runs the universal counter-forensic scheme illustrated in Figure 11.1.

[2]It is worth pointing out that, strictly speaking, the analysis of Chapter 7 has been made for the case of known sources only and then the $\mathcal{D}$ function is the optimum log-likelihood test function in that case. We argue that, as it happens for the single observation decision setup, the analysis of the training data case would lead to a more refined (generalized) log-likelihood function (with respect to $\mathcal{D}$).

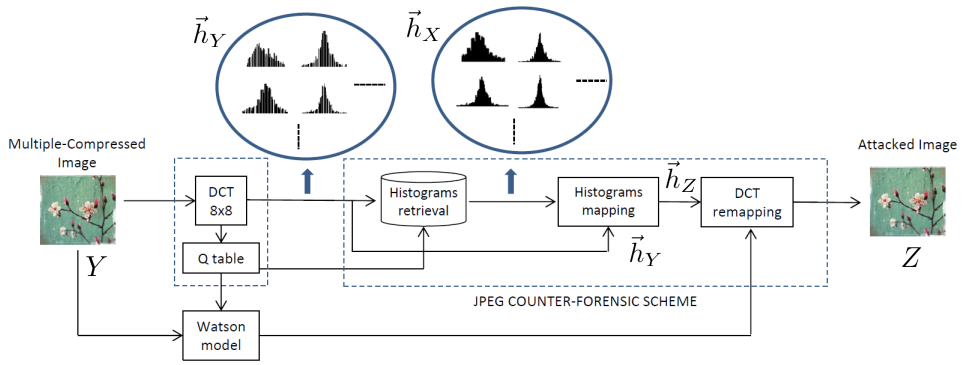


Figure 11.1: The block diagram of the proposed universal JPEG counter-forensic algorithm.

11.4.1 Retrieval phase

We assume that the Attacker has access to a database (DB) of images that have been JPEG compressed only once. Given the multiple compressed image Y_q with quantization matrix $Q_Y = \{q_Y(i,j)\}_{i,j=1}^{8}$, AD searches in the DB of images the one whose vector of DCT histograms is most similar to the histogram vector $\vec{h}_Y = (h_{Y_{11}}, h_{Y_{12}}, ..., h_{Y_{88}})$.
For any frequency (i,j), the similarity between an histogram $h_{X_{ij}}$ and $h_{Y_{ij}}$ is measured by the chi-square distance χ^2, defined as follows:

$$\chi^2(h_{X_{ij}}, h_{Y_{ij}}) = \frac{1}{2} \sum_{m \in \mathcal{C}} \frac{(h_{X_{ij}}(m) - h_{Y_{ij}}(m))^2}{(h_{X_{ij}}(m) + h_{Y_{ij}}(m))},$$

where $\mathcal{C}$ denotes the set of all the values taken by the DCT coefficients.[3] While in the spatial domain these values range from 0 to 255 (pixel values), in the frequency domain the DCT coefficients vary in $[-1024, 1023]$.
We can distinguish between two methods for performing the choice of the 64 DCT target histograms, depending on how the overall χ^2 distance is computed:

- *joint search*: for each image X in the DB the Attacker sums each contribution $\chi^2(h_{X_{ij}}, h_{Y_{ij}})$ provided by each couple of histograms and chooses the vector of 64 DCT histograms minimizing the overall distance. That is, the Attacker looks for the vector of histograms $\vec{h}_X$ which minimizes $\sum_{(i,j)} \chi^2(h_{X_{ij}}, h_{Y_{ij}})$;

[3]Experiments show that using χ^2 in place of $\mathcal{D}$ in this phase lightens the computation without significantly affecting the results.

- *disjoint search*: for each DCT subband, AD searches in the DB the histogram associated with the minimum of the χ^2 distance; i.e., for each (i, j), AD chooses the $h_{X_{ij}}$ which minimizes $\chi^2(h_{X_{ij}}, h_{Y_{ij}})$.

It is evident that, in the second case, the target DCT histograms retrieved from the DB probably belong to different images, i.e., different histogram vectors of the DB. However, in our model, which is confined to the analysis of first-order statistics, this fact does not arise any contradiction, being consistent with the optimum strategy for AD. Besides, it must be stressed that performing the choice in the second way allows to find, for each subband, target histograms which are closer to the source ones with respect to those found in the first way. There are some important considerations we need to do about the retrieval phase. First, we notice that the Attacker can hardly resort directly to a DB of single compressed images, since the corresponding quantization matrices would be probably different from the input quantization matrix Q_Y. Instead of storing thousands of versions of the same image quantized with all possible tables, the Attacker can more practically consider a DB of never-compressed images and, depending on the quantization matrix of the Y_q under analysis, adapt the DB "on-the-fly". This means that, for a given input image Y_q, the Attacker *simulates* the single JPEG compression by quantizing the DCT coefficients according to the input quantization matrix Q_Y. The second observation still concerns practicality: since the search is conducted on the vector of DCT histograms and not on images, only the histograms of unquantized DCT coefficients need to be stored in the DB. This allows to reduce both the size of the dataset and the execution time.

11.4.2 Mapping phase

According to our previous discussion, in this phase the Attacker has to determine the histograms $v_{Z_{i,j}}$ which minimizes the quantity $\sum_{(i,j)} \mathcal{D}(v_{Z_{ij}} || v_{X_{ij}})$, subject to a distortion constraint imposed in order to maintain the final image visually similar to the initial one. In order to characterize this constraint in the frequency domain we rely on the concept of Just Noticeable Distortion (JND). It is reasonable to take the JND as maximum value for the distortion that AD can introduce in the coefficients of the transformed image Y. A commonly used model for the JND is Watson's model [184], described in Section 11.2, which provides a 8×8 sensitivity matrix $W = \{W(i, j)\}_{i,j=1}^{8}$. Each entry of the matrix $W_q(i, j)$ provides the maximum amount of distortion which can be introduced in the quantized DCT coefficients of the subband (i, j) without generating annoying artifacts. Let $W_q = \{\text{round}\,(W(i, j)/q_Y(i, j))\}_{i,j=1}^{8}$ denote the quantized Watson's matrix, approx-

imated to integer values[4].[5] The maximum distortion for the (i,j) coefficient is given by $K(i,j) = W_q(i,j) \cdot L$ for some $L \geq 1$ (larger L allow to obtain more accurate mapping at the price of a higher visual distortion). Interestingly, since distortion constraints are defined per subband, the problem can be solved as 64 separate minimizations:

$$\min_{|Z(i,j)-Y(i,j)|\leq K(i,j)} \mathcal{D}(v_{Z_{ij}}||v_{X_{ij}}), \quad \forall (i,j), 1 \leq i,j \leq 8. \tag{11.7}$$

Let us focus on a single DCT subband and analyze the corresponding problem. It is useful to introduce the *transportation matrix* $N_{ij} = \{n_{ij}(m,r)\}_{m,r=1}^{|\mathcal{C}|}$, where each term $n_{ij}(m,r)$ indicates the number of elements in $h_{Y_{ij}}$ which must be moved from the m-th to the r-th bin. Let n_{ij} be the total number of blocks in the image (i.e., the number of DCT coefficients for each frequency (i,j)). Each constrained optimization problem in (11.7) is quite similar to the one in Section 10.2.2 and, similarly, can be rephrased in function of the $n_{ij}(m,n)$ variables as follows:

$$\min_{n_{ij}(m,r)} \sum_{r=1}^{|\mathcal{C}|} \frac{(\sum_m n_{ij}(m,r))}{n} \cdot \log \frac{(\sum_m n_{ij}(m,r))}{n v_{X_{ij}}(r)}, \tag{11.8}$$

subject to

$$\begin{cases} \sum_r n_{ij}(m,r) = h_{Y_{ij}}(m) \ \forall i \\ n_{ij}(m,r) = 0, \ \forall (m,r) \in \mathcal{I} : |m-r| > K(i,j) \\ n_{ij}(m,r) \geq 0 \quad \forall m,r \\ n_{ij}(m,r) \ \in \mathbb{N} \end{cases} \tag{11.9}$$

where the histogram $h_{Y_{ij}}$ and the distortion constraint were rewritten in terms of $n_{ij}(m,r)$ variables. Solving problem (11.8)–(11.9) provides the optimum map N_{ij}^*, from which we obtain the final attacked histogram $h_{Z_{ij}}$ by computing $\sum_m n_{ij}^*(m,r)$ for each r. Problem (11.8)–(11.9) is a convex mixed integer non-linear problem (MINLP) [161] for which a global optimum solution exists and efficient solvers are available for the resolution. It is worth observing that the number of optimization variables is given by $|\mathcal{C}|$, that is the cardinality of the alphabet of the DCT coefficients ($|\mathcal{C}| = 2048$), and it does not depend on the size of the image. This value seems to be significantly larger compared to the one in the pixel domain (i.e., 256); however, since the statistics of the DCT coefficients are usually peaked around the mean value [187], the number of variables can be noticeably reduced by cutting off the bins below

[4] Performing the rounding for computing W_q may cause a slight violation of the JND constraint, but it is preferable for the remapping operation.

[5] In this phase, we do not account for the refined model of the sensitivity matrix, which would complicate significantly the analysis, by introducing a block-wise dependence.

$m_{\min}$ (where $m_{\min}$ is s.t. $h_{Y_{ij}}(m) = 0\ \forall m < m_{\min}$) and above $m_{\max}$ (where $m_{\max}$ is s.t. $h_{Y_{ij}}(m) = 0\ \forall m > m_{\min}$). Let $\mathcal{E}$ be the set of the empty bins within the interval $[m_{\min}, m_{\max}]$. It is easy to argue that the actual number of variables of the (i,j)-th minimization is $2K(i,j)\cdot((m_{\max}-m_{\min})-|\mathcal{E}|)$, which is usually much lower than $|\mathcal{C}|$. Moreover, since JPEG compression quantizes more heavily the high-frequency DCT coefficients, the complexity of the minimizations will decrease at higher frequencies, because histograms will tend to cluster around zero.

Similarly to the optimization problem (10.7)–(10.8), the problem in (11.8)–(11.9) has very close ties with the classical *transportation problem*: the difference is that, according to the definition of the attacker's strategy, the Attacker is satisfied with any distortion less than $K(i,j)$, that is, he/she is not concerned about minimizing the distortion provided that it is less that $K(i,j)$. In this way, the optimum attacking strategy in (11.8)–(11.9) provides a distortion-limited map N_{ij} even when the classical transportation problem, which moves $v_{Y_{ij}}$ exactly into $v_{Z_{ij}}$, would introduce too much distortion into the image (i.e., more than $K(i,j)$).

To sum up, the mapping phase provides the Attacker with the 64 matrixes N^*_{ij}, $1 \leq i,j \leq 8$; each matrix N^*_{ij} defines the modifications that must be made on the DCT coefficients at freqeuncy (i,j) in order to obtain the optimum attacked histogram $h_{Z_{ij}}$.

11.4.3 Implementation of the mapping

After obtaining the transportation matrices, it is necessary for the Attacker to implement the mapping in such a way to reduce as much as possible the visual distortion introduced in the image. Notice that, since the forensic detector relies on the histograms of the DCT coefficients, the result of the attack in terms of detectability of the produced forgery depends only on the results of the mapping phase, and it is not affected by the modifications performed in this phase. In the following, we describe an approach that allows the Attacker to implement the modifications set by the matrixes N^*_{ij}'s in a perceptually convenient way. The basic idea is to exploit the different sensitivity of the Human Visual System to the DCT coefficients of the different blocks in order to first modify the coefficients in those blocks where the HVS is less sensitive. To do so, we exploit the values of the JND provided by Watson's model which, as described in Section 11.2, are indeed block-dependent. Again, modifications are implemented separately on the DCT coefficients of each frequency subband.

Below, we describe the main steps of the proposed scheme for the implementation of the transportation matrix N^*_{ij} in the generic subband (i,j):

1. Set all the coefficients as "admissible";

2. Rank the blocks based on the value of the threshold $T(i,j)$ in decreasing order: block k such that $T(i,j;k)$ is maximum is ranked first, and so on;

3. For each couple of values (m,r) such that $n_{ij}(m,r) \neq 0$ proceed as follows:

   (a) find the blocks with admissible DCT coefficients having value m;

   (b) select the first $n_{ij}(m,r)$ according to the order established by the ranking;

   (c) substitute them with r;

   (d) remove selected coefficients from the admissible ones[6];

The procedure is applied to all the 64 DCT subbands.
Notice that, according to the above scheme, the Attacker computes the thresholds of the JND only once, without updating them to account for the variations caused by incremental modifications. In principle, lower distortion can be introduced by iteratively updating the thresholds. However, since Watson's model is mainly concerned with the average luminance and energy of each block, the benefit obtained by iterative updates is not relevant enough to justify the increased computational complexity, and for this reason this feature was not implemented.
At the end of the procedure, the adversary gets the transformed image Z_q with the quantized 'remapped' DCT coefficients, whose DCT histograms are, by construction, the 64 target histograms $h_{Z_{ij}}$, $1 \leq i,j \leq 8$, obtained in the mapping phase. Computing the de-quantized coefficients and applying the inverse DCT transform yields the final attacked image $\underline{z}$ in the pixel domain. The image will appear visually close to the input one, but its histograms will show traces of just one compression step.

11.5 Experimental validation

In this section we put the technique described in the previous sections at work, in order to show that it actually conceals the traces of multiple compression in the histograms of the DCT coefficients. To test the effectiveness of the proposed scheme, we implemented a simple and common double compression detector based on the so-called calibration technique, borrowed from steganalysis [171]. Calibration is a procedure allowing to estimate the original distribution of the quantized DCT coefficients by removing a small number of rows/column (in the spatial domain) to disrupt the block structure of JPEG images. The calibration-based detector simply works by calculating the "expected" histograms for quantized DCT coefficients and

[6] This avoids multiple substitutions of the same coefficients.

comparing them to the histograms of observed DCT coefficients in the given image.[7] If the image was compressed only once, the expected histogram is quite similar to the observed one (the χ^2 distance is used to compare histograms); on the other hand, if multiple compressions were performed, the expected histogram differs significantly from the observed one. We limit the detector to consider the first 12 DCT coefficients (in the JPEG zig-zag ordering), because higher frequency coefficients are not reliable for this kind of analysis, due to the sparsity of histograms induced by quantization. To enforce the fact that the proposed scheme is *universal*, we also consider two data-driven detectors which work on first order statistics of the DCT coefficients. Notice that, with respect to the detector based on calibration, these detectors works exclusively with first-order statistics of the DCT coefficients, thus belonging to the class of detectors considered in the theoretical analysis. Specifically, the undetectability of our method is validated against the detector in [155], based on the analysis of the distribution of the first significant digit (FSD) of DCT coefficients absolute values, and an SVM-based detector directly fed with the histograms of the block-DCT coefficients, inspired by the detector in [152]. The idea behind the latter detector is simple: rather than considering specific features derived from the first order statistics of DCT coefficients, we can directly feed the SVM with a feature vector formed by the histograms of block-DCT coefficients. To build the feature vector, before concatenating the DCT histograms, each of them is arranged on a reference support which is determined so to be large enough to accommodate the histogram content, for all the quality factors considered. Specifically, or the DC histogram, we consider the range determined by the JPEG 100% (4096 bins), whereas, to save the length of the feature vector, a worst-case range extent for the histograms of the AC coefficients is determined experimentally.

Besides, we evaluate the perceptual similarity between the input image and the one obtained after the implementation of the mapping.

