Adversarial Detection: Theoretical Foundations and Applications to Multimedia Forensics

Benedetta Tondi
Post-Doc researcher
Department of Information Engineering and Mathematics, University of Siena (Italy)
Summary

- Introduction to Adversarial Signal Processing
- Adversarial Binary Detection
- Theoretical analysis:
  - General framework for the Binary Detection problem in the presence of adversary (simple case)
- [left out] Practical analysis:
  - Applications to Multimedia Forensics
Adversarial Signal Processing (AvdSP)

Motivations:
• Every digital system is exposed to malicious threats
• Security-oriented disciplines have to cope with the presence of adversaries
  ▪ Watermarking - fingerprinting
  ▪ Multimedia forensics
  ▪ Spam filtering
  ▪ intrusion detection
  ▪ ....and many others

• Researchers have started looking for countermeasures, with limited interaction.
Adversarial Signal Processing (AvdSP)

• These fields face similar problems
  ▪ e.g. oracle attacks (in watermarking, in biometrics, in machine learning)
• …and countermeasures are similar

Idea: a **unified view**

✓ catch the real essence of the problems
✓ work out effective and general solutions
✓ avoid the cat&mouse….

Tools: *Game Theory* -> a good fit!
Game Theory in a nutshell

Two players, strategic game

\[ G(S_1, S_2, u_1, u_2) \]

\[ S_1 = \{ s_{1,1}, s_{1,2}, \ldots, s_{1,m_1} \} \] Set of strategies of Player 1

\[ S_2 = \{ s_{2,1}, s_{2,2}, \ldots, s_{2,m_1} \} \] Set of strategies of Player 2

\[ u_1(s_{1,i}, s_{2,j}) \] Payoff of Player 1 for a given profile \((s_{1,i}, s_{2,j})\)

\[ u_2(s_{1,i}, s_{2,j}) \] Payoff of Player 2 for a given profile \((s_{1,i}, s_{2,j})\)

Competitive (zero-sum) game

\[ u_1(\cdot, \cdot) = -u_2(\cdot, \cdot) = u \]

In game theory we are interested in the optimal choices of rationale players.
Game Theory in a nutshell

Nash equilibrium

None of the players gets an advantage by changing his strategy (assuming the other does not change his own)

- Very Popular
- Often unsatisfactory (for the players)

Rationalizable equilibrium

The profile which survives to iterative elimination of strictly dominated strategies (for dominance-solvable games)

Dominated strategy

$$u_1(s_{1,k}, s_{2,j}) > u_1(s_{1,i}, s_{2,j}) \quad \forall s_{2,j} \in S_2$$

$s_{1,i}$ is strictly dominated by $s_{1,k}$
Binary Detection: a recurrent problem in SP

• Was a given image taken by a given camera?
• Was this image resized/compressed twice?
• Is this image a stego or a cover?
• Does this face/fingerprint/iris belong to Mr X?
• Is this e-mail spam or not?
• Is traffic level indicating the presence of an anomaly/intrusion?
• Is X a malevolent or fair user?
  ▪ Recommender systems, reputation handling
  ▪ Cognitive radio

Common element: the presence of an adversary aiming at making the test fail
Detection problem: basic setup

$H_0 / P_X \xrightarrow{X^n} \cdot \cdot \cdot \xrightarrow{Z^n} D$

$H_1 / P_Y \xrightarrow{Y^n} A$

$P_X$ and $P_Y$: pmf’s of discrete memoryless sources $X$ and $Y$

- **Goal of the Defender (D):** decide if a sequence has been generated by $P_X$ (under $H_0$) or $P_Y$ (under $H_1$)

- **Goal of the Attacker (A):** modify a sequence generated by $P_Y$ so that it looks as if it were generated by $P_X$ subject to a distortion constraint
A motivating example from Image Forensics

Camera Y → Does it come from X ? → Camera X

attack
Detection problem: basic setup

\[ H_0 / P_X \quad x^n \quad \cdots \quad \cdots \quad z^n \quad D \]

\[ H_1 / P_Y \quad y^n \quad A \]

\( P_X \) and \( P_Y \): pmf’s of discrete memoryless sources \( X \) and \( Y \)

- **Goal of the Defender (D):** decide if a sequence has been generated by \( P_X \) (under \( H_0 \)) or \( P_Y \) (under \( H_1 \))
- **Goal of the Attacker (A):** modify a sequence generated by \( P_Y \) so that it looks as if it were generated by \( P_X \) subject to a distortion constraint
Starting from this setup….

