Nonzero-sum adversarial hypothesis testing games

Sarath A Y, Patrick Loiseau (Grenoble)

25 Sep 2019

Motivation

- Classification in the presence of an adversary:
  - Traditionally, nature is unaware of the classification algorithm.
  - An adversary generates data to mislead the classifier.
  - Classifier wants to detect the presence of the adversary and the adversary wants the classifier to make an error.

Motivation

- Classification in the presence of an adversary:
  - Traditionally, nature is unaware of the classification algorithm.
  - An adversary generates data to mislead the classifier.
  - Classifier wants to detect the presence of the adversary and the adversary wants the classifier to make an error.
- Applications:
  - Network security
  - Multimedia forensics
  - Biometrics

Motivation

- Classification in the presence of an adversary:
  - Traditionally, nature is unaware of the classification algorithm.
  - An adversary generates data to mislead the classifier.
  - Classifier wants to detect the presence of the adversary and the adversary wants the classifier to make an error.
- Applications:
  - Network security
  - Multimedia forensics
  - Biometrics
- Two rational agents: the adversary and the classifier.

Motivation

- Classification in the presence of an adversary:
  - Traditionally, nature is unaware of the classification algorithm.
  - An adversary generates data to mislead the classifier.
  - Classifier wants to detect the presence of the adversary and the adversary wants the classifier to make an error.
- Applications:
  - Network security
  - Multimedia forensics
  - Biometrics
- Two rational agents: the adversary and the classifier.
- Propose and analyse a game-theoretic model in this context of adversarial classification.

Motivation

- Classification in the presence of an adversary:
  - Traditionally, nature is unaware of the classification algorithm.
  - An adversary generates data to mislead the classifier.
  - Classifier wants to detect the presence of the adversary and the adversary wants the classifier to make an error.
- Applications:
  - Network security
  - Multimedia forensics
  - Biometrics
- Two rational agents: the adversary and the classifier.
- Propose and analyse a game-theoretic model in this context of adversarial classification.
- Adversarial hypothesis testing–adversary picks a distribution and data is generated from this.

System model

- Bayesian setting: the external agent is an attacker with probability θ (H_1), and a normal user with probability $1 - \theta$ (H_0).

System model

- Bayesian setting: the external agent is an attacker with probability θ (H_1), and a normal user with probability $1-\theta$ (H_0).
- Normal user is not strategic. Generates n i.i.d. samples from a distribution p on $\mathcal{X} = \{1, 2, \ldots, d\}$.

System model

- Bayesian setting: the external agent is an attacker with probability θ (H_1), and a normal user with probability $1-\theta$ (H_0).
- Normal user is not strategic. Generates n i.i.d. samples from a distribution p on $\mathcal{X} = \{1, 2, \ldots, d\}$.
- Attacker picks a $q \in Q \subset M_1(\mathcal{X})$ (but there is a cost for this) and generates n i.i.d. samples from q.

System model

- Bayesian setting: the external agent is an attacker with probability θ (H_1), and a normal user with probability $1 - \theta$ (H_0).
- Normal user is not strategic. Generates n i.i.d. samples from a distribution p on $\mathcal{X} = \{1, 2, \ldots, d\}$.
- Attacker picks a $q \in Q \subset M_1(\mathcal{X})$ (but there is a cost for this) and generates n i.i.d. samples from q.
- Defender: upon observing $\mathbf{x}^n$, decide H_0 or H_1.

System model

- Bayesian setting: the external agent is an attacker with probability θ (H_1), and a normal user with probability $1-\theta$ (H_0).
- Normal user is not strategic. Generates n i.i.d. samples from a distribution p on $\mathcal{X} = \{1, 2, \ldots, d\}$.
- Attacker picks a $q \in Q \subset M_1(\mathcal{X})$ (but there is a cost for this) and generates n i.i.d. samples from q.
- Defender: upon observing $\mathbf{x}^n$, decide H_0 or H_1.
- The attacker and defender are strategic–we propose a game-theoretic model for this problem.

The game $\mathcal{G}^B(d, n)$

- Two players: the attacker and defender.