To generate the reference database for the Attacker, we computed the histograms of each DCT coefficient from more than 2000 grayscale uncompressed images, obtaining 64 histograms per image. The database is generated from the UCID database (1338 images of sizes 512×384) and images from a personal database of images in raw format (700 images) of sizes 1072×712.[8]

Then, 25 grayscale uncompressed images were chosen from different sources (a

[7] The expected histogram is obtained by estimating the histograms of unquantized coefficients (using calibration), then quantizing them according to the quantization factors available in the JPEG header of the file.

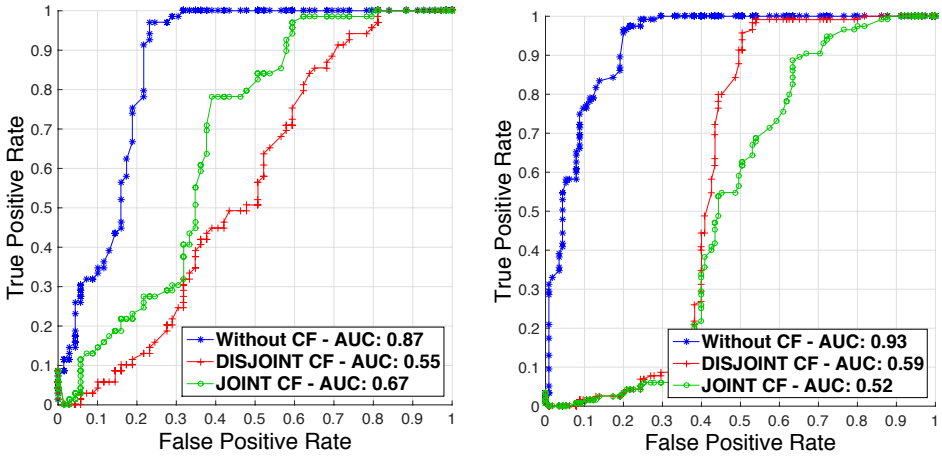
[8] We built a personal dataset of images in tiff format in order to increase the size of the database, due to the little availability of image datasets in uncompressed format at the time of these experiments.

different personal dataset of images of the same size) to perform the tests. Both the database and the test images are available in the website (http://clem.dii.unisi.it/~vipp/index.php/download/imagerepository). For a multiple compressed image, consistently with the notation introduced in Section 11.4, we denote by $\{QF_1,QF_2,...,QF_k\}$ the quality factor used for the first, second, ..., k-th compression step. Each test image was used to generate, using the `imwrite` function of Matlab, the following images: three double-compressed versions of the image, with quality couples $\{65,85\}$, $\{75,90\}$ and $\{85,95\}$; five triple-compressed versions, with quality triplets $\{65,85,90\}$, $\{70,75,95\}$, $\{70,80,95\}$, $\{75,85,95\}$, $\{80,85,95\}$; for each of the above multiple-compressed images, one single-compressed image with quality given by QF_k (these images serve to test the discrimination capability of a forensic detector).

Then, we applied the JPEG counter-forensic scheme to each of the above images, using $L = 4$; the experiment was performed using both the *disjoint* and *joint* search (as defined in Section 11.4.1) in order to compare the performance.

We now describe the experiments we did to validate our method against the calibration-based detector. In the first experiment, the detector was used to discriminate between double-compressed and single-compressed images, generated as detailed above, whereas in the second experiment we tried to discriminate between single- vs. triple- compressed images. To measure the performance, we computed the Receiver Operating Characteristic (ROC) curve of the detector before and after the application of our JPEG counter-forensic attack, along with the Area Under the Curve (AUC). As we can see in Figure 11.2, the detector behaves reasonably well in absence of counter-forensic attacks, while its performance dramatically drops after the proposed scheme is applied. Moreover, we see that both the *disjoint* and *joint search* methods lead to reasonably good performance in terms of deceiving the calibration-based detector, with the former slightly favored at small probabilities of false alarm. It is worth pointing that, the *joint* search approach would be preferable to the *disjoint* one because it is forensically more secure; however, it suffers more the limited size of the database. Then, quite expectedly, when the joint search approach is adopted, the algorithm needs significantly larger DB for getting good performances.

The results for the case of single- vs. triple- compressed images are plotted in Figure 11.2: we see that good CF performance are obtained in the leftmost part of the ROC, corresponding to low false alarm probability. For false alarm probabilities over 0.4, the detector manages to distinguish between single- and triple- compressed images even in the presence of counter forensic attacks. This fact is mainly due to the different distribution of quality factors between triple compressed and single compressed images in the considered experiments; from a forensic point of view, however, false alarm probabilities as high as 0.4 are not of interest.



(a) (b)

Figure 11.2: ROC curve for the calibration-based detector for single-vs-double (a) and single-vs-triple (b), before and after application of the proposed method.

Experiment	Mean SSIM	Std. dev. SSIM
Double compression - Disjoint	0.920	0.033
Double compression - Joint	0.903	0.046
Triple compression - Disjoint	0.945	0.025
Triple compression - Joint	0.935	0.027

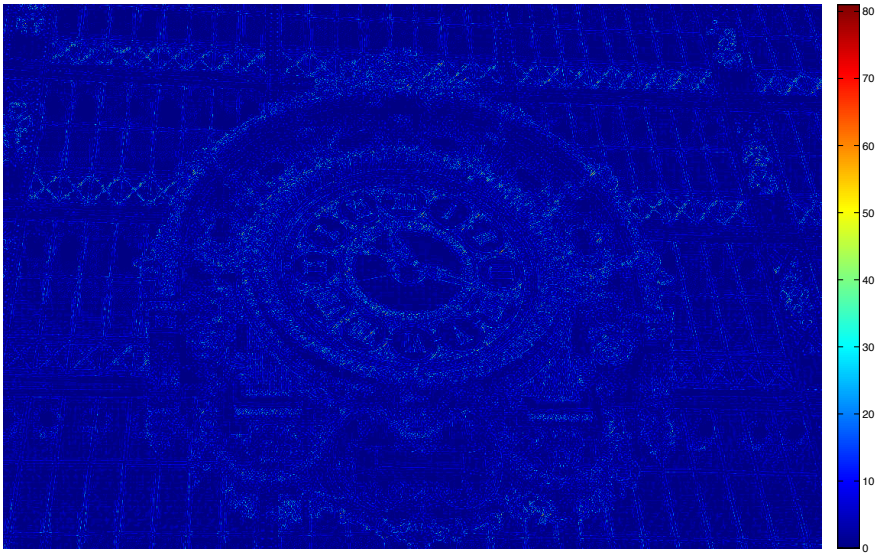
Table 11.1: Performance of the proposed method in terms of perceptual quality. Each row shows the mean and the standard deviation of the SSIM obtained for a given experiment. For double compression a total of 75 images were processed, 125 for triple compression.

Let us now turn to consider the perceptual quality of the attacked images. We evaluated the quality by means of the Structural Similarity (SSIM) index [165], computed between the original image and the image at the output of the proposed scheme. Results are given in Table 11.1. We can confirm that using the disjoint search on the database (as defined in Section 11.4.1) allows the Attacker to obtain better results in terms of perceptual quality of the produced image. It may seem counter-intuitive to the reader that a better similarity was obtained in the case of triple compression: this is actually not surprising if we keep in mind that the similarity is computed between the input and the output of the CF scheme, and it is easier to keep fidelity

to an image whose quality was not so high from the beginning (as it is a triple compressed images). A practical comparison between a multiple-compressed image and the counter-forensic version is shown in Figure 11.3; the DCT histograms at some frequency are showed in Figure 11.4.

(a)

(b)



(d)

Figure 11.3: Example of an application of the JPEG counter forensic algorithm: (a) triple-compressed image with qualities {70, 80, 95} and (b) its counter-forensic version; (c) absolute difference between (a) and (b).

Experiments were also conducted against the first significant digit (FSD) features-based detector [155] for the case of single-vs-double compression and the SVM-based

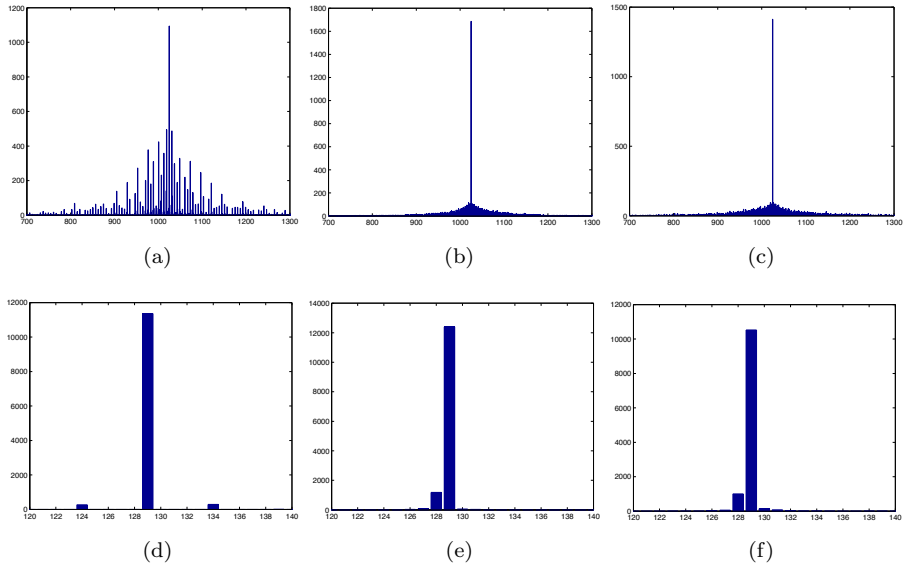


Figure 11.4: DCT Histograms for the example in Figure 11.3: (a) histogram of DCT coefficients at frequency (1,2) coming from the multiple-compressed image, (b) the corresponding target histogram from the DB and (e) the remapped version obtained with the proposed scheme. (g) - (i), histograms of DCT coefficients at frequency (3,4), ordered as in the previous line.

detector fed with the histograms of the block-DCT coefficients. To test our method against these detectors, we enlarged the set of grayscale never-compressed images to 40. A random subpart (about 70%) of these images in these two sets are then used for training the SVM. Specifically, the set of single and double-compressed images were generated as above; then, the feature vectors were computed from the images of both classes and used to train the classifiers. The remaining images (about 30%) were used as test set; we first produced the single and double compressed images, then we ran the universal C-F attack with $L = 3$ to hide the traces of recompression in the double compressed images. Cross validation was performed by repeating the experiments 10 times. The result of the classification against the detector in [155] is shown in Table 11.2 for the case of detection of double compression:[9] whereas the detector is able to perfectly tell apart double compressed images from single compressed ones, its performance drops after the application of the counter-forensic algorithm. The average

[9]This detector only consider the first 15 DCT coefficients taken in zig-zag order in the feature vector.

perceptual quality of the final attacked images is also reported. These experiments show that the proposed method is able to fool the detector in [155], although not tailored for this purpose: This is expected due to the universality of the method (as long as first-order statistics are considered).

Even the second, more general data-driven detector based on the block-DCT histograms, inspired by [152], is fooled by the universal C-F attack. To assess the validity of the attack against this detector, we used grayscale images from the RAISE dataset [188]. This dataset consists of 8156 high-resolution images (of sizes (3008×2000, 4288×2848 and 4928×3264), uncompressed and guaranteed to be camera-native, belonging to various categories (e.g., outdoor, landscape,..).[10]

Specifically: a subset of 2000 images were used as database for the Attacker; in addition, 1000 images were selected for evaluating the performance of the detector. We built the training set by randomly selecting 900 images (100 of which were used for 5-fold cross validation), while the remaining 300 images were used for the test. We considered the following pairs of quality factors for the first and second compression: $(QF_1, QF_2) \in \{(65;85);(70;85);(75,90);(85;95)\}$. The images in the training set were compressed once with QF_2 to build the set of single compressed and twice with (QF_1, QF_2) to build the double compressed set. From the test set, single and double compressed images were generated as above. Then, for the images in the double compressed set we ran the C-F scheme with $L = 4$, to build the set of the attacked images for both the joint and disjoint search case.[11]

As a result, while the double compressed images are correctly classified (AUC ≈ 0.99), the classification of the attacked images fails both in the case of joint and disjoint search. The AUC is about 0.51 and 0.42 respectively. Figure 11.5 shows the ROC curves before (single vs double compressed images) and after (single compressed vs attacked images) the application of the algorithm for the case of joint search.

We also verified that, not surprisingly, the SVM-based detector makes ineffective the attacks in [177] and [176], focused on the FSD domain, yielding almost ideal detection performance. The concealment of the traces in the FSD domain, in fact, leaves traces back in the histograms of the DCT coefficients that the SVM classifier is able to recognize.

[10] The RAISE dataset was not yet available at the time of our previous experiments.
[11] As before, cross validation is performed by repeating the experiment 10 times.

Test	AUC	PSNR	SSIM
No CF	1	inf	1
JOINT CF	0.52	31.78	0.88
DISJOINT CF	0.49	32.48	0.91

Table 11.2: Performance of the proposed method against the FSD features-based detector. When no CF scheme is applied, by default, PSNR= inf and SSIM = 1.

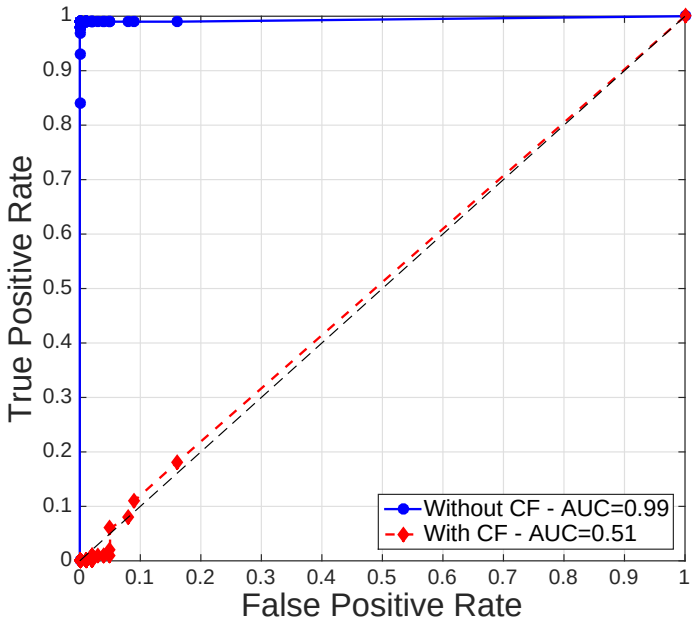


Figure 11.5: ROC curve for the detector based on block-DCT histograms before and after application of the proposed CF method. The joint search approach is considered.

Chapter 12

The Security Margin Concept in Image Forensics

The Security Margin (see Chapter 5) is a powerful concept which permits to summarise into a single quantity the asymptotic behaviour of the game between the Attacker and the Defender. Its practical application, however, poses a number of problems due to the assumptions behind the definition. In this chapter, we first discuss a possible practical meaning of the Security Margin concept (Section 12.1) and present its possible use within a multimedia forensics scenario (Section 12.2).

12.1 Practical meaning of the $\mathcal{SM}$

Binary detection is one of the most common problems in image and multimedia forensics. In fact, gathering information about the device that was used to produce a certain image plays a crucial role in many investigations. In a similar way, the analyst may be interested to decide if a certain processing operator has been applied to a given image, that is to distinguish between the class of images that underwent a certain processing and those which did not. Within this framework, according to the theoretical analysis, estimating the $\mathcal{SM}$ between the classes of original and processed images may help to understand how difficult it is for the adversary to completely delete the traces left by the processing operator and to answer the following question: who, between the FA and the AD, is going to win the multimedia forensic game?

