Theoretical Foundations of Adversarial Binary Detection

Other titles in Foundations and Trends® in Communications and Information Theory

Polynomial Methods in Statistical Inference
Yihong Wu and Pengkun Yang
ISBN: 978-1-68083-730-8

Information-Theoretic Foundations of Mismatched Decoding
Jonathan Scarlett, Albert Guillen i Fabregas, Anelia Somekh-Baruch and Alfonso Martinez
ISBN: 978-1-68083-712-4

Coded Computing: Mitigating Fundamental Bottlenecks in Large-Scale Distributed Computing and Machine Learning
Songze Li and Salman Avestimehr
ISBN: 978-1-68083-704-9

Theoretical Foundations of Adversarial Binary Detection

Mauro Barni
Department of Information Engineering and Mathematics
University of Siena
Italy
barni@dii.unisi.it

Benedetta Tondi
Department of Information Engineering and Mathematics
University of Siena
Italy
benedettatondi@gmail.com

now
the essence of knowledge
Boston — Delft

Foundations and Trends® in Communications and Information Theory

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

M. Barni and B. Tondi. *Theoretical Foundations of Adversarial Binary Detection*. Foundations and Trends® in Communications and Information Theory, vol. 18, no. 1, pp. 1–172, 2021.

ISBN: 978-1-68083-765-0

Foundations and Trends® in Communications and Information Theory

Volume 18, Issue 1, 2021

Editorial Scope

Topics

Foundations and Trends® in Communications and Information Theory publishes survey and tutorial articles in the following topics:

- Coded modulation
- Coding theory and practice
- Communication complexity
- Communication system design
- Cryptology and data security
- Data compression
- Data networks
- Demodulation and Equalization
- Denoising
- Detection and estimation
- Information theory and statistics
- Information theory and computer science
- Joint source/channel coding
- Modulation and signal design
- Multiuser detection
- Multiuser information theory
- Optical communication channels
- Pattern recognition and learning
- Quantization
- Quantum information processing
- Rate-distortion theory
- Shannon theory
- Signal processing for communications
- Source coding
- Storage and recording codes
- Speech and Image Compression
- Wireless Communications

Information for Librarians

Foundations and Trends® in Communications and Information Theory, 2021, Volume 18, 4 issues. ISSN paper version 1567-2190. ISSN online version 1567-2328 . Also available as a combined paper and online subscription.

Contents

Theoretical Foundations of Adversarial Binary Detection

Mauro Barni[1] and Benedetta Tondi[2]

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

ABSTRACT

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

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

1

Introduction

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

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

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

1.1 Application Areas

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

In network monitoring applications, for instance, a common binary detection problem consists in detecting if there is an on-going Denial of Service (DoS) attack. In the simplest case, the presence of the attack can be detected by relying on a few traffic characteristics like the traffic rate, the provenance of data packets and the frequency of traffic bursts [6]. In the likely case that the hacker responsible for the DoS attack is aware of the presence of a network monitoring service, he will try to shape the traffic resulting from the attack in such a way that its characteristics are as close as possible to those of the benign traffic loading the network in the absence of attacks (while of course retaining the effectiveness of the attack). In this way, the hacker is going to alter the statistics of the observed traffic in the presence of the attack, thus impacting heavily the performance of the monitoring service in case the service had been designed without taking into account the presence of the attacker. Of course, the designer of the monitoring service does not know exactly how the hacker will shape the traffic. In turn, the hacker may not know the exact features the traffic monitoring service is going to rely on to make his decision. This uncertainty, or lack of knowledge, characterizing both the network analyst and the hacker, must be properly taken into account by both parties to optimise the actions they are going to take. It is the goal of adversarial detection theory to model the interplay between the analyst and the hacker to suggest the *best* way for them to reach their (opposite) goals, and derive the performance the monitoring service can achieve despite the presence of the adversary.