The game $\mathcal{G}^B(d, n)$

- Two players: the attacker and defender.
- Strategy spaces
  - Attacker: the set of probability distributions Q on $\mathcal{X}$
  - Defender: $\Phi_n = \{\varphi : \mathcal{X}^n \to [0, 1]\}$; $\varphi(\mathbf{x}^n)$ denotes acceptance probability of H_1.

The game $\mathcal{G}^B(d, n)$

- Two players: the attacker and defender.
- Strategy spaces
  - Attacker: the set of probability distributions Q on $\mathcal{X}$
  - Defender: $\Phi_n = \{\varphi : \mathcal{X}^n \to [0, 1]\}$; $\varphi(\mathbf{x}^n)$ denotes acceptance probability of H_1.
- Utility function of the attacker:

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

The game $\mathcal{G}^B(d,n)$

- Two players: the attacker and defender.
- Strategy spaces
  - Attacker: the set of probability distributions Q on $\mathcal{X}$
  - Defender: $\Phi_n = \{\varphi : \mathcal{X}^n \to [0,1]\}$; $\varphi(\mathbf{x}^n)$ denotes acceptance probability of H_1.
- Utility function of the attacker:

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

- Utility function of the defender

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

The game $\mathcal{G}^B(d, n)$

- Two players: the attacker and defender.
- Strategy spaces
  - Attacker: the set of probability distributions Q on $\mathcal{X}$
  - Defender: $\Phi_n = \{\varphi : \mathcal{X}^n \to [0,1]\}$; $\varphi(\mathbf{x}^n)$ denotes acceptance probability of H_1.
- Utility function of the attacker:

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

- Utility function of the defender

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

- Goal: analyse the above game.
  What is the most likely outcome of this game?
  How much revenue do each players get?

Nash equilibrium

- Widely used solution concept for non-cooperative games.

Nash equilibrium

- Widely used solution concept for non-cooperative games.
- Nash equilibrium (NE): unilateral deviations do not help.

Nash equilibrium

- Widely used solution concept for non-cooperative games.
- Nash equilibrium (NE): unilateral deviations do not help. $(\hat{q}, \hat{\varphi})$ is a NE of $\mathcal{G}^B(d, n)$ if

$$u_n^A(\hat{q}, \hat{\varphi}) \geq u_n^A(q, \hat{\varphi}) \ \forall q \in Q, \text{ and}$$
$$u_n^D(\hat{q}, \hat{\varphi}) \geq u_n^D(\hat{q}, \varphi) \ \forall \varphi \in \Phi_n.$$

Nash equilibrium

- Widely used solution concept for non-cooperative games.
- Nash equilibrium (NE): unilateral deviations do not help.
  $(\hat{q}, \hat{\varphi})$ is a NE of $\mathcal{G}^B(d, n)$ if

$$u_n^A(\hat{q}, \hat{\varphi}) \geq u_n^A(q, \hat{\varphi}) \; \forall q \in Q, \text{ and}$$
$$u_n^D(\hat{q}, \hat{\varphi}) \geq u_n^D(\hat{q}, \varphi) \; \forall \varphi \in \Phi_n.$$

- But they may not always exist.

Nash equilibrium

- Widely used solution concept for non-cooperative games.
- Nash equilibrium (NE): unilateral deviations do not help. $(\hat{q}, \hat{\varphi})$ is a NE of $\mathcal{G}^B(d, n)$ if

$$
\begin{aligned}
u_n^A(\hat{q}, \hat{\varphi}) &\geq u_n^A(q, \hat{\varphi}) \ \forall q \in Q, \text{ and} \\
u_n^D(\hat{q}, \hat{\varphi}) &\geq u_n^D(\hat{q}, \varphi) \ \forall \varphi \in \Phi_n.
\end{aligned}
$$

- But they may not always exist.
- However mixed equilibria always exist for $\mathcal{G}^B(d, n)$.

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

- Assume:
  - (A1) Q is closed in $M_1(\mathcal{X})$.
  - (A2) c is continuous on Q.