In practice, a large $\mathcal{SM}$ means that if the adversary wants to be sure to make the forensic analysis fail, he has to introduce a large distortion ($L > \mathcal{SM}$), thus possibly compromising the visual quality of the forgery. By adopting the opposite (defender's) perspective, the $\mathcal{SM}$ provides a qualitative measure of the goodness of the performance of a detector: a detector which is able to distinguish well between two classes, should not be fooled if the adversary introduces a distortion less than the $\mathcal{SM}$.

As already discussed in Section 9.3, when a statistical model for the two classes of images is not available, as it is often the case in multimedia signal processing applications, the theoretical analysis developed in the first part of this thesis cannot be

applied straightforwardly, but needs to be adapted so to fit the practical scenario. Similarly, the concept of the $\mathcal{SM}$ between the two classes cannot be directly calculated as detailed in Chapter 5 and needs to be readapted. Such adaptation should consider the data-driven approach followed by the analyst, usually based on machine learning techniques, whereby the characteristics of the image classes are derived from a number of examples.

12.2 $\mathcal{SM}$ in data-driven Image Forensics

By sticking to the notation introduced in Section 9.3, let $\mathcal{C}$ and $\mathcal{C}'$ be two classes of images, for instance images acquired by a scanner and images produced by a camera. Given that the statistical model of the two classes is unknown, the analyst relies on two sets of training images belonging to $\mathcal{C}$ and $\mathcal{C}'$, let us call such sets $\mathcal{S}$ and $\mathcal{S}'$.

Given a test image I, the goal of the Defender is to accept or reject the hypothesis that I belongs to $\mathcal{C}$, by relying on the first order statistics of I, that is the image histogram h_I; on the other hand, the goal of the Attacker is to take an image J belonging to $\mathcal{C}'$ and modify it in such a way that the Defender classifies it as belonging to $\mathcal{C}$. From the theoretical definition of $\mathcal{SM}$ (see Section 5.1), we can argue that *in some sense* the Security Margin between J and $\mathcal{S}$ (which is the only available representation of $\mathcal{C}$) is the minimum *EMD* between h_J and the histograms of the images in $\mathcal{S}$, namely

$$\mathcal{SM}(J, \mathcal{S}) = \min_{I \in \mathcal{S}} EMD(h_J, h_I). \tag{12.1}$$

In fact, if the distortion allowed to the Attacker is larger than $\mathcal{SM}(J, \mathcal{S})$, A can modify J in such a way that its histogram is equal to the histogram of one of the images in $\mathcal{S}$, thus making a reliable distinction impossible. In the same way, we could define the $\mathcal{SM}$ between two classes of images as the average minimum *EMD* between the histograms of the images in one class and those of the images in the other class:

$$\mathcal{SM}(\mathcal{S}', \mathcal{S}) = \frac{1}{|\mathcal{S}'|} \sum_{J \in \mathcal{S}'} \min_{I \in \mathcal{S}} EMD(h_J, h_I). \tag{12.2}$$

A similar analysis can be applied when the distinction between the classes $\mathcal{C}$ and $\mathcal{C}'$ is carried out in a transformed domain, e.g., the block DCT domain.

In the next paragraph, we exemplify the above ideas by applying them to two well-known problems in image forensics: detection of contrast enhancement and double JPEG compression.

12.2.1 Histogram-based detection of contrast enhancement

As already pointed out, given that most contrast enhancement operators work directly on the image histogram, forensic tools for contrast-enhancement detection usually rely on the analysis of the image histogram and hence fit well the theoretical setup considered in this thesis (e.g., [149, 136]). In this framework, estimating the $\mathcal{SM}$ between the classes of original and contrast-enhanced images as specified in equations (12.1) and (12.2) may help to understand how difficult is to make the enhancement operation undetectable.

To exemplify the above ideas we considered the images contained in the MIRFLICKR dataset [189]. These are 25,000 JPEG images of size 333×500.[1] We randomly split the images in two sets $\mathcal{S}$ containing 24,000 images and $\mathcal{S}'$ with 1,000 images[2]. We use set $\mathcal{S}$ as evidence for the class of never processed images. Then, we contrast-enhanced the images in $\mathcal{S}'$ by applying a gamma correction operator with various γ [190]. Eventually, we used equation (12.1) to compute the $\mathcal{SM}$ between the images in $\mathcal{S}'$ and $\mathcal{S}$. The results we obtained are reported in Figure 12.1 where we show the distribution of the $\mathcal{SM}$ across all the images in $\mathcal{S}'$ for both the cases of squared Euclidean distance and maximum distance for $\gamma = 0.8$. The $\mathcal{SM}$ ranges from a minimum of 1.6 to a maximum of 195.3 for the square Euclidean distance, and from 5 to 85 for the L_∞ case. By looking at the figures we see that the images for which the $\mathcal{SM}$ takes large values are very few. In fact, the 95th percentile is 24.5 and 64 respectively in the two cases. It is worth noticing that such outliers correspond to very bright images which are actually not suited to be γ-corrected with a low γ (resulting in a restriction rather than in an expansion of the histogram). In such a case, unless $\mathcal{S}$ is very large and contains also some almost saturated images, it is plausible that a larger distortion is needed to map the histogram into the closest histogram retrieved from $\mathcal{S}$. Figure 12.2 shows the original image with the maximum $\mathcal{SM}$ ($\mathcal{SM} = 85$) for the case of L_∞ distance and the corresponding histograms before and after gamma correction. For the same motivation, large $\mathcal{SM}$'s are also obtained with very dark images when we perform the enhancement with a large γ.

In Table 12.1, we show the average $\mathcal{SM}$, computed as stated in (12.2), for different values of γ. The values in the table suggest that, for instance, with a strength of the enhancement of 1.8, the Attacker must introduce an average square distortion in the order of 26 and a maximum (non-squared) distortion in the order of 15 for a perfect concealment of the traces left by the γ correction operator.

Figure 12.3 reports the distribution of the $\mathcal{SM}$ across all the images in $\mathcal{S}'$ for the

[1] We consider the MIRFLICKR dataset because of its very large size.

[2] The reason for the unbalanced sizes of the sets is the following: in order to have a good estimate of the value of the $\mathcal{SM}$ between a processed image $J \in \mathcal{S}'$ and the class of the never processed images $\mathcal{S}$, we need a descriptive characterization of $\mathcal{S}$.

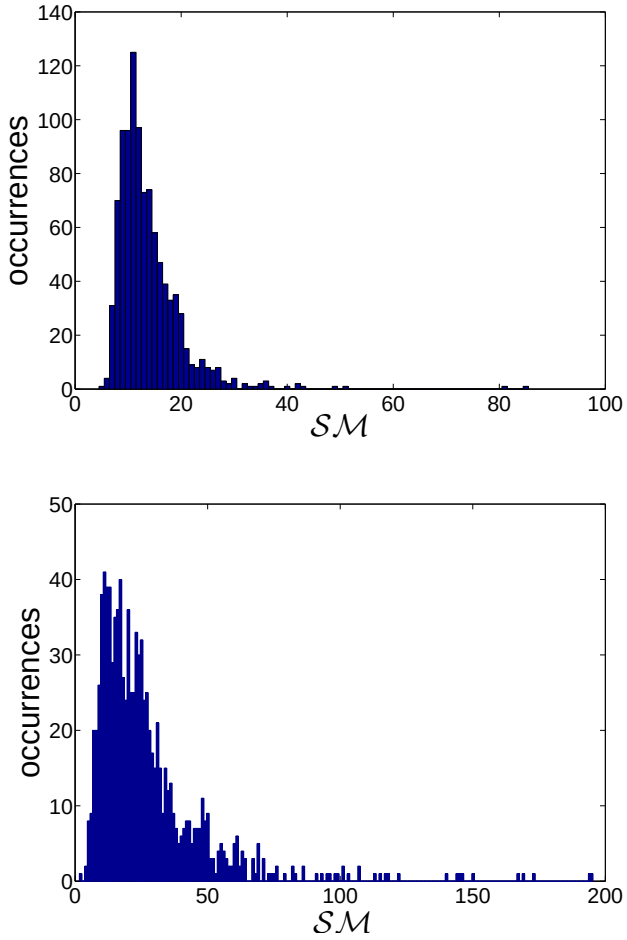


Figure 12.1: Distribution of the $\mathcal{SM}$ across the images in $\mathcal{S}'$ for the case of L_∞ (above) and L_2^2 (below) distance. The strength of the enhancement operator is $\gamma = 0.8$.

Table 12.1: Average $\mathcal{SM}$ between $\mathcal{S}$ and $\mathcal{S}'$ for various values of γ.

	γ					
	0.3	0.8	1.3	1.8	2.3	2.8
$\mathcal{SM}_{L_\infty}$	27.3	13.8	13.8	14.9	16.1	17.4
$\mathcal{SM}_{L_2^2}$	48.8	27.4	25.8	26.1	26.3	26.8

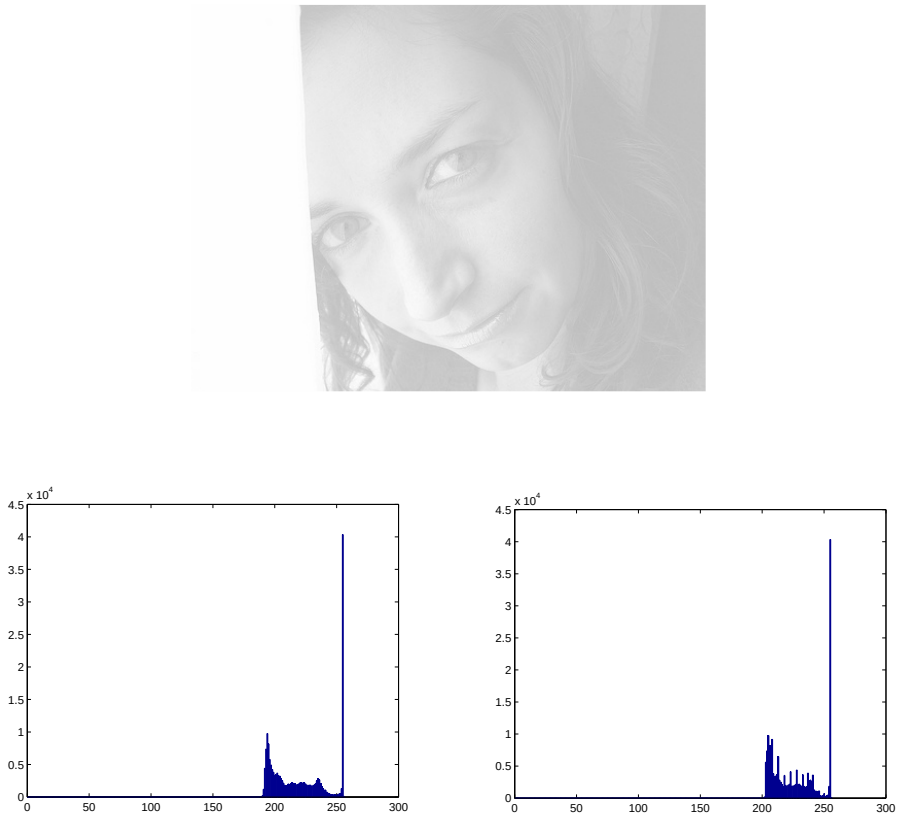


Figure 12.2: Image in $\mathcal{S}'$ which yields $\mathcal{SM} = 85$ for the case of γ correction. Original image (top). Histogram of the original and γ-corrected image with strength 0.8 (bottom).

case of maximum distance when the images are enhanced through histogram stretching to build $\mathcal{S}'$. Even in this case, we notice the presence of some large $\mathcal{SM}$ values corresponding to images with bright (almost white) areas in dark backgrounds or viceversa, which are not suited to be enhanced with stretching of the histogram. The original image corresponding to the maximum $\mathcal{SM}$ ($\mathcal{SM} = 76$) and the histograms before and after the histogram stretching are shown in Figure 12.4.

We conclude this section by observing that the values given Figure 12.1 and Table 12.1, as well as those in Figure 12.3, must be interpreted with care. First of all, the results depend on the used database and in particular on its size; the values

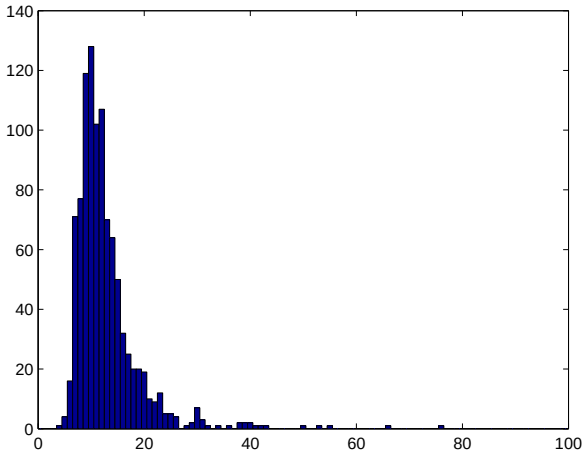


Figure 12.3: Distribution of the $\mathcal{SM}$ across the images in $\mathcal{S}'$ for the case of L_∞ distance. The enhancement is performed trough histogram stretching.

of the $\mathcal{SM}$ may decrease by considering larger database. Secondly, introducing a distortion equal to the $\mathcal{SM}$ ensures the success of the attack asymptotically and in the presence of an optimum detector. Deceiving practical forensic operators may be significantly easier, and hence may require a considerably lower distortion. This is exactly the case when the optimum attack is pursued against the detector in [136], as the experimental results in Chapter 10 show. Secondly, the visual impact of the attack can not be measured only in terms of L_2^2 or even L_∞ distance, since it also depends on how the attack is implemented in the pixel domain, that is on which specific pixels are chosen to implement the mapping defined by the NWC rule (the practical implementation of histogram mapping proposed in Section 10.2.3 provides insights on the visual impact. However, the proposed method is only a possibility of performing the remapping in the pixel domain and in principle other solutions leading to a smaller visual impact could be found).

12.2.2 Detection of double JPEG compression

Here we focus on the distinction between images which have been JPEG-compressed once from those which have been compressed twice, which, as discussed in the previous chapter, is another common problem in image forensics. The most frequent approach consists of analysing the histogram of block-DCT coefficients, since the double quantization entailed by double JPEG compression leaves peculiar artifacts

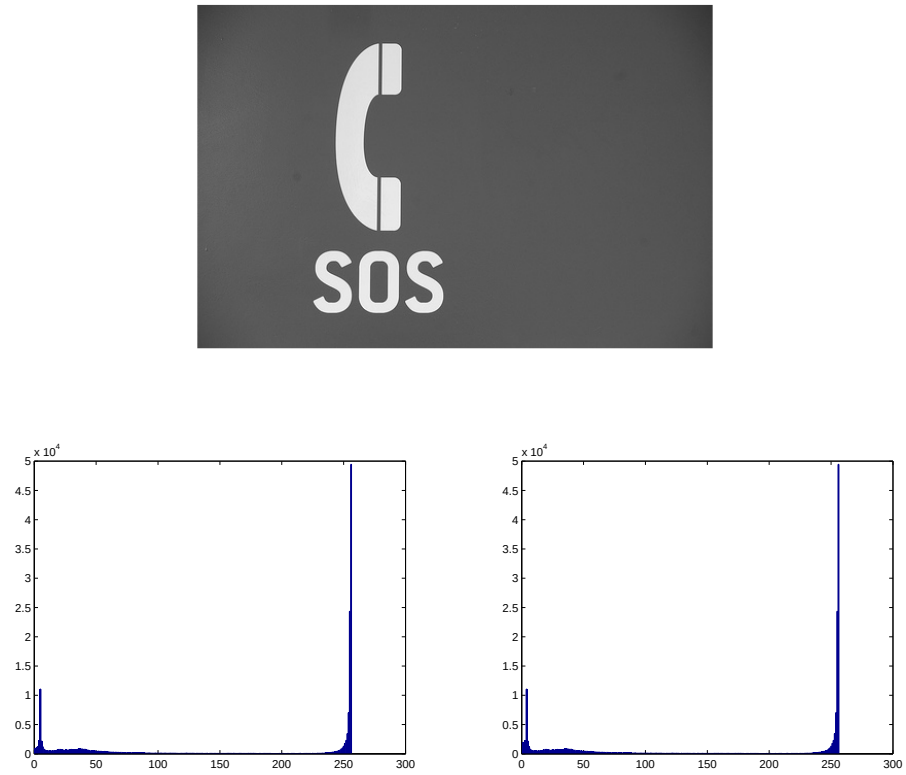


Figure 12.4: Image in $\mathcal{S}'$ which yields $\mathcal{SM} = 76$ for the case of histogram stretching. Original image (top). Histogram of the original and enhanced image (bottom).

in the histogram of low-to-medium frequency coefficients. When such an approach is adopted, the possible correlation between DCT coefficients of different image blocks and between coefficients at different frequencies is discarded thus justifying the assumption that the Defender relies only on first order statistics. As many examples of detectors of double JPEG compression are based on the analysis of the histograms of block-DCT coefficients, finding an estimation of the $\mathcal{SM}$ between the single and double compressed class is a worth investigation.