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

Another relevant scenario, even closer to the theory presented in this monograph, is Multimedia Forensics (MF) [9]. Most problems in MF can be formulated as a binary detection or hypothesis testing problem. For instance, the MF analyst may be asked to distinguish between synthetic and natural images, or to decide if a given image has been captured by a specific device or not. In other cases, the analysis aims at deciding if an image or a video has been compressed once or multiple times, since the compression history of the image/video may reveal important aspects of the processing chain the image/video has been subject to. In yet other cases, binary detection requires understanding if a certain media has been manipulated since it has been captured or not. Since the very first days of MF research, it has been recognised that forensic analysis had to cope with the opposite effort, usually referred to as counter-forensics, made by a counterfeiter [10]. From this perspective, counter-forensics can then be defined as a way to degrade the performance of the hypothesis test envisaged by the analyst. In an attempt to avoid a never-ending loop wherein new defenses and attacks are developed iteratively, and to an extent anticipating the theory developed here, the authors of [10] argued that the Kullback–Leibler distance between the probability density functions of the observed signals after the application of the counter-forensic attack is a proper way to measure the effectiveness of the attack itself. Noticeably, such measure does not depend on the particular technique adopted by the analyst. Even though the formulation in [10] does not explicitly use the game-theoretic approach, this can be seen as the first step towards the definition of the equilibrium point of a general multimedia forensics game.

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

Biometric authentication is yet another discipline which is often faced with binary decision in settings wherein the presence of an adversary cannot be ignored. In biometric-based user verification, for instance, the authenticating system must decide whether a biometric template (a face image, a fingerprint, an iris image or any other biometric trait) belongs to a certain individual, despite the opposite efforts of an attacker aiming at building a fake template that passes the authentication test. In other cases, the owner of the biometric template modifies the template to avoid being recognized [13]. In both cases, the distortion introduced within the template as a consequence of the attack should be minimal impede the detection of the attack. Another problem pertaining to biometric security that is naturally modelled as an adversarial binary detection problem, is anti-spoofing. A spoofing attack refers to a situation wherein the attacker attempts to impersonate the target by presenting to the authentication system a synthetic copy of the biometric signal used for authentication. In the case of face-based authentication, for instance, a spoofing attack is easily implemented by showing to the authentication system the face of the victim displayed on the screen of a mobile phone (rebroadcast attack). In this framework, the goal of the anti-spoofing system is to distinguish between natural and rebroadcast images. In his turn, the attacker will try to generate the image or video to be rebroadcast in such a way that it is judged as a natural one by the spoofing detection system. In doing so, the attacker must preserve the quality of the displayed image/video since otherwise it would fail to be recognized as the victim of the attack [14].

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

1.2 Scope of the Theory

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

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

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

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

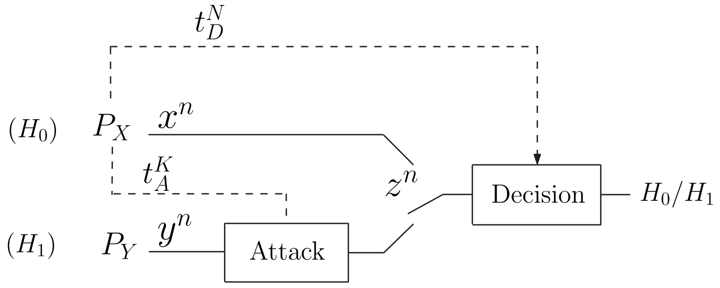

Figure 1.1: General setup of the adversarial binary decision test. P_X and P_Y govern the generation of the test sequence under H_0 and H_1 respectively. P_X also underlies the generation of the training sequences t_D^N and t_A^K for the case of binary decision based on training data.