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

- Assume:
  - (A1) Q is closed in $M_1(\mathcal{X})$.
  - (A2) c is continuous on Q.
- Both Q and Φ_n are compact metric spaces.

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

- Assume:
  - (A1) Q is closed in $M_1(\mathcal{X})$.
  - (A2) c is continuous on Q.
- Both Q and Φ_n are compact metric spaces.
- Define randomisations over them: $M_1(Q)$ and $M_1(\Phi_n)$ denote the spaces of probability measure on Q and Φ_n, respectively.

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

- Assume:
  - (A1) Q is closed in $M_1(\mathcal{X})$.
  - (A2) c is continuous on Q.
- Both Q and Φ_n are compact metric spaces.
- Define randomisations over them: $M_1(Q)$ and $M_1(\Phi_n)$ denote the spaces of probability measure on Q and Φ_n, respectively.
- Mixed strategy: $(\sigma^A, \sigma^D) \in M_1(Q) \times M_1(\Phi_n)$.
  $u_n^A(\sigma^A, \sigma^D) = \int u_n^A(q, \varphi)\sigma^A(dq)\sigma^D(d\varphi)$; similarly u_n^D.

Mixed strategies

- Q is equipped with the standard Euclidean topology.
- Φ_n is equipped with the "sup-norm" distance

$$d_n(\varphi_1, \varphi_2) = \max_{\mathbf{x}^n \in \mathcal{X}^n} |\varphi_1(\mathbf{x}^n) - \varphi_2(\mathbf{x}^n)|,$$

- Assume:
  - (A1) Q is closed in $M_1(\mathcal{X})$.
  - (A2) c is continuous on Q.
- Both Q and Φ_n are compact metric spaces.
- Define randomisations over them: $M_1(Q)$ and $M_1(\Phi_n)$ denote the spaces of probability measure on Q and Φ_n, respectively.
- Mixed strategy: $(\sigma^A, \sigma^D) \in M_1(Q) \times M_1(\Phi_n)$.
  $u_n^A(\sigma^A, \sigma^D) = \int u_n^A(q, \varphi)\sigma^A(dq)\sigma^D(d\varphi)$; similarly u_n^D.
- A strategy $(\hat{\sigma}^A, \hat{\sigma}^D)$ is a mixed strategy Nash equilibrium if

$$u_n^A(\hat{\sigma}^A, \hat{\sigma}^D) \geq u_n^A(\sigma^A, \hat{\sigma}^D) \ \forall \sigma^A \in M_1(Q), \text{ and}$$
$$u_n^D(\hat{\sigma}^A, \hat{\sigma}^D) \geq u_n^D(\hat{\sigma}_A, \sigma^D) \ \forall \sigma^D \in M_1(\Phi_n).$$

Existence and partial characterisation of mixed NE

Proposition

Assume (A1) and (A2). 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 $\varphi_{\hat{\sigma}_n^D} = \int \varphi(\mathbf{x}^n)\hat{\sigma}_n^D(d\varphi)$.

- Follows from the Glicksberg fixed point theorem.
- Randomisation over Φ_n is needed so show existence of NE.
- $q_{\hat{\sigma}_n^A}$ need not be a product of elements from Q.

How to study the utilities at NE

- $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE. What can we say about $u_n^A(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ and $u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$?
- Consider the classification error: $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.

How to study the utilities at NE

- $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE. What can we say about $u_n^A(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ and $u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$?
- Consider the classification error: $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- Decision rule φ^δ: accept H_0 when the empirical distribution of $\mathbf{x}^n(\mathcal{P}_{\mathbf{x}^n})$ falls in a δ-neighbourhood of p.
  (Picture on board)

How to study the utilities at NE

- $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE. What can we say about $u_n^A(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ and $u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$?
- Consider the classification error: $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- Decision rule φ^δ: accept H_0 when the empirical distribution of $\mathbf{x}^n(\mathcal{P}_{\mathbf{x}^n})$ falls in a δ-neighbourhood of p.
  (Picture on board)
- By the law of large numbers, one expects that $e_n(\hat{\sigma}_n^A, \varphi^\delta) \to 0$ as $n \to \infty$.

