

NONZERO-SUM ADVERSARIAL HYPOTHESIS TESTING GAMES

Sarath Yasodharan[†] and Patrick Loiseau[‡]

[†]ECE Department, Indian Institute of Science, Bangalore 560 012, India

[‡]INRIA Grenoble Rhône-Alpes, LIG, 38400 St Martin d'Hères, France



Introduction

- Classification in the presence of an adversary who can generate data to mislead the classifier.
- Two frameworks:
 - Adversarial classification: adversary generates data vectors directly.
 - Adversarial hypothesis testing: adversary picks a probability distribution, and data is generated independently from this distribution.
- Our contributions:
 - A nonzero-sum game to model adversarial hypothesis testing problems.
 - Show existence of mixed strategy Nash Equilibrium (NE) in such games.
 - Show concentration properties of NE and error exponent associated with classification error.

A Game-Theoretic Model

- \mathcal{X} : The alphabet set with d elements. The external agent can be an attacker (with probability θ ; H_1), or a normal user (with probability $1 - \theta$; H_0).
- A normal user generates n independent samples from the probability distribution $p \in M_1(\mathcal{X})$.
- Strategy space of the attacker: $Q \subseteq M_1(\mathcal{X})$. Attacker picks a $q \in Q$ and generates n independent samples from q .
- Strategy space of the defender: $\Phi_n = \{\varphi : \mathcal{X}^n \rightarrow [0, 1]\}$. $\varphi(\mathbf{x}^n)$ denotes the probability that hypothesis H_1 is accepted.
- Attacker wants to maximize misclassification. But, there is a cost of picking an element from Q :

$$u_n^A(q, \varphi) = \sum_{\mathbf{x}^n} (1 - \varphi(\mathbf{x}^n)) q(\mathbf{x}^n) - c(q).$$

- Defender wants to minimize misclassification:

$$u_n^D(q, \varphi) = - \left(\sum_{\mathbf{x}^n} (1 - \varphi(\mathbf{x}^n)) q(\mathbf{x}^n) + \gamma \sum_{\mathbf{x}^n} \varphi(\mathbf{x}^n) p(\mathbf{x}^n) \right).$$

- $\mathcal{G}^B(d, n)$ is the above two-player game.

Assumptions on the model

- (A1) Q is a closed subset of $M_1(\mathcal{X})$, and $p \notin Q$.
- (A2) $p(i) > 0$ for all $i \in \mathcal{X}$. Furthermore, for each $q \in Q$, $q(i) > 0$ for all $i \in \mathcal{X}$.
- (A3) c is continuous on Q , and there exists a unique $q^* \in Q$ such that

$$q^* = \arg \min_{q \in Q} c(q).$$

- (A4) The point p is distant from the set Q relative to the point q^* , i.e.,

$$\{\mu \in M_1(\mathcal{X}) : D(\mu||p) \leq D(\mu||q^*)\} \cap Q = \emptyset,$$

Results

Existence and partial characterization of NE

Proposition 1. Assume (A1)-(A3). Then, there exists a mixed strategy Nash equilibrium for $\mathcal{G}^B(d, n)$. If $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE, then so is $(\hat{\sigma}_n^A, \hat{\varphi}_n)$ where $\hat{\varphi}_n$ is the likelihood ratio test given by

$$\hat{\varphi}_n(\mathbf{x}^n) = \begin{cases} 1, & \text{if } q_{\hat{\sigma}_n^A}(\mathbf{x}^n) - \gamma p(\mathbf{x}^n) > 0, \\ \varphi_{\hat{\sigma}_n^D}, & \text{if } q_{\hat{\sigma}_n^A}(\mathbf{x}^n) - \gamma p(\mathbf{x}^n) = 0, \\ 0, & \text{if } q_{\hat{\sigma}_n^A}(\mathbf{x}^n) - \gamma p(\mathbf{x}^n) < 0, \end{cases}$$

where $q_{\hat{\sigma}_n^A}(\mathbf{x}^n) = \int q(\mathbf{x}^n) \hat{\sigma}_n^A(dq)$, and $\varphi_{\hat{\sigma}_n^D} = \int \varphi(\mathbf{x}^n) \hat{\sigma}_n^D(d\varphi)$.

- Proof uses Glicksberg's fixed point theorem. Existence of pure NE is not clear.