Then, in this setting, we use equations (12.1) and (12.2) to estimate the distortion that the Attacker needs to introduce at each frequency to make the DCT histograms of double compressed images equal to those of single-compressed images. We used again the images in the MIRFLICKR dataset to exemplify the above ideas. In our experiments, we considered the following pairs of quality factors:

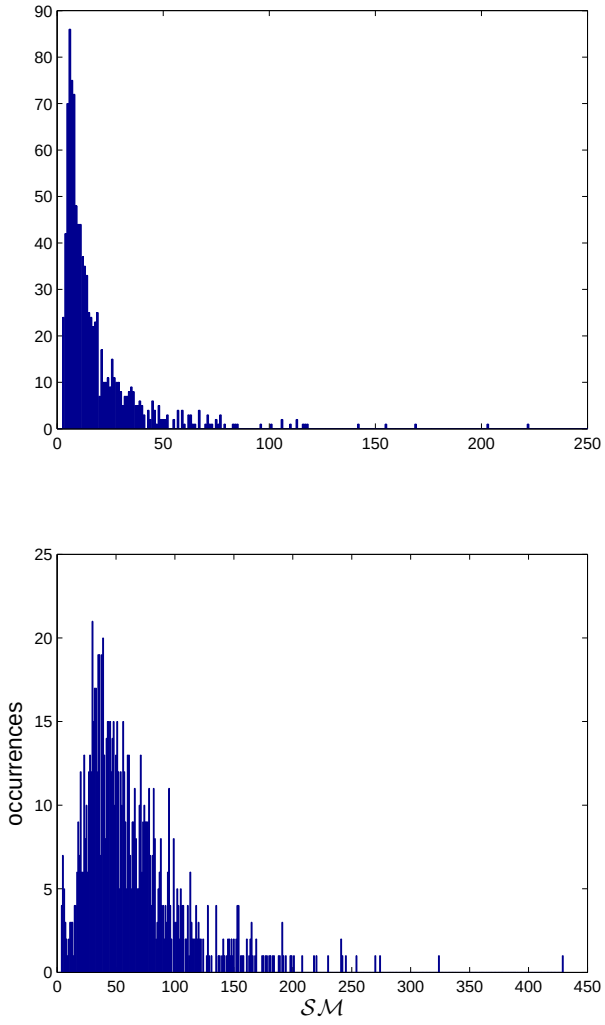


Figure 12.5: Distribution of the $\mathcal{SM}$ with the L_∞ distance at the DCT frequencies (1,1), above, and (2,1), below. The plot refers to the case of double JPEG compression with first and second quality factor 85 and 95 respectively.

$(QF_1, QF_2) \in \{(65, 85), (70, 85), (75, 90), (85, 95)\}$. For any given pair (QF_1, QF_2), the images are compressed once with QF_2 to build set $\mathcal{S}$, and twice, first with QF_1, secondly with QF_2) to build $\mathcal{S}'$ [3]. Figure 12.5 shows the distribution of the $\mathcal{SM}$

[3] We verified experimentally that for a double compressed histogram with a given QF_2, the closest

17.362	66.375	50.09	21.094	18.188	8.083	5.523	4.35
63.284	46.516	37.525	15.613	9.202	4.313	3.773	3.434
50.624	36.206	16.076	13.91	6.389	3.968	3.18	3.117
40.784	16.182	13.776	8.755	5.094	2.751	2.81	3.109
18.177	13.358	6.138	4.133	3.39	2.518	2.412	2.555
15.774	6.193	3.897	3.909	2.841	2.535	2.233	2.333
5.451	3.579	2.85	2.591	2.452	2.216	2.086	2.13
3.792	2.805	2.394	2.397	2.245	2.219	2.128	2.029

Table 12.2: Average $\mathcal{SM}_{L_\infty}$ between the set of never processed and double compressed images in the DCT domain. The images in $\mathcal{S}'$ are double compressed with quality factors 85 and 95 respectively.

across the images in $\mathcal{S}'$ for the case $(85, 95)$ for two DCT frequencies when the L_∞ norm is adopted. Table 12.2 and 12.3 show the average $\mathcal{SM}$ for all the 64 DCT frequencies for the pairs of quality factors $(85, 95)$ and $(65, 85)$ when the L_∞ distance is considered (the values taken by $\mathcal{SM}$ with the other pairs of quality factors are intermediate between these two), whereas Table 12.4 and 12.5 report the 95th percentile of the values assumed by the $\mathcal{SM}$ for the same pair of quality factors.

Quite expectedly, in both cases, the $\mathcal{SM}$ is lower at high frequencies, because of the lower range extent of the distribution of the DCT coefficients, which becomes very picked around 0. Notice that the values of the $\mathcal{SM}$ are smaller for the case $(65, 85)$. This is due to the fact that the value of QF_2 is smaller and then the quantized histograms take values on smaller supports. Finally, the average values of the $\mathcal{SM}$ when the L_2^2 is adopted are reported in Tables 12.6–12.7 for two pairs of quality factors. Even in this case, the values taken by the $\mathcal{SM}$ are larger at low frequencies and pretty small at the very high frequencies.

We remark that, in this case, the values of the $\mathcal{SM}$ refer to the mapping among DCT histograms; quantifying the visual distortion entailed in the spatial domain is not easy and needs to account for the peculiarity of the HVS.

To conclude, we notice that, as for the case of contrast enhancement detection, deceiving practical detectors of double JPEG compression may be easier in practice (see Chapter 11). On the other hand, i.e., from the analyst's perspective, this means that the forensic detection (based on first order statistics) has rooms for improvements, as attacks introducing a distortion (significantly) less than the $\mathcal{SM}$ should be detected in principle.

single compressed one is the one corresponding to the same original image quantized with QF_2.

7.426	22.555	16.905	8.535	5.411	3.026	2.194	1.853
16.053	11.742	9.478	5.259	3.608	2.043	1.662	1.583
12.721	9.025	6.508	4.166	2.424	1.919	1.528	1.471
10.228	6.629	4.073	3.043	2.089	1.463	1.338	1.443
7.485	3.943	2.62	1.829	1.632	1.223	1.161	1.244
4.744	2.571	1.763	1.641	1.407	1.213	1.116	1.114
2.235	1.58	1.421	1.357	1.174	1.089	1.044	1.057
1.639	1.329	1.205	1.202	1.101	1.084	1.027	1.031

Table 12.3: Average $\mathcal{SM}_{L_\infty}$ between the set of never processed and double compressed images in the DCT domain with $(QF_1, QF_2) = (65, 85)$.

56	154	126	54	50	23	15	11
152	120	92	40	25	10	10	8
124	83	40	38	17	10	7	7
105	41	35	22	14	6	6	7
50	34	15	10	8	5	5	5
43	16	9	10	6	5	4	4
13	8	6	5	5	4	3	4
9	6	4	4	4	4	3	3

Table 12.4: 95th percentile of the values taken by the $\mathcal{SM}_{L_\infty}$ at the various frequencies when $(QF_1, QF_2) = (85, 95)$.

22	51	42	22	14	8	5	4
38	30	23	13	9	4	3	3
31	21	16	11	6	4	3	3
26	17	10	7	5	2	2	3
20	10	6	4	3	2	2	2
12	6	3	3	2	2	2	2
5	3	2	2	2	2	2	2
3	2	2	2	2	2	2	2

Table 12.5: 95th percentile of the values taken by the $\mathcal{SM}_{L_\infty}$ at the various frequencies when $(QF_1, QF_2) = (65, 85)$.

0.56	0.62	0.54	0.44	0.59	0.52	0.43	0.34
1.12	1.04	0.98	0.53	0.62	0.38	0.36	0.38
1.03	1.1	0.43	0.59	0.51	0.37	0.31	0.38
0.99	0.43	0.58	0.53	0.44	0.35	0.34	0.32
0.43	0.59	0.58	0.45	0.54	0.28	0.29	0.36
0.60	0.57	0.45	0.4	0.36	0.30	0.24	0.31
0.45	0.39	0.45	0.33	0.29	0.22	0.20	0.21
0.33	0.38	0.27	0.26	0.23	0.23	0.22	0.18

Table 12.6: Average $\mathcal{SM}_{L_2^2}$ between the set of never processed and double compressed images in the DCT domain with $(QF_1, QF_2) = (85, 95)$.

0.51	0.61	0.43	0.36	0.34	0.24	0.21	0.16
0.45	0.40	0.38	0.31	0.28	0.21	0.18	0.11
0.41	0.38	0.34	0.33	0.24	0.19	0.15	0.11
0.41	0.36	0.33	0.24	0.22	0.14	0.11	0.13
0.46	0.33	0.28	0.16	0.16	0.11	0.1	0.11
0.34	0.29	0.16	0.18	0.13	0.1	0.08	0.06
0.23	0.18	0.15	0.13	0.11	0.07	0.062	0.063
0.15	0.1	0.10	0.09	0.07	0.07	0.06	0.05

Table 12.7: Average $\mathcal{SM}_{L_2^2}$ between the set of never processed and double compressed images in the DCT domain with $(QF_1, QF_2) = (65, 85)$.

Chapter 13

Conclusion

In this thesis we provided a general theoretical framework for the study of the Adversarial Binary Decision. With specific reference to the multimedia forensic field, we also put some of the theoretical findings into practice. In this chapter, we summarise the main contributions of our work and outline some possible paths for future research.

13.1 Summary

When we began our research activity, a multitude of techniques had been developed, in different security-oriented areas, to address different security threats. In hindsight, in all the research fields, the problems addressed were often the same in disguise, but because of the lack of a unifying view, some similar solutions were often re-invented many times and basic concepts misunderstood.

The need for a general theory which allows to retrieve a global view of the problems had just been claimed in [6], where a unifying framework for *Adversarial Signal Processing* (Adv-SP) was proposed.

Motivated by this need, in the first part of the thesis we have studied one of the most prominent problems in adversarial signal processing, that is, binary detection or Hypothesis Testing. As a first attempt in this sense, we have introduced a general framework to analyze the achievable performance of binary decision in an adversarial setting, i.e., in the presence of an adversary with the explicit goal of degrading the performance of the test. We did so by casting the detection problem into a game-theoretic framework. In this way, in fact, we have been able to define rigorously the goals and constraints of the two contenders, namely the Defender and the Attacker. More specifically, we addressed several versions of the binary decision game, depending on: i) the specific decision setup: decision based on one single observation or on multiple observations (decision fusion setup); ii) the knowledge of the system available to the Defender and the Attacker: full knowledge of the statistical model of the system under the two hypotheses or partial knowledge of the statistics

achieved by means of training data; iii) the possibility for the Attacker to interfere with the learning phase by corrupting the training data on which the Defender bases the decision.

From a more technical point of view, for the various versions of the *DT* game, we derived the asymptotic equilibrium point and analyzed the achievable payoff at the equilibrium. The analysis of the limiting performance of the various games allowed us to summarize in a single quantity, named *Security Margin*, the distinguishability of two information sources under adversarial conditions; the Security Margin provides a measure of the 'vulnerability' of a system to the attacks, quantifying how difficult attacking the system is. For the adversarial settings in which the training data are also corrupted, we derived another interesting parameter, i.e., the *blinding corruption percentage*, which, together with the Security Margin, characterizes the distinguishability of two sources.

In addition to shedding a new light on the achievable performance of binary detection in an adversarial environment, the analysis we carried out has the merit to show the potentiality of the use of game-theoretic concepts coupled with typical tools of information theory and statistics; essentially, the method of types.

In the second part of the thesis, we applied the theoretical findings to some practical problems in a real scenario, namely Image Forensics. Due to the lack of the statistical models, we considered the theoretical tools developed under partial knowledge of the statistics (i.e., knowledge based on training samples), which could be easily adapted to model the real image forensic scenarios based on a data-driven approach. By playing the role of the Attacker, leveraging on the theoretical results, we were able to devise universal counter-forensic techniques for both spatial and transformed domain forensic methods. The performance of these universal techniques has been validated against targeted state-of-the-art detectors in two specific application scenarios, namely contrast enhancement and multiple JPEG compression detection.

Overall, the analysis provided in the second part of the thesis contributes to fill the gap between the simplicity of theoretical models and the complexity of real life applications.

13.2 Open issues

With the work of this thesis, we laid the basis for the study of Adversarial Signal Processing, by addressing several variants of the binary detection problem. The study of signal processing in adversarial setup, though, is an open research field and several directions for future research can be pointed out.

From a more focused perspective, relaxing the assumptions behind the theoretical

analysis would represent a significant contribution. As discussed in the final section of Chapter 3, efforts can be made to remove the memoryless assumption for the sources, by considering more realistic models, e.g., Markov sources. 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 to a wider variety of practical applications where, because of the inherent dependence among the observation samples, looking at higher order statistics helps in making a correct decision. We also mention the opportunity of extending the analysis to the case of continuous sources, which is actually an ongoing work of our research. While the general ideas would remain the same, passing from discrete to continuous sources is not a trivial step, since our analysis relies heavily on the method of types, whose extension to continuous sources, though possible, comes with a number of additional difficulties. Finally, a main characteristic of our analysis is its asymptotic nature; the strategic interaction between Defender and Attacker for finite n would be also worth studying, since it would better fit with the practical requirements in application scenarios.

More in general, adversarial *classification* or *multiple hypothesis testing* is another interesting problem which is worth studying under a unified framework. The extension of our analysis to this case is then an interesting research directions, which would extend the applicability of the theory to a large number of practical applications where the detector must distinguish among different classes including biometric recognition, fingerprinting, multimedia forensics (multiple-camera identification, multiple JPEG compression,....), and many others.

We expect that in the near future a general framework for Adv-SP will be developed, by combining elements of game, detection, machine learning, optimization and complexity theories. We envisage that Adv-SP will become a stimulating and challenging field whose developments will find a significant number of applications.

Bibliography

[1] McAfee.

[2] R. Wortley and S. Smallbone, "Child pornography on the internet." [Online]. Available: http://www.popcenter.org/problems/child_pornography/print/

[3] [Online]. Available: https://www.fbi.gov/investigate/cyber

[4] I. J. Cox, J. Kilian, T. Leighton, and T. Shamoon, "A secure, robust watermark for multimedia," in *International Workshop on Information Hiding*. Springer, 1996, pp. 185–206.

[5] A. K. Jain, A. Ross, and S. Prabhakar, "An introduction to biometric recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 4–20, 2004.