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

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

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

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

1.2.1 Adversarial Binary Detection Setups

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

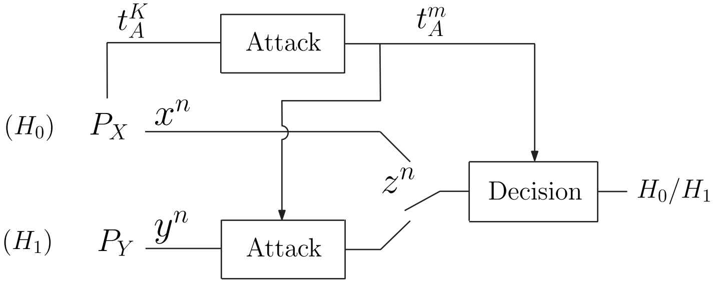

Figure 1.2: Setup of the adversarial binary decision test with corruption of the training set. P_X and P_Y rule the generation of the test sequence under H_0 and H_1 respectively. P_X also rules the generation of the training sequence t_A^K.

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

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

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

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

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

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

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

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

1.3 Related Work

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

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

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

Another extension of the theory presented in this monograph concerns the case of binary detection based on multiple observations. This scenario is relevant in several applications, including multimedia forensics, data fusion, distributed hypothesis testing and detection, sensor networks, and cognitive radio networks. In all these cases, a fusion center has to take a binary decision about the status of a system by relying on a number of observations made available by different sensors or a number of traces detected by different investigation tools. In many situations, it is possible that an attacker corrupts the observations or deliberately provides misleading data to induce a decision error at the fusion center. The binary detection game with multiple observations studied in [20] models several situations that can be traced back to the above general formulation, accounting for attackers altering a different number of observations and with different attacking capabilities.

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

As a last remark, we mention interesting relationships – deserving further investigation – between adversarial binary detection with training data and the vast body of research devoted to studying the security of Machine Learning (ML) [29], [30]. Despite the difficulty of applying the theory described in this monograph to practical applications, due to the difficulty of building precise statistical models to describe the kind of data ML systems usually involve, such a theory can be conveniently used to get useful insights about the security level that can be reached by binary detectors in practice. An example of such an analysis applied to image forensics is described in [31]. The theoretical framework behind Generative Adversarial Networks (GANs) also presents interesting connections to adversarial detection with training data. As explained in the seminal work by Goodfellow *et al.* [32], GANs are based on a game played by a generator and a discriminator, the former aiming at generating samples that mimic those of a certain class (e.g., natural images), in such a way that the discriminator can not distinguish between natural samples and samples produced by the generator. The generator, in turn, iteratively updates its decision strategy by learning the characteristics of the samples output by the generator. Interestingly, [32] shows that the equilibrium point of the game is reached when the data produced by the generator minimizes the Jensen–Shannon divergence between the distributions of natural and synthetic samples, which is by any means equivalent to the generalized log-likelihood ratio function appearing in Theorem 5.3 defining the equilibrium point of the binary detection game with training data.

1.4 Outline of the Monograph

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

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

References

[1] M. Barni and F. Pérez-González, "Coping with the enemy: Advances in adversary-aware signal processing," in *ICASSP 2013, IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 8682–8686, Vancouver, Canada, 2013.

[2] M. Barni and B. Tondi, "The source identification game: An information-theoretic perspective," *IEEE Transactions on Information Forensics and Security*, vol. 8, no. 3, pp. 450–463, 2013b.

[3] M. Barni and B. Tondi, "Binary hypothesis testing game with training data," *IEEE Transactions on Information Theory*, vol. 60, no. 8, pp. 4848–4866, 2014. DOI: 10.1109/TIT.2014.2325571.

[4] M. Barni and B. Tondi, "Source distinguishability under distortion-limited attack: An optimal transport perspective," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 10, pp. 2145–2159, 2016. DOI: 10.1109/TIFS.2016.2570739.

[5] M. Barni and B. Tondi, "Adversarial source identification game with corrupted training," *IEEE Transactions on Information Theory*, vol. 64, no. 5, pp. 3894–3915, 2018. DOI: 10.1109/TIT.2018.2806742.