Concentration properties of equilibrium

Lemma 1. Assume (A1)-(A3). Let $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)_{n \geq 1}$ be a sequence such that, for each $n \geq 1$, $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a mixed strategy Nash equilibrium for $\mathcal{G}^B(d, n)$. Then, $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \rightarrow 0$ as $n \rightarrow \infty$.

- Bound the error using a strategy whose acceptance region is a small neighborhood of p .

Lemma 2. Assume (A1)-(A3), and let $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)_{n \geq 1}$ be as in Lemma 1. Then, $\hat{\sigma}_n^A \rightarrow \delta_{q^*}$ weakly as $n \rightarrow \infty$.

- Uses the fact that q^* is the unique minimiser of q .

Lemma 3. Assume (A1)-(A4), and let $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)_{n \geq 1}$ be as in Lemma 1. Let $(q_n)_{n \geq 1}$ be a sequence such that $q_n \in \text{supp}(\hat{\sigma}_n^A)$ for each $n \geq 1$. Then, $q_n \rightarrow q^*$ as $n \rightarrow \infty$.

- Acceptance region of H_0 does not intersect with Q , for large n .

Error exponent

Theorem 1. Assume (A1)-(A4), and let $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)_{n \geq 1}$ be as in Lemma 1. Then,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \log e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -\Lambda_0^*(0).$$

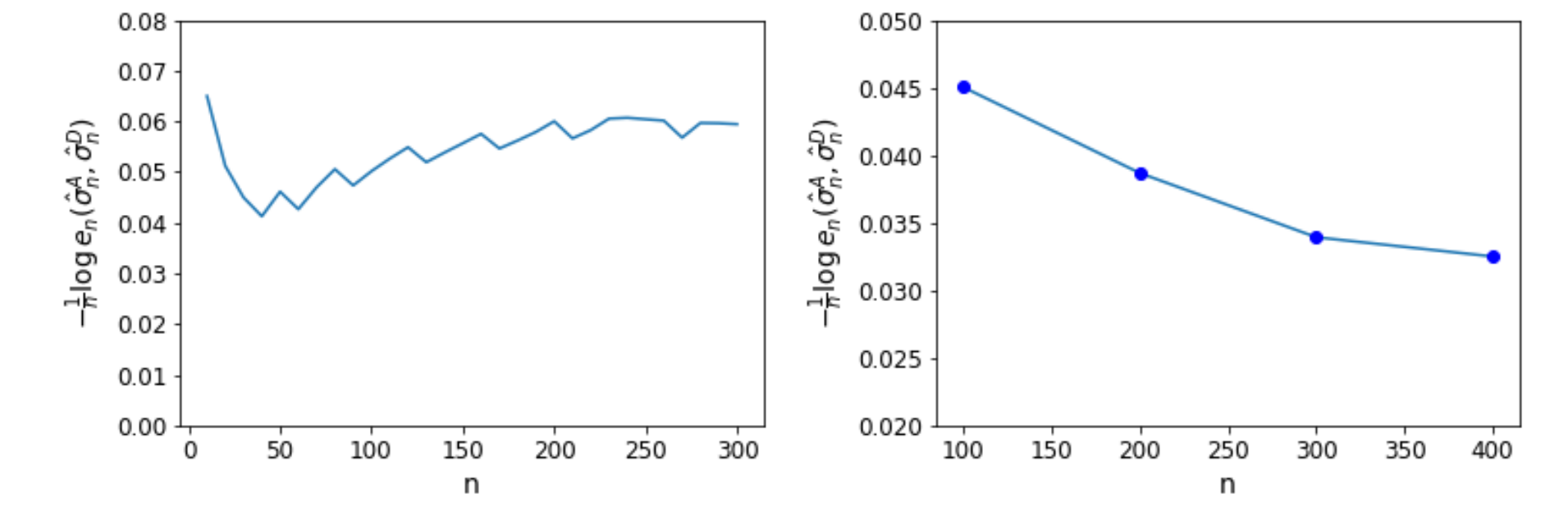
- Here, Λ_0^* denotes the convex dual of $\Lambda_0(\lambda) = \log \sum_{x \in \mathcal{X}} \exp \left(\lambda \frac{q^*(x)}{p(x)} \right) p(x)$.
- Lower bound: let the attacker play the point q^* ; Upper bound: let the defender play a specific decision rule and use Lemmas 1-3.
- Same error exponent for the classical binary hypothesis testing of p vs q^* .