[6] 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*, Vancouver, Canada, 26-31 May 2013, pp. 8682–8686.

[7] 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.

[8] L. Pérez-Freire, P. Comesana, J. R. Troncoso-Pastoriza, and F. Pérez-González, "Watermarking security: a survey," in *Transactions on Data Hiding and Multimedia Security I*. Springer, 2006, pp. 41–72.

[9] C. Cachin, "An information-theoretic model for steganography," in *IH98, Second International Workshop on Information Hiding*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 1998, vol. 6958, pp. 306–318.

[10] N. J. Hopper, J. Langford, and L. Von Ahn, "Provably secure steganography," in *Annual International Cryptology Conference*. Springer, 2002, pp. 77–92.

[11] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *Proceedings of EUSIPCO'05, European Signal Processing Conference*, 2005, pp. 469–472.

[12] J. Deng, R. Han, and S. Mishra, "Countermeasures against traffic analysis attacks in wireless sensor networks," in *First International Conference on Security and Privacy for Emerging Areas in Communications Networks (SECURECOMM'05)*. IEEE, 2005, pp. 113–126.

[13] K. M. C. Tan, K. S. Killourhy, and R. A. Maxion, "Undermining an anomaly-based intrusion detection system using common exploits," in *Recent Advances in Intrusion Detection (RAID)*, Zurich, Switzerland, October 16-18, 2002, pp. 54–73.

[14] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proceedings of the 24th ACM Symposium on Applied Computing*, Honolulu, Hawai, USA, 9-12 March, 2009.

[15] W. Wang, H. Li, Y. Sun, and Z. Han, "Securing collaborative spectrum sensing against untrustworthy secondary users in cognitive radio networks," *EURASIP Journal on Advances in Signal Processing*, vol. 2010, p. 4, 2010.

[16] T.-T. Do, E. Kijak, T. Furon, and L. Amsaleg, "Deluding image recognition in SIFT-based CBIR systems," in *Proceedings of the 2nd ACM Workshop on Multimedia in Forensics, Security and Intelligence*. ACM, 2010, pp. 7–12.

[17] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in *Proceedings of the 2006 ACM Symposium on Information, Computer and Communications Security*, ser. ASIACCS '06. New York, NY, USA: ACM, 2006, pp. 16–25. [Online]. Available: http://doi.acm.org/10.1145/1128817.1128824

[18] P. Laskov and R. Lippmann, "Machine learning in adversarial environments," *Machine Learning*, vol. 81, no. 2, pp. 115–119, 2010.

[19] B. Biggio, G. Fumera, and F. Roli, "Security evaluation of pattern classifiers under attack," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, no. 4, pp. 984–996, 2014.

[20] G. L. Wittel and S. F. Wu, "On attacking statistical spam filters," in *CEAS*, 2004.

[21] P. Fogla, M. I. Sharif, R. Perdisci, O. M. Kolesnikov, and W. Lee, "Polymorphic blending attacks." in *USENIX Security*, 2006.

[22] D. Wagner and P. Soto, "Mimicry attacks on host-based intrusion detection systems," in *ACM Conference on Computer and Communications Security*, Washington, DC, USA, November 2002, pp. 255–264.

[23] N. Dalvi, P. Domingos, P. Mausam, S. Sanghai, and D. Verma, "Adversarial classification," in *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 2004, pp. 99–108.

[24] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*. ACM, 2011, pp. 43–58.

[25] N. Bao, O. P. Kreidl, and J. Musacchio, "A network security classification game," in *International Conference on Game Theory for Networks*. Springer, 2011, pp. 265–280.

[26] L. Dritsoula, P. Loiseau, and J. Musacchio, "A game-theoretical approach for finding optimal strategies in an intruder classification game," in *51st IEEE Conference on Decision and Control (CDC)*. IEEE, 2012, pp. 7744–7751.

[27] L. Dritsoula, P. Loiseau, and J. Musacchio, "Computing the nash equilibria of intruder classification games," in *International Conference on Decision and Game Theory for Security*. Springer, 2012, pp. 78–97.

[28] M. Brückner and T. Scheffer, "Nash equilibria of static prediction games," in *Advances in Neural Information Processing Systems*, 2009, pp. 171–179.

[29] N. Hopper, L. von Ahn, and J. Langford, "Provably secure steganography," *IEEE Transactions on Computers*, vol. 58, no. 5, pp. 662–676, 2009.

[30] J. Zöllner, H. Federrath, H. Klimant, A. Pfitzmann, R. Piotraschke, A. Westfeld, G. Wicke, and G. Wolf, "Modeling the security of steganographic systems," in *International Workshop on Information Hiding*. Springer, 1998, pp. 344–354.

[31] R. Böhme and A. Westfeld, "Breaking cauchy model-based JPEG steganography with first order statistics," in *European Symposium on Research in Computer Security*. Springer, 2004, pp. 125–140.

[32] J. M. Ettinger, "Steganalysis and game equilibria," in *International Workshop on Information Hiding*. Springer, 1998, pp. 319–328.

[33] A. D. Ker, "Batch steganography and the threshold game," in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 650 504–650 504.

[34] E. Franz, "Steganography preserving statistical properties," in *International Workshop on Information Hiding*. Springer, 2002, pp. 278–294.

[35] P. Schöttle and R. Böhme, "A game-theoretic approach to content-adaptive steganography," in *International Workshop on Information Hiding*. Springer, 2012, pp. 125–141.

[36] B. Johnson, P. Schöttle, and R. Böhme, "Where to hide the bits?" in *International Conference on Decision and Game Theory for Security*. Springer, 2012, pp. 1–17.

[37] T. Denemark and J. Fridrich, "Detection of content adaptive lsb matching: A game theory approach," in *IS&T/SPIE Electronic Imaging*. International Society for Optics and Photonics, 2014, pp. 902 804–902 804.

[38] A. S. Cohen and A. Lapidoth, "The gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 48, no. 6, pp. 1639–1667, June 2002.

[39] P. Moulin and M. K. Mihcak, "The parallel-Gaussian watermarking game," *IEEE Transactions on Information Theory*, vol. 50, no. 2, pp. 272–289, February 2004.

[40] A. Somekh-Baruch and N. Merhav, "On the capacity game of public watermarking systems," *IEEE Transactions on Information Theory*, vol. 50, no. 3, pp. 511–524, March 2004.

[41] 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, March 2013.

[42] 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*, Coventry, UK, 6-7 September 2012, pp. 97–104.

[43] M. Barni, M. Fontani, and B. Tondi, "Universal counterforensics of multiple compressed JPEG images," in *International Workshop on Digital Watermarking*. Springer, 2014, pp. 31–46.

[44] 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.

[45] J. S. Merlevede and T. Holvoet, "Game theory and security: Recent history and future directions," in *International Conference on Decision and Game Theory for Security*. Springer, 2015, pp. 334–345.

[46] A. Nochenson and C. L. Heimann, "Simulation and game-theoretic analysis of an attacker-defender game," in *International Conference on Decision and Game Theory for Security*. Springer, 2012, pp. 138–151.

[47] A. Clark, K. Sun, L. Bushnell, and R. Poovendran, "A game-theoretic approach to ip address randomization in decoy-based cyber defense," in *International Conference on Decision and Game Theory for Security*. Springer, 2015, pp. 3–21.

[48] B. Johnson, J. Grossklags, N. Christin, and J. Chuang, "Are security experts useful? Bayesian Nash equilibria for network security games with limited information," in *European Symposium on Research in Computer Security*. Springer, 2010, pp. 588–606.

[49] S. Rass, S. König, and S. Schauer, "Uncertainty in games: Using probability-distributions as payoffs," in *International Conference on Decision and Game Theory for Security*. Springer, 2015, pp. 346–357.

[50] A. Bensoussan, M. Kantarcioglu, and S. C. Hoe, "A game-theoretical approach for finding optimal strategies in a botnet defense model," in *International Conference on Decision and Game Theory for Security*. Springer, 2010, pp. 135–148.

[51] E. Delp, N. Memon, and M. Wu, "Special issue on digital forensics," *IEEE Signal Processing Magazine*, vol. 26, no. 2, March 2009.

[52] M. Martinez-Diaz, J. Fierrez-Aguilar, F. Alonso-Fernandez, J. Ortega-Garcia, and J. A. Siguenza, "Hill-climbing and brute-force attacks on biometric systems: A case study in match-on-card fingerprint verification," in *Proceedings of 40th Annual IEEE International Carnahan Conferences Security Technology*, 2006, pp. 151–159.

[53] P. Moulin, A. Ivanovic, and A. Ivanovic, "Game-theoretic analysis of watermark detection." in *ICIP (3)*, 2001, pp. 975–978.

[54] 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 Trans. Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[55] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications*. Cambridge University Press, 2009.

[56] I. J. Cox and J.-P. M. Linnartz, "Public watermarks and resistance to tampering," in *International Conference on Image Processing (ICIP97)*, 1997, pp. 26–29.

[57] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "The return of the sensitivity attack," in *International Workshop on Digitial Watermarking*, Siena, Italy, September 2005, pp. 260–274.

[58] P. Comesaña, L. Pérez-Freire, and F. Pérez-González, "Blind Newton sensitivity attack," *IEE Proceedings on Information Security*, vol. 153, no. 3, pp. 115–125, September 2006.

[59] M. Martínez-Díaz, J. Fiérrez-Aguilar, F. Alonso-Fernández, J. Ortega-García, and J. A. Sigüenza, "Hill-climbing and brute-force attacks on biometric systems: A case study in match-on-card fingerprint verification," in *Annual IEEE International Carnahan Conferences Security Technology*, Lexington, KY, USA, October 2006, pp. 151–159.

[60] J. Galbally, J. Fiérrez, J. Ortega-García, C. McCool, and S. Marcel, "Hill-climbing attack to an eigenface-based face verification system," in *IEEE International Conference on Biometrics, Identity and Security*, Tampa, FL, USA, September 2009, pp. 1–6.

[61] E. Maiorana, G. E. Hine, and P. Campisi, "Hill-climbing attacks on multibiometrics recognition systems," *IEEE Transactions on Information Forensics and Security*, vol. 10, no. 5, pp. 900–915, May 2015.

[62] P. Fogla and W. Lee, "Evading network anomaly detection systems: Formal reasoning and practical techniques," in *ACM Conference on Computer and Communications Security*, Alexandria, VA, USA, October-November 2006, pp. 59–68.

[63] D. Lowd and C. Meek, "Adversarial learning," in *ACM SIGKDD International Conference on Knowledge Discovery in Data Mining*, Chicago, IL, USA, August 2005, pp. 641–647.

[64] M. Backes and C. Cachin, "Public-key steganography with active attacks," in *Theory of Cryptography Conference*. Springer, 2005, pp. 210–226.

[65] R. Venkatesan and M. H. Jakubowski, "Randomized detection for spread-spectrum watermarking: Defending against sensitivity and other attacks." in *ICASSP (2)*, 2005, pp. 9–12.

[66] K. Wang, J. J. Parekh, and S. J. Stolfo, "Anagram: A content anomaly detector resistant to mimicry attack," in *International Workshop on Recent Advances in Intrusion Detection*. Springer, 2006, pp. 226–248.

[67] T. Furon and P. Duhamel, "An asymmetric watermarking method," *IEEE Transactions on Signal Processing*, vol. 51, no. 4, pp. 981–995, 2003.

[68] A. Globerson and S. Roweis, "Nightmare at test time: Robust learning by feature deletion," in *International Conference on Machine Learning*, Pittsburgh, PA, USA, June 2006, pp. 353–360.

[69] B. Biggio, I. Corona, Z.-M. He, P. P. K. Chan, G. Giacinto, D. S. Yeung, and F. Roli, "One-and-a-half-class multiple classifier systems for secure learning against evasion attacks at test time," in *International Workshop on Multiple Classifier Systems*, Günzburg, Germany, June-July 2015, pp. 168–180.

[70] G. F. Cretu, A. Stavrou, M. E. Locasto, S. J. Stolfo, and A. D. Keromytis, "Casting out demons: Sanitizing training data for anomaly sensors," in *IEEE Symposium on Security and Privacy*, Oakland, CA, USA, May 2008, pp. 81–95.

[71] M. Barni, P. Comesaña-Alfaro, F. Pérez-González, and B. Tondi, "Are you threatening me?: Towards smart detectors in watermarking," in *SPIE Electronic Imaging*, San Fransciso, CA, USA, February 2014.

[72] B. Tondi, P. Comesaña-Alfaro, F. Pérez-González, and M. Barni, "On the effectiveness of meta-detection for countering oracle attacks in watermarking," in *IEEE International Workshop on Information Forensics and Security*, Rome, Italy, November 2015, pp. 1 – 6.

[73] Y. Yang, Y. L. Sun, S. Kay, and Q. Yang, "Defending online reputation systems against collaborative unfair raters through signal modeling and trust," in *Proceedings of the 2009 ACM symposium on Applied Computing*. ACM, 2009, pp. 1308–1315.

[74] H. Jin, J. Lotspiech, and N. Megiddo, "Efficient coalition detection in traitor tracing," in *IFIP International Information Security Conference*. Springer, 2008, pp. 365–380.

[75] R. Feldman, "Techniques and applications for sentiment analysis," *Communications of the ACM*, vol. 56, no. 4, pp. 82–89, 2013.

[76] R. Avenhaus, B. Von Stengel, and S. Zamir, "Inspection games," *Handbook of Game Theory with Economic Applications*, vol. 3, pp. 1947–1987, 2002.

[77] J. Von Neumann and O. Morgenstern, *Theory of games and economic behavior*. Princeton University Press, 2007.

[78] M. J. Osborne, *An introduction to game theory*. Oxford University Press New York, 2004, vol. 3, no. 3.

[79] J. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[80] M. J. Osborne and A. Rubinstein, *A Course in Game Theory*. MIT Press, 1994.

[81] J. v. Neumann, "Zur theorie der gesellschaftsspiele," *Mathematische Annalen*, vol. 100, pp. 295–320, 1928. [Online]. Available: http://eudml.org/doc/159291

[82] 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, November 2007.

[83] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.

[84] P. Weirich, *Equilibrium and rationality: game theory revised by decision rules.* Cambridge University Press, 2007.

[85] E. L. Lehmann and J. P. Romano, *Testing statistical hypotheses.* Springer Science & Business Media, 2006.

[86] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory.* Prentice Hall, 1998.

[87] J. Munkres, *Topology*, ser. Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. [Online]. Available: https://books.google.it/books?id=XjoZAQAAIAAJ

[88] J. Henrikson, "Completeness and total boundedness of the Hausdorff metric," *MIT Undergraduate Journal of Mathematics*, vol. 1, pp. 69–80, 1999.

[89] I. Csiszar, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, October 1998.

[90] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* New York: Wiley Interscience, 1991.

[91] 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, January 2008.

[92] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, April 1965.

[93] C. Villani, *Optimal Transport: Old and New.* Berlin: Springer-Verlag, 2009.

[94] G. Monge, *Mémoire sur la théorie des déblais et des remblais.* De l'Imprimerie Royale, 1781.

[95] R. Ansari, N. Memon, and E. Ceran, "Near-lossless image compression techniques," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 486–494, 1998. [Online]. Available: http://dx.doi.org/10.1117/1.482591

[96] 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.

[97] I. Csiszar and P. C. Shields, "Redundancy rates for renewal and other processes," *IEEE Transactions on Information Theory*, vol. 42, no. 6, pp. 2065–2072, November 1996.

[98] M.Chen, J. Fridrich, M. Goljan, and J. Lukas, "Determining image origin and integrity using sensor noise," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 1, pp. 74–90, March 2008.

[99] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, March 1989.