[6] G. Carl, G. Kesidis, R. R. Brooks, and Suresh Rai, "Denial-of-service attack-detection techniques," *IEEE Internet Computing*, vol. 10, no. 1, pp. 82–89, 2006.

[7] G. Cormack, "Email spam filtering: A systematic review," *Foundations and Trends in Information Retrieval*, vol. 1, pp. 335–455, Jan. 2006. DOI: 10.1561/1500000006.

[8] T. Guzella and W. M. Caminhas, "A review of machine learning approaches to spam filtering," *Expert Systems with Applications*, vol. 36, pp. 10 206–10 222, Sep. 2009. DOI: 10.1016/j.eswa.2009.02.037.

[9] F. Marturana, S. Tacconi, and G. Italiano, "Handbook of digital forensics of multimedia data and devices," *A Machine Learning-Based Approach to Digital Triage,* John Wiley & Sons, Ltd., pp. 94–132, 2015.

[10] R. Böhme and M. Kirchner, "Counter-forensics: Attacking image forensics," in *Digital Image Forensics*, H. T. Sencar and N. Memon, Eds., Springer, Berlin/Heidelberg, 2012.

[11] J. Fridrich, *Steganography in Digital Media: Principles, Algorithms, and Applications.* Cambridge University Press, 2009.

[12] C. Cachin, "An information-theoretic model for steganography," in *Proceedings of IH98, Second International Workshop on Information Hiding,* Lecture Notes in Computer Science Series, vol. 6958, Springer, Berlin/Heidelberg, 1998, pp. 306–318.

[13] A. K. Jain, A. Ross, and U. Uludag, "Biometric template security: Challenges and solutions," in *Proceedings of EUSIPCO'05, European Signal Processing Conference*, pp. 469–472, 2005.

[14] B. Zhang, B. Tondi, and M. Barni, "Adversarial examples for replay attacks against CNN-based face recognition with anti-spoofing capability," *Computer Vision and Image Understanding*, vol. 197–198, 102988, 2020. DOI: 10.1016/j.cviu.2020.102988.

[15] L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein, and J. D. Tygar, "Adversarial machine learning," in *Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence*, ACM, pp. 43–58, 2011.

[16] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning," *Pattern Recognition*, vol. 84, pp. 317–331, 2018. DOI: 10.1016/j.patcog.2018.07.023.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing, Volume 2: Detection Theory.* Prentice Hall, 1998.

[18] B. Tondi, N. Merhav, and M. Barni, "Detection games under fully active adversaries," *Entropy*, vol. 21, no. 1, 2019. DOI: 10.3390/e21010023.

[19] S. Yasodharan and P. Loiseau, "Nonzero-sum adversarial hypothesis testing games," in *Conference on Neural Information Processing Systems (NeurIPS 2019)*, pp. 7312–7322, 2019.

[20] M. Barni and B. Tondi, "Multiple-observation hypothesis testing under adversarial conditions," in *2013 IEEE International Workshop on Information Forensics and Security (WIFS)*, pp. 91–96, 2013a. DOI: 10.1109/WIFS.2013.6707800.

[21] L. Lamport, R. Shostak, and M. Pease, "The Byzantine generals problem," in *Concurrency: The Works of Leslie Lamport*, 2019, pp. 203–226.

[22] A. Abrardo, M. Barni, K. Kallas, and B. Tondi, "A game-theoretic framework for optimum decision fusion in the presence of Byzantines," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 6, pp. 1333–1345, 2016.

[23] A. Abrardo, M. Barni, K. Kallas, and B.Tondi, *Information Fusion in Distributed Sensor Networks with Byzantines*. Signals and Communication Technology, Springer, Singapore, 2021.

[24] S. Marano, V. Matta, and L. Tong, "Distributed detection in the presence of Byzantine attacks," *IEEE Transactions on Signal Processing*, vol. 57, no. 1, pp. 16–29, 2009.