• We studied the problem of the Adversarial Binary Detection in different scenarios depending on:
  ▪ Threat setup: attack under $H_0$ only or under both $H_0$ and $H_1$
  ▪ Decision setup: based on single or multiple observations
  ▪ Knowledge available to Defender and Attacker (full or based on training data)
  ▪ Possibility for the attacker of corrupting the training data

What we will cover….

• Binary Detection Game with known sources
  ▪ Attack under $H_1$ only, known statistics, single observation-based decision
Binary Detection Game with known sources ($\text{DT}_k$)
The DT$_{ks}$ game

Set of strategies for D

$S_D = \{ \Lambda^n : P_{FP} \leq 2^{-\lambda n} \}$

$\Lambda^n$ defined by relying on $P_{zn}$ (first-order analysis)

$\lambda$ decay rate (asymptotic analysis)

Set of strategies for A

$S_A = \{ g(\cdot) : d(y^n, g(y^n)) \leq nL \}$

$L$, maximum average per letter distortion

Payoff (zero-sum game)

$u(\Lambda^n, g) = -P_{FN} = -\sum_{y^n : g(y^n) \in \Lambda^n} P_Y(y^n)$
The DT_{ks} game: equilibrium point

**Lemma** (optimum defence strategy)

\[ \Lambda^{n,*} = \left\{ P_{zn} : \mathcal{D}(P_{zn} || P_X) < \lambda - |\mathcal{X}| \frac{\log(n + 1)}{n} \right\} \]

is a *dominant strategy* for the Defender.

**Remarks:**

- regardless of the attacking strategy (the optimum strategy is *dominant*)
- regardless of \( P_Y \) (the optimum strategy is *universal* w.r.t. \( Y \))
The $\text{DT}_{ks}$ game: equilibrium point

Optimum attack strategy

Given that D will play the dominant strategy, A must solve a minimization problem

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

**Theorem (equilibrium point):** the profile $(\Lambda^{n,*}, g^*)$ is the only rationalizable equilibrium of the game
The DT$_{ks}$ game: who wins?

**Theorem** (asymptotic payoff at the equilibrium)

Given $P_X$, $\lambda$ and $L$, it is possible to define a region $\Gamma$ for which we have:

$$\begin{cases} P_Y \in \Gamma, & \text{then } P_{FN} \to 1 \\ P_Y \notin \Gamma, & \text{then } P_{FN} \to 0 \end{cases}$$

In the latter case we have:

$$\varepsilon = \min_{P \in \Gamma} \mathcal{D}(P || P_Y)$$

$\Gamma$ -> *indistinguishability region of the test* (set of the pmf’s $P$ that cannot be distinguished from $P_X$)
The Security Margin (in the DT_{ks} setup)

Given P_x and P_y.....

Security Margin between P_X and P_Y = maximum L for which P_X and P_Y can be \textit{reliably} distinguished, \( SM(P_Y, P_X) \)

SM and Optimal Transport

If we interpret P_Y and P_X as two different ways of piling up a certain amount of soil.....

The Earth Mover Distance (EMD) is the \textit{minimum cost} necessary to transform P_Y into P_X

\[ SM(P_Y, P_X) = EMD(P_Y, P_X) \]
Further work

• Extension to
  – higher-order statistics (adversary-aware data driven classification)
  – continuous sources (on-going)
  – sources with memory
• Multiple-hypothesis testing or classification
• Applications to other fields (not only MM-Forensics)
REFERENCES

CONFERENCE PUBLICATIONS


References

JOURNAL PUBLICATIONS


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
Thank you for your attention