[100] M. Kendall and S. Stuart, *The Advanced Theory of Statistics, vol. 2, 4th edition.* New York: MacMillan, 1979.

[101] I. Csiszár and P. Shields, *Information Theory and Statistics: a Tutorial.* Now Publishers Inc., 2004.

[102] W. A. Sutherland, *Introduction to metric and topological spaces.* Oxford University Press, 1975.

[103] M. Goljan, J. Fridrich, and M. Chen, "Sensor noise camera identification: countering counter forensics," in *SPIE Conference on Media Forensics and Security, San Jose, CA*, 2010.

[104] 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, November 2000.

[105] S. T. Rachev, *Mass Transportation Problems: Volume I: Theory.* Springer, 1998, vol. 1.

[106] C. Villani, *Topics in Optimal Transportation.* Graduate Studies in Mathematics Series: American Mathematical Society, 2003, vol. 58.

[107] S. T. Rachev, "The Monge-Kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[108] 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, 2001, pp. 251–256 vol.2.

[109] O. Pele and M. Werman, "Fast and robust Earth Mover's distances," in *Proceedings ICCV'09, 12th IEEE International Conference on Computer Vision*, 2009, pp. 460–467.

[110] F. L. Hitchcock, "The distribution of a product from several sources to numerous localities," *Journal of Mathematical Physics*, vol. 20, pp. 224–230.

[111] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications.* Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1993.

[112] V. Chvatal, "Linear programming," *A Series of Books in the Mathematical Sciences, New York: Freeman, 1983*, vol. 1, 1983.

[113] A. Hoffman, "On simple linear programming problems," in *Proceedings of Symposia in Pure Mathematics*, vol. 7. World Scientific, 1963, pp. 317–327.

[114] 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.

[115] J. B. Orlin, "A faster strongly polynomial minimum cost flow algorithm," *Operations research*, vol. 41, no. 2, pp. 338–350, 1993.

[116] A. C. Williams, "A treatment of transportation problems by decomposition," *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 1, pp. pp. 35–48.

[117] A. Irpino and E. Romano, "Optimal histogram representation of large data sets: Fisher vs piecewise linear approximation." in *EGC*, ser. Revue des Nouvelles Technologies de l'Information, M. Noirhomme-Fraiture and G. Venturini, Eds., vol. RNTI-E-9. Cepadues-Editions, 2007, pp. 99–110.

[118] K. Košmelj and L. Billard, "Mallows' L_2 distance in some multivariate methods and its application to histogram-type data," *Metodoloski Zvezki*, vol. 9, no. 2, pp. 107–118, 2012.

[119] F. Willems, T. Kalker, J. Goseling, and J.-P. Linnartz, "On the capacity of a biometrical identification system," in *IEEE International Symposium on Information Theory*, 2003, pp. 82–82.

[120] P. K. Varshney, *Distributed Detection and Data Fusion*. Springer-Verlag, 1997.

[121] J.-F. Chamberland and V. V. Veeravalli, "Decentralized detection in sensor networks," *IEEE Transactions on Signal Processing*, vol. 51, no. 2, pp. 407–416, February 2003.

[122] I. F. Akyildiz, B. F. Lo, and R. Balakrishnan, "Cooperative spectrum sensing in cognitive radio networks: A survey," *Physical Communication*, vol. 2011, no. 4, pp. 40–62, 2011.

[123] M. Fontani, T. Bianchi, A. De Rosa, A. Piva, and M. Barni, "A framework for decision fusion in image forensics based on Dempster-Shafer theory of evidence," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 4, pp. 593–607, April 2013.

[124] B. Kailkhura, S. Brahma, Y. S. Han, and P. K. Varshney, "Optimal distributed detection in the presence of byzantines," in *ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, Vancouver, Canada, 27-31 May 2013.

[125] R. G. Gallager, *Information theory and reliable communication*. Springer, 1968, vol. 2.

[126] 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, February 2011.

[127] Y. Liu and Y. L. Sun, "Anomaly detection in feedback-based reputation systems through temporal and correlation analysis," in *Proceedings of 2nd IEEE International Conference on Social Computing*, August 2010.

[128] M. Goljan, J. Fridrich, and M. Chen, "Defending against fingerprint-copy attack in sensor-based camera identification," *IEEE Transactions on Information Forensics and Security*, vol. 6, no. 1, pp. 227–236, March 2011.

[129] T. Gloe, M. Kirchner, A. Winkler, and R. Bohme, "Can we trust digital image forensics ?" in *ACM Multimedia 2007, Augsburg, Germany*, September 2007, pp. 78–86.

[130] M. Barni and F. Bartolini, *Watermarking systems engineering: enabling digital assets security and other applications.* CRC Press, 2004.

[131] N. Khanna, A. K. Mikkilineni, G. T. Chiu, J. P. Allebach, and E. J. Delp, "Forensic classification of imaging sensor types," in *Electronic Imaging 2007.* International Society for Optics and Photonics, 2007, pp. 65 050U–65 050U.

[132] J. Fridrich, "Digital image forensics," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 26–37, 2009.

[133] H. Farid, "Image forgery detection," *IEEE Signal Processing Magazine*, vol. 26, no. 2, pp. 16–25, 2009.

[134] M. Kirchner and R. Böhme, "Tamper hiding: Defeating image forensics," in *International Workshop on Information Hiding.* Springer, 2007, pp. 326–341.

[135] G. Cao, Y. Zhao, R. Ni, and H. Tian, "Anti-forensics of contrast enhancement in digital images," in *Proceedings of MM&Sec 2010, 12th ACM Workshop on Multimedia and Security (MM&Sec '10)*, 2010.

[136] M. Stamm and K. J. R. Liu, "Blind forensics of contrast enhancement in digital images," in *Proceedings of the 15th IEEE International Conference on Image Processing*, 2008, pp. 3112–3115.

[137] M. Kirchner and R. Bohme, "Hiding traces of resampling in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 582–292, December 2008.

[138] M. C. Stamm and K. J. R. Liu, "Wavelet-based image compression anti-forensics," in *Proceedings of ICIP 2010, IEEE International Conference on Image Processing*, 2010, pp. 1737–1740.

[139] M. C. Stamm, S. K. Tjoa, W. S. Lin, and K. J. R. Liu, "Anti-forensics of JPEG compression," in *Proceedings of ICASSP 2010, IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2010, pp. 1694–1697.

[140] M. C. Stamm, S. Tjoa, W. Lin, and K. Liu, "Undetectable image tampering through JPEG compression anti-forensics," in *Proceedings of ICIP 2010, IEEE International Conference on Image Processing*, 2010, pp. 2109 –2112.

[141] M. C. Stamm and K. Liu, "Anti-forensics for frame deletion/addition in MPEG video," in *Proceedings of ICASSP 2011, IEEE International Conference on Acoustics, Speech and Signal Processing*, 2011, pp. 1876 –1879.

[142] W. Wang and H. Farid, "Exposing digital forgeries in video by detecting double MPEG compression," in *Proceedings of MM&Sec 2006, 8th ACM workshop on Multimedia & Security*, 2006, pp. 37–47.

[143] G. Valenzise, V. Nobile, M. Tagliasacchi, and S. Tubaro, "Countering JPEG anti-forensics," in *Proceedings of ICIP 2011, IEEE International Conference on Image Processing*, 2011, pp. 1949–1952.

[144] S. Lai and R. Bhme, "Countering counter-forensics: The case of JPEG compression," in *Information Hiding (13^{th} International Conference)*, ser. Lecture Notes in Computer Science, T. Filler, T. Pevny, S. Craver, and A. Ker, Eds. Berlin: Springer, 2011, vol. 6958, pp. 285–298, publication status: Published.

[145] M. C. Stamm, W. S. Lin, and K. J. R. Liu, "Forensics vs anti-forensics: a decision and game theoretic framework," in *ICASSP 2012, IEEE International Conference on Acoustics, Speech and Signal Processing*, Kyoto, Japan, 25-30 March 2012.

[146] M. Barni, "A game theoretic approach to source identification with known statistics," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2012, pp. 1745–1748.

[147] S. W. Zucker and D. Terzopoulos, "Finding structure in co-occurrence matrices for texture analysis," *Computer Graphics and Image Processing*, vol. 12, no. 3, pp. 286–308, 1980.

[148] J. Fowler and Q. Du, "Anomaly detection and reconstruction from random projections," *IEEE Transactions on Image Processing*, vol. 21, no. 1, pp. 184–195, Jan 2012.

[149] 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, March 2014.

[150] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *Computational Photography (ICCP), 2012 IEEE International Conference on*, April 2012, pp. 1–10.

[151] X. Pan, X. Zhang, and S. Lyu, "Detecting splicing in digital audios using local noise level estimation," in *Acoustics, Speech and Signal Processing (ICASSP), 2012 IEEE International Conference on*, March 2012, pp. 1841–1844.

[152] T. Pevny and J. Fridrich, "Detection of double-compression in JPEG images for applications in steganography," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 2, pp. 247–258, June 2008.

[153] B. Li, Y. Shi, and J. Huang, "Detecting doubly compressed JPEG images by using mode based first digit features," in *Proceedings of MMSP 2008, IEEE Workshop on Multimedia Signal Processing,*, Oct 2008, pp. 730–735.

[154] D. Fu, Y. Q. Shi, and W. Su, "A generalized benford's law for JPEG coefficients and its applications in image forensics," in *SPIE Conference on Security, Steganography, and Watermarking of Multimedia Contents*, E. J. Delp and P. W. Wong, Eds., vol. 6505, 2007.

[155] S. Milani, M. Tagliasacchi, and S. Tubaro, "Discriminating multiple JPEG compression using first digit features," in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 2012, pp. 2253–2256.

[156] R. Cogranne, C. Zitzmann, L. Fillatre, F. Retraint, I. Nikiforov, and P. Cornu, "A cover image model for reliable steganalysis," in *International Workshop on Information Hiding*. Springer, 2011, pp. 178–192.

[157] A. N. Harutyunyan, N. Grigoryan, S. Voloshynovskiy, and O. J. Koval, "A new biometric identification model and the multiple hypothesis testing for arbitrarily varying objects." in *CAST Workshop - Biometrie, BIOSIG2011*, 2011, pp. 305–312.

[158] F.-W. Fu and S.-Y. Shen, "Hypothesis testing for arbitrarily varying source with exponential-type constraint," *IEEE Transactions on Information Theory*, vol. 44, no. 2, pp. 892–895, Mar 1998.

[159] M. Bussieck and A. Pruessner, "Mixed-integer nonlinear programming," *SIAG/OPT Newsletter: Views & News*, vol. 14, no. 1, pp. 19–22, 2003.

[160] M. Bussieck and S. Vigerske, "MINLP solver software," 2011.

[161] P. Bonami, M. Kilinc, J. Linderoth *et al.*, "Algorithms and software for convex mixed integer nonlinear programs," Computer Sciences Department, University of Wisconsin-Madison, Tech. Rep., 2009.

[162] D. Bertsimas and J. Tsitsiklis, *Introduction to Linear Optimization*, 1st ed. Athena Scientific, 1997.

[163] A. Wächter and L. T. Biegler, "On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming," *Mathematical Programming*, vol. 106, pp. 25–57, 2006. [Online]. Available: http://dx.doi.org/10.1007/s10107-004-0559-y

[164] O. Pele and M. Werman, "The Quadratic-Chi histogram distance family," in *Proceedings of ECCV 2010, European Conference on Computer Vision*, 2010.

[165] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[166] J. Fridrich and R. Du, "Secure steganographic methods for palette images," in *International Workshop on Information Hiding*. Springer, 1999, pp. 47–60.

[167] G. Schaefer, "An uncompressed benchmark image dataset for colour imaging," in *Proceedings of ICIP 2010, IEEE International Conference on Image Processing*, 2010, pp. 3537–3540.

[168] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of MIR '08, ACM International Conference on Multimedia Information Retrieval*. New York, NY, USA: ACM, 2008.

[169] A. C. Popescu and H. Farid, "Statistical tools for digital forensics," in *Proceedings of IH 2005, International Conference on Information Hiding*. Springer, 2005, pp. 128–147.

[170] J. Lukáš and J. Fridrich, "Estimation of primary quantization matrix in double compressed JPEG images," in *Proceedings Digital Forensic Research Workshop*, 2003, pp. 5–8.

[171] J. Fridrich, M. Goljan, and D. Hogea, "Steganalysis of JPEG images: Breaking the F5 algorithm," in *Information Hiding*. Springer, 2003, pp. 310–323.

[172] J. Fridrich, "Feature-based steganalysis for JPEG images and its implications for future design of steganographic schemes," in *International Workshop on Information Hiding*. Springer, 2004, pp. 67–81.

[173] J. Kodovskỳ and J. Fridrich, "Calibration revisited," in *Proceedings of the 11th ACM workshop on Multimedia and security*. ACM, 2009, pp. 63–74.

[174] B. Mahdian and S. Saic, "Detecting double compressed JPEG images," in *3rd International Conference on Crime Detection and Prevention*. IET, 2009, pp. 1–6.

[175] D. Fu, Y. Q. Shi, and W. Su, "A generalized Benford's law for JPEG coefficients and its applications in image forensics," in *Electronic Imaging 2007*. International Society for Optics and Photonics, 2007, pp. 65 051L–65 051L.

[176] S. Milani, M. Tagliasacchi, and S. Tubaro, "Antiforensics attacks to benford's law for the detection of double compressed images," in *Proceedings of ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2013, pp. 3053–3057.

[177] C. Pasquini, P. Comesaña-Alfaro, F. Pérez-González, and G. Boato, "Transportation-theoretic image counterforensics to first significant digit histogram forensics," in *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, May 2014, pp. 2699–2703.

[178] M. Kirchner and S. Chakraborty, "A second look at first significant digit histogram restoration," in *IEEE International Workshop on Information Forensics and Security (WIFS)*. IEEE, 2015, pp. 1–6.

[179] C. Chen, Y. Q. Shi, and W. Su, "A machine learning based scheme for double JPEG compression detection," in *Pattern Recognition, 2008. ICPR 2008. 19th International Conference on*. IEEE, 2008, pp. 1–4.

[180] S. Lai and R. Böhme, "Block convergence in repeated transform coding: JPEG-100 forensics, carbon dating, and tamper detection," in *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, 2013.

[181] F. Huang, J. Huang, and Y. Q. Shi, "Detecting double JPEG compression with the same quantization matrix," *IEEE Transactions on Information Forensics and Security*, vol. 5, no. 4, pp. 848–856, 2010.

[182] J. Yang, J. Xie, G. Zhu, S. Kwong, and Y.-Q. Shi, "An effective method for detecting double JPEG compression with the same quantization matrix," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 11, pp. 1933–1942, 2014.

[183] G. K. Wallace, "The JPEG still picture compression standard," *IEEE Transactions on Consumer Electronics*, vol. 38, no. 1, pp. xviii–xxxiv, 1992.

[184] A. B. Watson, "DCT quantization matrices visually optimized for individual images," in *Proceedings of IS&T/SPIE's Symposium on Electronic Imaging: Science and Technology*. International Society for Optics and Photonics, 1993, pp. 202–216.

[185] I. Cox, M. Miller, J. Bloom, J. Fridrich, and T. Kalker, *Digital Watermarking and Steganography*, 2nd ed. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., 2008.

[186] H. R. Wu and K. R. Rao, *Digital video image quality and perceptual coding*. CRC Press, 2005.

[187] E. Y. Lam and J. W. Goodman, "A mathematical analysis of the DCT coefficient distributions for images," *IEEE Transactions on Image Processing*, vol. 9, no. 10, pp. 1661–1666, 2000.