[25] B. Kailkhura, Y. S. Han, S. Brahma, and P. K. Varshney, "Distributed Bayesian detection in the presence of Byzantine data," *IEEE Transactions on Signal Processing*, vol. 63, no. 19, pp. 5250–5263, 2015.

[26] A. S. Rawat, P. Anand, H. Chen, and P. K. Varshney, "Collaborative spectrum sensing in the presence of Byzantine attacks in cognitive radio networks," *IEEE Transactions on Signal Processing*, vol. 59, no. 2, pp. 774–786, 2011.

[27] J. Zhang, R. S. Blum, X. Lu, and D. D. Conus, "Asymptotically optimum distributed estimation in the presence of attacks," *IEEE Transactions on Signal Processing*, vol. 63, no. 5, pp. 1086–1101, 2014.

[28] A. Vempaty, T. Lang, and P. Varshney, "Distributed inference with Byzantine data: State-of-the-art review on data falsification attacks," *IEEE Signal Processing Magazine*, vol. 30, no. 5, pp. 65–75, 2013.

[29] M. Barreno, B. Nelson, A. D. Joseph, and J. D. Tygar, "The security of machine learning," *Machine Learning*, vol. 81, no. 2, pp. 121–148, 2010.

[30] X. Wang, J. Li, X. Kuang, Y. Tan, and J. Li, "The security of machine learning in an adversarial setting: A survey," *Journal of Parallel and Distributed Computing*, vol. 130, pp. 12–23, 2019.

[31] M. Barni, M. Fontani, and B. Tondi, "A universal technique to hide traces of histogram-based image manipulations," in *Proceedings of the ACM Multimedia and Security Workshop*, pp. 97–104, Coventry, UK, Jun. 2012.

[32] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, pp. 2672–2680, 2014.

[33] J. R. Munkres, *Topology,* Featured Titles for Topology Series. Prentice Hall, Incorporated, 2000. URL: https://books.google.it/books?id=XjoZAQAAIAAJ.

[34] J. Von Neumann and O. Morgenstern, *Theory of Games and Economic Behavior*. Princeton University Press, 2007.

[35] M. J. Osborne, *An Introduction to Game Theory*, vol. 3. Oxford University Press, New York, 2004.

[36] J. Nash, "Equilibrium points in n-person games," *Proceedings of the National Academy of Sciences*, vol. 36, no. 1, pp. 48–49, 1950.

[37] M. J. Osborne and A. Rubinstein, *A Course in Game Theory.* MIT Press, 1994.

[38] J. Von Neumann, "Zur theorie der gesellschaftsspiele," ger, *Mathematische Annalen*, vol. 100, pp. 295–320, 1928. URL: http://eudml.org/doc/159291.

[39] V. Chvatal, "Linear programming," *A Series of Books in the Mathematical Sciences,* New York: W. H. Freeman and Company 1983, vol. 1, 1983.

[40] A. Charnes and W. W. Cooper, "Management models and industrial applications of linear programming," *Management Science*, vol. 4, no. 1, pp. 38–91, 1957.

[41] C. Daskalakis, P. W. Goldberg, and C. H. Papadimitriou, "The complexity of computing a Nash equilibrium," *SIAM Journal on Computing*, vol. 39, no. 1, pp. 195–259, 2009.

[42] D. Bernheim, "Rationalizable strategic behavior," *Econometrica*, vol. 52, pp. 1007–1028, 1984.

[43] Y. C. Chen, N. Van Long, and X. Luo, "Iterated strict dominance in general games," *Games and Economic Behavior*, vol. 61, no. 2, pp. 299–315, 2007.

[44] P. Weirich, *Equilibrium and Rationality: Game Theory Revised by Decision Rules*. Cambridge University Press, 2007.

[45] R. B. Myerson, *Game Theory: Analysis of Conflict*. Harvard University Press, 1991. URL: http://www.jstor.org/stable/j.ctvjsf522.