Model Discussion

- Assumption (A4) is too strong. But, we have a counter example where (A4) does not hold and the conclusion of Theorem 1 fails.
- Our model is related to composite hypothesis testing. But our results are new and of a different flavour.
- Applications: Multimedia forensics, biometrics, etc.
- We have a game formulation in the Neyman-Pearson framework as well.

Numerical Results

Error Exponents

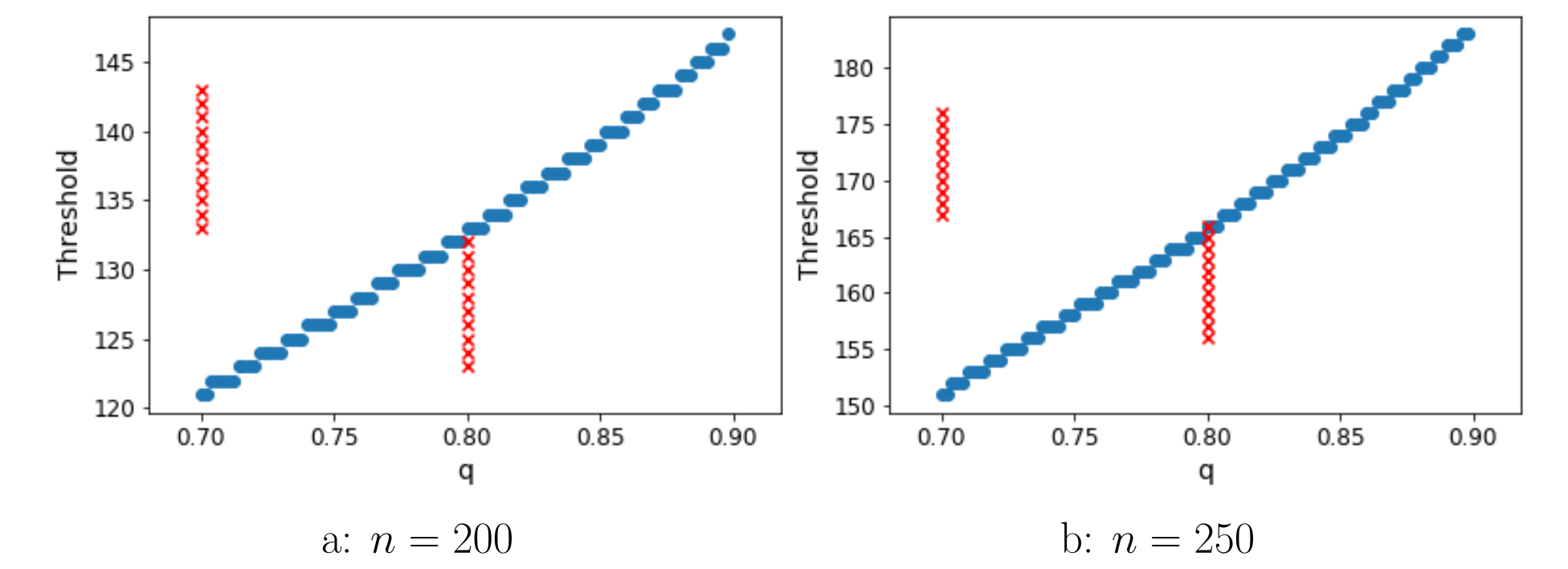


a: $Q = [0.7, 0.9]$, $c(q) = |q - 0.8|$, $\Lambda^*(0) = 0.054$
b: $Q = [0.6, 0.9]$, $c(q) = 3|q - 0.9|$, $\Lambda^*(0) = 0.111$

Figure 1: Error exponents as a function of n

- Figure 0a illustrates the conclusion in Theorem 1. Assumption (A4) does not hold in the Example in Figure 0b

Existence of a pure strategy NE for large n



a: $n = 200$ b: $n = 250$

Figure 2: Best response plots for $c(q) = |q - 0.8|$

- There is no pure NE in Figure 1a, but there is a pure NE in Figure 1b. This suggests that there is a pure NE for large n .

Future Work

- Error exponents when the cost function c has multiple minima: requires a finer analysis.
- Algorithms for computation of NE.
- Sequential hypothesis testing game: data samples arrive over time; defender needs to decide on how many samples to observe.
- Conditions for existence of pure NE.

Acknowledgements

The first author is partially supported by the Cisco-IISc Research Fellowship grant.