[188] D.-T. Dang-Nguyen, C. Pasquini, V. Conotter, and G. Boato, "RAISE: A raw images dataset for digital image forensics," in *Proceedings of the 6th ACM Multimedia Systems Conference*, ser. MMSys '15. New York, NY, USA: ACM, 2015, pp. 219–224. [Online]. Available: http://doi.acm.org/10.1145/2713168.2713194

[189] M. J. Huiskes and M. S. Lew, "The MIR Flickr retrieval evaluation," in *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. ACM, 2008, pp. 39–43.

[190] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc., 1992.

[191] S. I.N., "On the probability of large deviations of random variables," *Math. Sbornik*, vol. 42, pp. 11–44, 1957.

[192] A. Dembo and O. Zeitouni, *Large deviations techniques and applications*. Springer Science & Business Media, 2009, vol. 38.

[193] K. Kuratowski, *Topology*, ser. Topology. Academic Press, 1968, no. v. 2.

[194] G. Salinetti and R. J.-B. Wets, "On the convergence of sequences of convex sets in finite dimensions," *Siam review*, vol. 21, no. 1, pp. 18–33, 1979.

[195] D. Bertsimas and J. N. Tsitsiklis, *Introduction to linear optimization*. Athena Scientific Belmont, MA, 1997, vol. 6.

[196] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge University Press, 2004.

Appendix A

Generalization of Sanov's theorem

In this appendix, we generalize the classical Sanov's theorem, so to be able to apply it to more general sequences of sets, like the ones considered in this thesis in the computation of the payoff at the equilibrium for the DT_{ks}, DT_{tr} and $DT_{c\text{-}tr}$ games in Chapter 3, 4 and 6 respectively.

Let us consider a sequence of n i.i.d. random variables drawn according to a distribution P. We denote with $\hat{P}_n$ the empirical pmf of the sequence Let $E \subseteq \mathcal{P}$ be a set of probability distributions. Sanov's theorem [90, 191, 192] states that

$$\begin{aligned} -\inf_{Q \in int\ E} \mathcal{D}(Q||P) \leq & \liminf_{n\to\infty} \frac{1}{n} \log P(\hat{P}_n \in E) \\ \leq & \limsup_{n\to\infty} \frac{1}{n} \log P(\hat{P}_n \in E) \\ \leq & -\inf_{Q \in E} \mathcal{D}(Q||P), \end{aligned} \tag{A1}$$

where $int\ X$ denotes the interior part of the set X.

When $cl(E) = cl(int(E))$ [1], or equivalently, $E \subseteq cl(int(E))$, the left and right-hand side of (A7) coincide and we get the exact rate:

$$\lim_{n\to\infty} \frac{1}{n} \log P(\hat{P}_n \in E) = -\inf_{Q \in E} \mathcal{D}(Q||P). \tag{A2}$$

Obviously, if we define the set $E_n = E \cap \mathcal{P}_n$, we have that $P(\hat{P}_n \in E) = P(\hat{P}_n \in E_n)$ and we can rewrite Sanov's theorem accordingly,

$$\begin{aligned} -\inf_{Q \in int\ E} \mathcal{D}(Q||P) \leq & \liminf_{n\to\infty} \frac{1}{n} \log P(\hat{P}_n \in E_n) \\ \leq & \limsup_{n\to\infty} \frac{1}{n} \log P(\hat{P}_n \in E_n) \\ \leq & -\inf_{Q \in E} \mathcal{D}(Q||P), \end{aligned} \tag{A3}$$

Clearly, by construction, in Sanov's theorem we have that $cl(E) = cl(\cup_n E_n)$.

In the following, we extend Sanov's theorem to more general sequences of sets E_n for which, by letting $E = cl(\bigcup_n E_n)$, it does not necessary hold that $E_n = E \cap \mathcal{P}_n$.

[1] $cl(X)$ denotes the closure of the set X. Clearly, $cl(X) \equiv X$ if X is a closed set.

We start by introducing the notion of convergence for sequences of subsets due to Kuratowsky, which is a more general notion of convergence with respect to the one based on Hausdorff distance (see Section 3.1). Let (S, d) be a metric space. We first provide the definition of *lower closed limit* or Kuratowski limit inferior [193].

Definition 16. *A point p belongs to the lower limit $\underset{n\to\infty}{Li}\, K_n$ (or simply LiK_n) of a sequence of sets $K_1, K_2, ...$, if every neighborhood of p intersects all the K_n from a sufficiently great index n onward.*

The formula $p \in \underset{n\to\infty}{Li}\, K_n$ is equivalent to the existence of a sequence of points $\{p_n\}$ such that:[2]

$$p = \lim_{n\to\infty} p_n, \quad p_n \in K_n. \tag{A4}$$

Then,

$$\underset{n\to\infty}{Li}\, K_n = \{p \in S \text{ s.t. } \limsup_{n\to\infty} d(p, K_n) = 0\}. \tag{A5}$$

Similarly, we have the following definition of *upper closed limit* or Kuratowski limit superior [193].

Definition 17. *A point p belongs to the upper limit $\underset{n\to\infty}{Ls}\, K_n$ (or simply LsK_n) of a sequence of sets $K_1, K_2, ...$, if every neighborhood of p intersects an infinite set of the terms K_n.*

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

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

Then,

$$\underset{n\to\infty}{Li}\, K_n = \{p \in S \text{ s.t. } \liminf_{n\to\infty} d(p, K_n) = 0\}. \tag{A6}$$

It can be proved that the Kuratowski limit inferior and superior are closed sets (see [193]). Given the above, we can state the following.

Definition 18. *The sequence of sets $\{K_n\}$ is said to converge to K in the sense of Kuratowski, that is $K_n \xrightarrow{K} K$ if $LiK_n = K = LsK_n$, in which case we write $K = LimK_n$.*

It is worth noting that Kuratowski convergence is weaker than convergence in Hausdorff metric; that is, given a sequence of closed sets $\{K_n\}$, $K_n \xrightarrow{H} K$ implies $K_n \xrightarrow{K} K$ [194]. For compact metric spaces, the reverse implication also holds and the two kinds of convergence coincide.

In our arguments, we are interested in the space of probability distributions $\mathcal{P}$ defined over $\mathcal{X}$, i.e., the probability simplex in $\mathbb{R}^{|\mathcal{X}|}$. Being $\mathcal{P}$ a closed subset of $\mathbb{R}^{|\mathcal{X}|}$, $\mathcal{P}$ is complete for any metric d. Be d such that $\mathcal{P} \in \mathcal{L}(\mathbb{R}^{|\mathcal{X}|})$, that is, $\mathcal{P}$ is bounded with the metric d. The metric space $(\mathcal{P}, d)$ is a compact metric space. Accordingly, for our purposes, Kuratowski and Hausdorff convergence are equivalent.

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

[2] LiK_n is the set of the accumulation points of sequences in K_n.

Theorem 24 (Generalized Sanov's theorem)**.** *Let $\{E^{(n)}\}$ be a sequence of sets of probability distributions in $\mathcal{P}$ ($E^{(n)} \subseteq \mathcal{P}$), defined over $\mathcal{X}$. Then:*[3]

$$\begin{aligned} -\min_{Q\in\, Li\ (E^{(n)}\cap\mathcal{P}_n)} \mathcal{D}(Q||P) \leq & \liminf_{n\to\infty} \frac{1}{n}\log P(P_n\in E^{(n)}) \\ \leq & \limsup_{n\to\infty} \frac{1}{n}\log P(P_n\in E^{(n)}) \\ \leq & -\min_{Q\in\, LsE^{(n)}} \mathcal{D}(Q||P). \end{aligned} \tag{A7}$$

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

$$\lim_{n\to\infty} \frac{1}{n}\log P(P_n\in E^{(n)}) \to -\min_{Q\in LimE^{(n)}} \mathcal{D}(Q||P). \tag{A8}$$

Proof. We first prove the expression for the lower bound. Let $E_n = E^{(n)} \cap \mathcal{P}_n$. We have:

$$\begin{aligned} P(E^{(n)}) &= \sum_{Q\in E_n} P(T(Q)) \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\min_{Q\in E_n}\mathcal{D}(Q||P)} \\ &\leq (n+1)^{|\mathcal{X}|} 2^{-n\inf_{Q\in E^{(n)}}\mathcal{D}(Q||P)} \\ &= (n+1)^{|\mathcal{X}|} 2^{-n\min_{Q\in cl(E^{(n)})}\mathcal{D}(Q||P)}. \end{aligned} \tag{A9}$$

In the last inequality we exploited the fact that, being each $E^{(n)}$ a bounded subset of $\mathcal{P}$, and $\mathcal{D}$ lower bounded in $\mathcal{P}$, the infimum over $E^{(n)}$ corresponds to the minimum over its closure.

By taking the logarithm of each side in (A9) and dividing by n, we get:

$$\frac{1}{n}\log P(E^{(n)}) \leq -\min_{Q\in cl(E^{(n)})} \mathcal{D}(Q||P) + \frac{\log(n+1)^{|\mathcal{X}|}}{n}, \tag{A10}$$

We now prove that, if n is sufficiently large, we have

$$\min_{Q\in cl(E^{(n)})} \mathcal{D}(Q||P) \geq \min_{Q\in LsE^{(n)}} \mathcal{D}(Q||P). \tag{A11}$$

Firstly, according to the properties of the limit superior, $LsE^{(n)} = Ls(cl(E^{(n)}))$ [193]. By contradiction, let us assume that the left-hand side of (A11) is strictly lower than the right-hand side. Let Q_n be a point achieving the minimum of the left-hand side of (A11). Because of the strict inequality, it must be $Q_n \notin LsE^{(n)}$. This means that there exists a value δ, such that the neighborhood $B(Q_n, \delta)$ intersects only a finite number of $E^{(n)}$, that is, there exists a finite value n' such that $B(Q_n,\delta) \cap cl(E^{(n)}) = \emptyset\ \forall n \geq n'$. Hence, Q_n can not belong to $cl(E^{(n)})$ for $n \geq n'$, thus raising an absurd.

[3] We are assuming that $E^{(n)} \cap \mathcal{P}_n \neq \emptyset$.

Hence, by exploiting equation (A11) in (A10), we have that, for large n,

$$\frac{1}{n}\log P(E^{(n)}) \leq -\min_{Q\in LsE^{(n)}} \mathcal{D}(Q||P) + \frac{\log(n+1)^{|\mathcal{X}|}}{n}, \tag{A12}$$

and hence

$$\limsup_{n\to\infty}\frac{1}{n}\log P(E^{(n)}) \leq -\min_{Q\in LsE^{(n)}} \mathcal{D}(Q||P). \tag{A13}$$

Let us pass to the upper bound. We have:

$$\begin{aligned} P(E^{(n)}) &= \sum_{Q\in E_n} P(T(Q)) \\ &\geq P(T(Q_n)) \\ &\geq \frac{2^{-n\mathcal{D}(Q_n||P)}}{(n+1)^{|\mathcal{X}|}}, \end{aligned} \tag{A14}$$

for any Q_n in E_n. Let Q^* be a point achieving the minimum of the divergence over the set LiE_n. By definition of limit inferior, there exists a sequence of points $\{Q_n\}$, $Q_n \in E_n$ such that $Q_n \to Q^*$ as $n \to \infty$. Then, by exploiting the continuity of $\mathcal{D}$, it follows that

$$\mathcal{D}(Q_n||P) \leq D(Q^*||P) + \gamma, \tag{A15}$$

where γ can be made arbitrarily small for large n. Hence, we get

$$\begin{aligned} \frac{1}{n}\log P(E^{(n)}) &\geq -\mathcal{D}(Q_n||P) - |\mathcal{X}|\frac{\log(n+1)}{n}, \\ &\geq -\mathcal{D}(Q^*||P) - \gamma - |\mathcal{X}|\frac{\log(n+1)}{n}, \\ &= -\inf_{Q\in LiE_n} \mathcal{D}(Q||P) - \gamma - |\mathcal{X}|\frac{\log(n+1)}{n}, \end{aligned} \tag{A16}$$

and then

$$\liminf_{n\to\infty}\frac{1}{n}\log P(E^{(n)}) \geq -\inf_{Q\in LiE_n} \mathcal{D}(Q||P), \tag{A17}$$

which concludes the proof of the first part (relation (A7)).

For the proof of the second part, we observe that, when $LsE^{(n)} = Li(E^{(n)} \cap \mathcal{P}_n)$, the two bounds in (A7) coincides. Moreover, in such conditions, the following chain of inclusions holds, $LiE^{(n)} \subseteq LsE^{(n)} = Li(E^{(n)} \cap \mathcal{P}_n) \subseteq LiE^{(n)}$, and then we have that $LiE^{(n)} = LsE^{(n)} = LimE^{(n)}$, yielding (A8). □

We observe that, in general, the Kuratowski convergence of $E^{(n)}$ is a *necessary* condition for the existence of the generalized Sanov limit in (A8), but it is not sufficient (in fact, it could be $LiE^{(n)} \supseteq Li(E^{(n)} \cap \mathcal{P}_n)$, in which case the lower and upper bound in (A7) do not coincide). Theorem 24 has the following interesting corollary.

Corollary 4. *When $E^{(n)}$ is a sequence of subsets of $\mathcal{P}_n$, the generalized Sanov's limit holds whenever $E^{(n)} \xrightarrow{K} E$ for some set E, or equivalently, by exploiting the compactness of space $\mathcal{P}$, $E^{(n)} \xrightarrow{H} E$.*

Proof. The corollary follows immediately by observing that when $E^{(n)} \subseteq \mathcal{P}_n$, $E_n = E^{(n)} \cap \mathcal{P}_n = E^{(n)}$. □

The generalization of Sanov's theorem finds applications in the computation of the error exponent of the Type II error probability of the decision of the hypothesis test in the presence of adversary. The proof of Theorem 2 in Chapter 3, as well as the proof of Theorem 6 in Chapter 4 and the one of Theorem 4 in Chapter 6, can be regarded as a straightforward application of this theorem.

We conclude with the following observation.

Observation. When the sequence $\{E^{(n)}\} = E \ \forall n$ (or from a certain n onwards), the generalized Sanov's theorem corresponds to the *classical* Sanov's theorem. In fact, we have that $LsE^{(n)} = LsE = E$, while the set $Li(E_n)$ coincides with $Li(E \cap \mathcal{P}_n)$, i.e., the set of all the accumulation points of sequences in $E \cap \mathcal{P}$. Since $Li(E \cap \mathcal{P}_n) \supseteq intE$[4], we can write Sanov's bounds:

$$\begin{aligned} \inf_{Q \in E} \mathcal{D}(Q||P) &\leq -\limsup_{n \to \infty} \frac{1}{n} \log P(P_n \in E) \\ &\leq -\liminf_{n \to \infty} \frac{1}{n} \log P(P_n \in E) \\ &\leq \inf_{Q \in \ Li \ (E \cap \mathcal{P}_n)} \mathcal{D}(Q||P), \\ &\leq \inf_{Q \in intE} \mathcal{D}(Q||P). \end{aligned} \tag{A18}$$

[4]It is easy to show that every $p \in int(E)$ is an accumulation point for a sequence in $E \cap \mathcal{P}_n$.

Appendix B

Regularity properties of the admissibility set

To prove the theorems on the asymptotic behavior of the payoff in the various versions of the detection game, we need the following result, which holds under the assumption that the set of admissible maps $\mathcal{A}$ in (3.19) is determined by a set of linear constraints.