[46] I. L. Glicksberg, "A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points," *Proceedings of the American Mathematical Society*, vol. 3, no. 1, pp. 170–174, 1952. URL: http://www.jstor.org/stable/2032478.

[47] G. Monge, *Mémoire sur la Théorie des Déblais et des Remblais*. De l'Imprimerie Royale, 1781.

[48] C. Villani, *Topics in Optimal Transportation*, vol. 58. Graduate Studies in Mathematics Series: American Mathematical Society, 2003.

[49] S. T. Rachev, *Mass Transportation Problems: Volume I: Theory*, vol. 1. Springer, 1998.

[50] R. K. Ahuja, T. L. Magnanti, and J. B. Orlin, *Network Flows: Theory, Algorithms, and Applications*. Upper Saddle River, NJ: Prentice-Hall, Inc., 1993.

[51] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance as a metric for image retrieval," *International Journal on Computer Vision*, vol. 40, no. 2, pp. 99–121, 2000.

[52] C. Villani, *Optimal Transport: Old and New*. Berlin: Springer-Verlag, 2009.

[53] T. S. Rachev, "The Monge–Kantorovich mass transference problem and its stochastic applications," *Theory of Probability & Its Applications*, vol. 29, no. 4, pp. 647–676, 1985.

[54] E. Levina and P. Bickel, "The Earth Mover's Distance is the Mallows distance: Some insights from statistics," in *Proceedings of the 8th IEEE International Conference on Computer Vision*, vol. 2, pp. 251–256, 2001. DOI: 10.1109/ICCV.2001.937632.

[55] A. J. Hoffman, "On simple linear programming problems," in *Proceedings of Symposia in Pure Mathematics*, World Scientific, vol. 7, pp. 317–327, 1963.

[56] R. E. Burkard, B. Klinz, and R. Rudolf, "Perspectives of Monge properties in optimization," *Discrete Applied Mathematics*, vol. 70, no. 2, pp. 95–161, 1996.

[57] J. B. Orlin, "A faster strongly polynomial minimum cost flow algorithm," *Operations Research*, vol. 41, no. 2, pp. 338–350, 1993.

[58] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley Interscience, 1991.

[59] I. N. Sanov, "On the probability of large deviations of random variables," *Matematicheskii Sbornik*, vol. 42, pp. 11–44, 1957.

[60] I. Csiszar, "The method of types," *IEEE Transactions on Information Theory*, vol. 44, no. 6, pp. 2505–2523, 1998.

[61] A. Dembo and O. Zeitouni, *Large Deviations Techniques and Applications*, vol. 38. Springer Science & Business Media, 2009.

[62] K. Kuratowski, *Topology*, vol. 2. Academic Press, 1968.

[63] G. Salinetti and J. B. Wets, "On the convergence of sequences of convex sets in finite dimensions," *Siam Review*, vol. 21, no. 1, pp. 18–33, 1979.

[64] N. Merhav and E. Sabbag, "Optimal watermark embedding and detection strategies under limited detection resources," *IEEE Transactions on Information Theory*, vol. 54, no. 1, pp. 255–274, 2008.

[65] G. Cao, Y. Zhao, R. Ni, and X. Li, "Contrast enhancement-based forensics in digital images," *IEEE Transactions on Information Forensics and Security*, vol. 9, no. 3, pp. 515–525, 2014.

[66] X. Pan, X. Zhang, and S. Lyu, "Exposing image splicing with inconsistent local noise variances," in *2012 IEEE International Conference on Computational Photography (ICCP)*, IEEE, pp. 1–10, 2012.

[67] A. C. Popescu and H. Farid, "Statistical tools for digital forensics," in *Proceedings of the 6th International Conference on Information Hiding,* IH'04, pp. 128–147, Toronto, Canada: Springer-Verlag, 2004. DOI: 10.1007/978-3-540-30114-1_10.

[68] B. Tondi, *Theoretical Foundations of Adversarial Detection and Applications to Multimedia Forensics.* University of Siena: PhD Thesis, Sep. 2016.