To derive our results we need to define a distance measure between transportation maps, that is a function $d_s : \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \times \mathbb{R}^{|\mathcal{X}|\times|\mathcal{X}|} \to \mathbb{R}^+$. Let us (arbitrarily) choose 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 9. *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, $\delta_H(\mathcal{A}(L,P'), \mathcal{A}(L,P)) \leq |\mathcal{X}|^2 \cdot \tau$ [1], 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 $n > n^*$, $\delta_H(\mathcal{A}^n(L,P'), \mathcal{A}(L,P)) \leq \varepsilon$, $\forall P' \in B(P,\tau) \cap \mathcal{P}_n$, 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 limited, i.e. is a convex polytope [195, 196]. 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 $\{j : S_{PV}(i,j) \neq 0\}$ 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}) [2] both in the case of additive distortion constraint (see 3.20) and L_∞ distortion constraint (see 3.49), which are the cases we focus on in this thesis. Then, $S_{P'V'} \in \mathcal{A}(L,P')$. Besides, by construction $|S_{P'V'}(i,j) - S_{PV}(i,j)| \leq \tau$,

[1] We remind (Section 3.1) that $|\mathcal{X}|$ corresponds to the cardinality of the space the simplex $\mathcal{P}$ lives in.

[2] Remember that any move from a bin to itself does not increase the distortion.

$\forall i, j$. 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 [3].

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 $|S^n_{P'V'}(i, j) - S_{P'V}(i, j)| \leq 1/n$. In fact, for any fixed i, we can define $S^n_{P'V'}$ as:

$$S^n_{P'V'}(i, j) = \max\{k : k/n \leq S_{P'V}(i, j)\}/n, \quad \forall j \neq i, \tag{A1}$$

$$S^n_{P'V'}(i, i) = 1 - \sum_{j \neq i} S_{P'V}(i, j), \tag{A2}$$

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 (A2) 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

$$\begin{aligned} \delta_H(\mathcal{A}^n(L, P'), \mathcal{A}(L, P)) &\leq \delta_H(\mathcal{A}^n(L, P'), \mathcal{A}(L, P')) + \delta_H(\mathcal{A}(L, P'), \mathcal{A}(L, P)) \\ &\leq \delta_H(\mathcal{A}^n(L, P'), \mathcal{A}(L, P')) + |\mathcal{X}|^2 \cdot \tau \\ &\leq 2|\mathcal{X}|^2/n + |\mathcal{X}|^2 \cdot \tau. \end{aligned} \tag{A3}$$

Then, for a fixed ϵ, by choosing τ^* and n^* such that $|\mathcal{X}|^2 \cdot (2/n^* + \tau^*) = \varepsilon$, we obtain 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 25. *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 $V' \in B(V, \varepsilon)$, with $\varepsilon \leq |\mathcal{X}|^2 \cdot \tau$.*

Similarly, for any $\varepsilon' > 0$, there exists τ^* and n^* such that $\forall\ \tau < \tau^*$ 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 $V'_n \in B(V, \varepsilon')$, $\forall P' \in B(P, \tau) \cap \mathcal{P}_n$, 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, $\forall j$

$$\begin{aligned} V'(j) = \sum_i S_{P'V'}(i, j) &< \sum_i (S_{PV}(i, j) + |S_{P'V'}(i, j) - S_{PV}(i, j)|) \\ &\leq V(j) + d_s(S_{P'V'}, S_{PV}) \\ &\leq V(j) + \delta_H(\mathcal{A}(L, P'), \mathcal{A}(L, P)), \end{aligned} \tag{A4}$$

[3] 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, Section 3.1).

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 9, we get

$$|V'(j) - V(j)| < \delta_H(\mathcal{A}(L,P'),\mathcal{A}(L,P)) < |\mathcal{X}|^2 \cdot \tau, \tag{A5}$$

and hence $V' \in B(V, |\mathcal{X}|^2 \cdot \tau)$. Similarly, for the second part, we observe that, from Lemma 9, for a proper choice of the admissible map $S^n_{P'V'}$, we have

$$|V'_n(j) - V(j)| < \delta_H(\mathcal{A}^n(L,P'),\mathcal{A}(L,P)) \leq 2|\mathcal{X}|^2/n + |\mathcal{X}|^2 \cdot \tau. \tag{A6}$$

Then, for a fixed ε, we can choose τ^* and n^* such that $2|\mathcal{X}|^2/n^* + |\mathcal{X}|^2 \cdot \tau^* = \varepsilon$. □

Appendix C

Asymptotic behavior of the indistinguishability regions

C.1 Behavior of set Γ_{ks} and Γ_{tr} for $\lambda \to 0$.

We start by studying the behavior of $\Gamma_{ks}(P_X, \lambda, L)$ when $\lambda \to 0$. More specifically, we show that for small values of λ the set $\Gamma_{ks}(P_X, \lambda, L)$ approaches $\Gamma(P_X, L)$ smoothly.

As a first step, we highlight the following property.

Property 5. $\mathrm{EMD}(P, Q)$ *is a continuous and convex function of* P *and* Q.

Proof. Property 5 follows immediately if we look at the *EMD* as the solution of a Linear Programming (LP) problem (see Section 5.2.1), wherein P and Q are the known terms of the linear constraints. In fact, it is a known result in operations research that the minimum of the objective function of an LP problem is a continuous and convex function of the known terms of the linear constraints [162].

□

By exploiting the continuity of the divergence and the continuity and convexity of the *EMD*, we now show that when λ tends to 0, the set $\Gamma_{ks}(P_X, \lambda, L)$ tends to $\Gamma(P_X, L)$ regularly. More precisely, the following lemma holds.

Lemma 10. *Let* $X \sim P_X$ *be an information source and* L *the maximum allowable average per-letter distortion in the* DT_{ks} *game. The set* $\Gamma_{ks}(P_X, \lambda, L)$*, defined in* (5.3)*, satisfies the following property:*

$$\forall \tau > 0, \exists \lambda > 0 \ \textit{s.t.} \ \forall P \in \Gamma_{ks}(P_X, \lambda, L) \ \exists P' \in \Gamma(P_X, L) \ \textit{s.t.} \ P \in B(P', \tau),$$

where $\Gamma(P_X, L)$ *is defined as in* (5.4) *and* $B(P', \tau)$ *is a ball centered in* P' *with radius* τ.

Proof. Throughout the proof we will refer to Figure C.1 where all the sets and quantities involved in the proof are sketched. For any $\tau > 0$, we consider the set:

$$\Gamma_\tau(P_X, L) = \{P : \exists P' \in \Gamma(P_X, L) \text{ s.t. } P \in B(P', \tau)\}. \tag{A1}$$

With such a definition, we can rephrase (A1) as follows:

$$\forall \tau > 0, \exists \lambda > 0 \text{ s.t. } \Gamma_{ks}(P_X, \lambda, L) \subseteq \Gamma_\tau(P_X, L). \tag{A2}$$

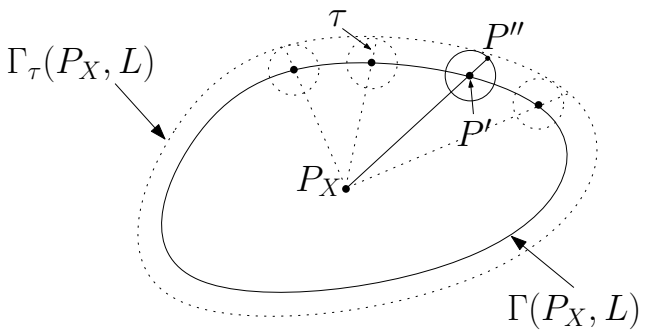


Figure C.1: Graphical representation of the set $\Gamma_\tau(P_X, L)$.

For sake of simplicity, we will prove a slightly stronger version of the lemma by means of the following two-step proof. First, we will show that a subset of $\Gamma_\tau(P_X, L)$ exists having the following form:

$$\Gamma_\tau^{sub}(P_X, L) = \{P : EMD(P, P_X) \leq L + \delta(\tau)\}, \tag{A3}$$

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 $\mathcal{B}(\Gamma(P_X, L))$, the boundary of $\Gamma(P_X, L)$. Among all the 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 C.1). By the convexity of the *EMD* (Property 5) 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 $\delta(\tau) = \min_{P' \in \mathcal{B}(\Gamma(P_X, L))} \mu(P', \tau)$. 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 $B(P^*, \tau)$ 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 $P \in B(P^*, \tau)$ as required.

Let us now pass to the second part of the proof. First, we notice that set $\Gamma_{ks}(P_X, \lambda, L)$ 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 $Q \in B(P_X, \kappa(\lambda))$ 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 10 holds also when $\Gamma_{ks}(P_X, \lambda, L)$ is replaced by $\Gamma_{tr}(Q, \lambda, L)$ and $\Gamma(P_X, L)$ by $\Gamma(Q, L)$ with a generic Q instead of P_X. To be convinced about that, it is sufficient to note that the only difference between Γ_{ks} 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 10 still holds.

C.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 11 (Extension of Lemma 10 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 5.3, satisfies the following property:*

$$\forall \tau > 0, \exists \lambda > 0 \text{ s.t., } \forall P \in \Gamma_{L_\infty}(P_X, \lambda, L) \ \exists P' \in \Gamma_{L_\infty}(P_X, L) \text{ s.t. } P \in B(P', \tau), \tag{A4}$$

where $B(P', \tau)$ is a ball centered in P' with radius τ.

Proof. We will prove the lemma by assuming that the distance defining the ball $B(P', \tau)$ 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 (3.28) 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:

$$\begin{aligned} &\gamma^+(j) = \gamma(j) && \text{if } P_X(j) - Q(j) \geq 0 \\ &\gamma^+(j) = 0 && \text{if } P_X(j) - Q(j) < 0, \end{aligned} \tag{A5}$$

and $\gamma^-(j)$ defined as

$$\begin{aligned} &\gamma^-(j) = -\gamma(j) && \text{if } P_X(j) - Q(j) < 0 \\ &\gamma^-(j) = 0 && \text{if } P_X(j) - Q(j) \geq 0. \end{aligned} \tag{A6}$$

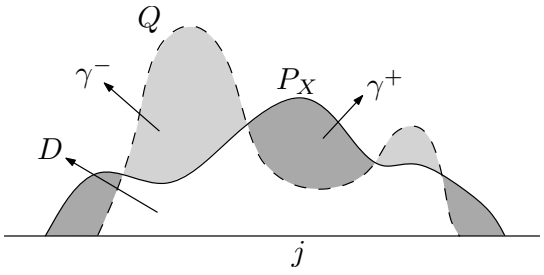


Figure C.2: Geometric interpretation of γ^+, γ^- and $D(j)$.

A graphical interpretation of γ^+ and γ^- is given in Figure C.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 which 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:

$$P(i) = \sum_j S^D_{PQ}(i,j) + \sum_j S^\gamma_{PQ}(i,j) = P_D(i) + P_\gamma(i), \tag{A7}$$

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

$$P'(i) = P_D(i) + \gamma^+(i) \tag{A8}$$

$$S'(i,j) = S^D_{PQ}(i,j) + \gamma^+(j)\delta(i,j), \tag{A9}$$

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

$$\max_{(i,j):S'(i,j)\neq 0} |i-j| \leq \max_{(i,j):S_{PQ}(i,j)\neq 0} |i-j| \leq L, \tag{A10}$$

(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. □

Appendix D

Convexity of $\mathcal{D}$ as a function of the displacement map

Let y^n be a n-length sequence. Let $N = \{n(i,j)\}_{i\in\mathcal{X}, j\in\mathcal{X}}$ be a displacement map, where $n(i,j)$ indicates the number of elements that should be moved from the i-th bin to the j-th bin. Let z^n denote a n-length sequence which results by applying the displacement map. We want to show that the objective function of the optimization problem expressed in (10.1) (corresponding to the divergence function $\mathcal{D}(P_{z^n}||P_X)$, for some pmf $P_X \in \mathcal{P}$), is convex in N. Let us indicate it by $g(N)$. Then, for any two maps (matrices) N_1 and N_2 and any two values $\alpha \in [0,1]$ and $\beta = 1-\alpha$, we have to prove that

$$g(\alpha N_1 + \beta N_2) \leq \alpha g(N_1) + \beta g(N_2). \tag{A1}$$

Let N^j be the j-th column of N and let $g_j(N^j)$ be defined as:

$$g_j(N^j) = \left(\sum_k n(k,j)\right) \cdot \log \frac{(\sum_k n(k,j))}{nP_X(j)}. \tag{A2}$$

We clearly have:

$$\begin{aligned} g(N) &= \sum_{j=1}^{|\mathcal{X}|} \frac{(\sum_k n(k,j))}{n} \cdot \log \frac{(\sum_k n(k,j))}{nP_X(j)} \\ &= \frac{1}{n}\sum_{j=1}^{|\mathcal{X}|} g_j(N^j). \end{aligned} \tag{A3}$$

By definition g_j does not depend on $n(k,i)\ \forall i \neq j$, hence, if relation (A1) holds for each g_j, then it also holds for the overall function $g(N)$. We have

$$\begin{aligned} g_j(\alpha N_1^j + \beta N_2^j) = &\sum_k (\alpha n_1(k,j) + \beta n_2(k,j)) \\ &\cdot \log \frac{\sum_k (\alpha n_1(k,j) + \beta n_2(k,j))}{nP_X(j)}, \end{aligned} \tag{A4}$$

that we conveniently rewrite as:

$$\begin{aligned} g_j(\alpha N_1^j + \beta N_2^j) = &\left(\alpha \sum_k n_1(k,j) + \beta \sum_k n_2(k,j)\right) \\ &\cdot \log \frac{\alpha \sum_k n_1(k,j) + \beta \sum_k n_2(k,j)}{\alpha n P_X(j) + \beta n P_X(j)}. \end{aligned} \tag{A5}$$

Being $n(k,j)$ nonnegative, we can apply the log-sum inequality [90] to (A5), obtaining

$$\begin{aligned} g_j(\alpha N_1^j + \beta N_2^j) &\leq \alpha \sum_k n_1(k,j) \cdot \log \frac{\alpha(\sum_k n_1(k,j))}{\alpha n P_X(j)} \\ &\quad + \beta \sum_k n_2(k,j) \cdot \log \frac{\beta(\sum_k n_2(k,j))}{\beta n P_X(j)} \\ &= \alpha g_j(N_1^j) + \beta g_j(N_2^j), \end{aligned} \tag{A6}$$

which completes the proof.

Every day we share our personal information with digital systems which are constantly exposed to threats. Security-oriented disciplines of signal processing have then received increasing attention in the last decades: multimedia forensics, digital watermarking, biometrics, network intrusion detection, steganography and steganalysis are just a few examples. Even though each of these fields has its own peculiarities, they all have to deal with a common problem: the presence of adversaries aiming at making the system fail. It is the purpose of Adversarial Signal Processing to lay the basis of a general theory that takes into account the impact of an adversary on the design of effective signal processing tools.

By focusing on the most prominent problem of Adversarial Signal Processing, namely binary detection or Hypothesis Testing, we contribute to the above mission with a general theoretical framework for the binary detection problem in the presence of an adversary. We resort to Game Theory and Information Theory concepts to model and study the interplay between the decision function designer, a.k.a. Defender, and the adversary, a.k.a. Attacker. We analyze different scenarios depending on the adversary's behavior, the decision setup and the players' knowledge about the statistical characterization of the system. Then, we apply some of the theoretical findings to specific problems in multimedia forensics: the detection of contrast enhancement and multiple JPEG compression.

UNIVERSITÀ
DI SIENA
1240

The Ph.D. School of Information Engineering of the University of Siena is a school aiming at educating scholars in a number of fields of research in the Information Engineering area. The Ph.D. School of Information Engineering is part of the Santa Chiara High School of the University of Siena. A Scientific Committee of external experts recognized Ph.D. Schools belonging to Santa Chiara as excellent, according to their degree of internationalization, their research, and educational activities.