[69] W. Hoeffding, "Asymptotically optimal tests for multinomial distributions," *The Annals of Mathematical Statistics*, vol. 36, no. 2, pp. 369–401, 1965.

[70] M. R. Bussieck and A. Pruessner, "Mixed-integer nonlinear programming," *SIAG/OPT Newsletter: Views & News*, vol. 14, no. 1, pp. 19–22, 2003.

[71] M. Bussieck and S. Vigerske, "MINLP solver software," 2011. DOI: 10.1002/9780470400531.eorms0527.

[72] P. Bonami, M. Kilinc, and J. Linderoth, "Algorithms and software for convex mixed integer nonlinear programs," *Tech. Rep.* Computer Sciences Department, University of Wisconsin-Madison, 2009.

[73] S. Mallat, *A Wavelet Tour of Signal Processing.* Elsevier, 1999.

[74] F. F. Leimkuhler, *Introduction to Operations Research.* Taylor & Francis Group, 1968.

[75] R. Ansari, N. Memon, and E. Ceran, "Near-lossless image compression techniques," *Journal of Electronic Imaging*, vol. 7, no. 3, pp. 486–495, 1998.

[76] Z. Wei and K. N. Ngan, "Spatio-temporal just noticeable distortion profile for grey scale image/video in DCT domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 19, no. 3, pp. 337–346, 2009.

[77] J. Lubin, "A visual discrimination model for imaging system design and evaluation," *Vision Models for Target Detection and Recognition*, vol. 2, pp. 245–357, 1995.

[78] O. Pele and M. Werman, "Fast and robust Earth Mover's Distances," in *Proceedings ICCV'09, 12th IEEE International Conference on Computer Vision*, pp. 460–467, 2009.

[79] A. C. Williams, "A treatment of transportation problems by decomposition," English, *Journal of the Society for Industrial and Applied Mathematics*, vol. 10, no. 1, pp. 35–48, 1962.

[80] M. Kendall and S. Stuart, *The Advanced Theory of Statistics,* vol. 2, 4th edition. New York: MacMillan, 1979.

[81] M. Gutman, "Asymptotically optimal classification for multiple tests with empirically observed statistics," *IEEE Transactions on Information Theory*, vol. 35, no. 2, pp. 401–408, 1989.

[82] J. Lin, "Divergence measures based on the Shannon entropy," *IEEE Transactions on Information Theory*, vol. 37, no. 1, pp. 145–151, 1991.

[83] I. Csiszár and P. Shields, *Information Theory and Statistics: A Tutorial.* Now Publishers Inc., 2004.

[84] W. A. Sutherland, *Introduction to Metric and Topological Spaces.* Oxford University Press, 1975.

[85] M. Barni, K. Kallas, and B. Tondi, "A new backdoor attack in CNNs by training set corruption without label poisoning," *Proceedings of 2019 IEEE International Conference on Image Processing, ICIP 2019, arXiv:1902.11237*, 2019.

[86] B. Biggio, G. Fumera, P. Russu, L. Didaci, and F. Roli, "Adversarial biometric recognition: A review on biometric system security from the adversarial machine-learning perspective," *IEEE Signal Processing Magazine*, vol. 32, no. 5, pp. 31–41, 2015.

[87] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," *arXiv preprint arXiv:1712.05526*, 2017.

[88] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *25th Annual Network and Distributed System Security Symposium, NDSS 2018,* San Diego, CA, USA, February 18–22, 2018, The Internet Society, 2018.

[89] N. Merhav, “Statistical physics and information theory,” *Foundations and Trends in Communications and Information Theory*, vol. 6, pp. 1–212, 2010. DOI: 10.1561/0100000052.

[90] D. Bertsimas and J. N. Tsitsiklis, *Introduction to Linear Optimization*, vol. 6. Athena Scientific Belmont, MA, 1997.

[91] S. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge University Press, 2004.