How to study the utilities at NE

- $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE. What can we say about $u_n^A(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ and $u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$?
- Consider the classification error: $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- Decision rule φ^δ: accept H_0 when the empirical distribution of $\mathbf{x}^n(\mathcal{P}_{\mathbf{x}^n})$ falls in a δ-neighbourhood of p.
  (Picture on board)
- By the law of large numbers, one expects that $e_n(\hat{\sigma}_n^A, \varphi^\delta) \to 0$ as $n \to \infty$.
- We have $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \leq e_n(\hat{\sigma}_n^A, \varphi^\delta) \to 0$ as $n \to \infty$.

How to study the utilities at NE

- $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE. What can we say about $u_n^A(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ and $u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$?
- Consider the classification error: $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) = -u_n^D(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- Decision rule φ^δ: accept H_0 when the empirical distribution of $\mathbf{x}^n(\mathcal{P}_{\mathbf{x}^n})$ falls in a δ-neighbourhood of p.
  (Picture on board)
- By the law of large numbers, one expects that $e_n(\hat{\sigma}_n^A, \varphi^\delta) \to 0$ as $n \to \infty$.
- We have $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \le e_n(\hat{\sigma}_n^A, \varphi^\delta) \to 0$ as $n \to \infty$.
- Thus, we anticipate that $\hat{\sigma}^A$ to concentrate on the set $\{q^* \in Q : c(q^*) \le c(q) \, \forall q \in Q\}$.

Concentration of NE

- To proceed further, we need another assumption:
  - (A3) There exists a unique $q^* \in Q$ such that

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

Concentration of NE

- To proceed further, we need another assumption:
  - (A3) There exists a unique $q^* \in Q$ such that

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

Lemma

Assume (A1)-(A3). Then, $\hat{\sigma}_n^A \to \delta_{q^*}$ weakly as $n \to \infty$:

$$\int_Q f(q)\hat{\sigma}_n^A(dq) \to f(q^*)$$

for all bounded continuous functions $f : Q \to \mathbb{R}$.

Concentration of NE

- To proceed further, we need another assumption:
  - (A3) There exists a unique $q^* \in Q$ such that

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

Lemma

Assume (A1)-(A3). Then, $\hat{\sigma}_n^A \to \delta_{q^*}$ weakly as $n \to \infty$:

$$\int_Q f(q)\hat{\sigma}_n^A(dq) \to f(q^*)$$

for all bounded continuous functions $f : Q \to \mathbb{R}$.

- Proof idea: Subsequential limits of $\{\hat{\sigma}_n^A\}_{n \geq 1}$ exist (Prohorov). Use $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \to 0$ and the fact that $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$ is a NE to show that every limit point coincides with δ_{q^*}.

Support of $\hat{\sigma}_n^A$

- Lemma 1 does not imply that $\mathrm{dist}(supp(\hat{\sigma}_n^A), q^*) \to 0$.
  (Picture on board)
- One more assumption:
  (A4) The point p is distant from the set Q relative to the point q^*, i.e.,

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

Lemma

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

- Proof idea: Show that $\sup_{q\in Q} e_n(q, \hat{\sigma}_n^D) \to 0$.
  Then use uniqueness of q^*.

Main result: error exponents

- Define

$$\Lambda_0(\lambda) = \log \sum_{i \in \mathcal{X}} \exp\left(\lambda \frac{q^*(i)}{p(i)}\right) p(i), \; \lambda \in \mathbb{R},$$

the log-moment generating function of $\frac{q^*(X)}{p(X)}$ under H_0, i.e., when $X \sim p$, and its convex dual

$$\Lambda_0^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda_0(\lambda)\}, \; x \in \mathbb{R}.$$

Main result: error exponents

- Define

$$\Lambda_0(\lambda) = \log \sum_{i \in \mathcal{X}} \exp\left(\lambda \frac{q^*(i)}{p(i)}\right) p(i),\ \lambda \in \mathbb{R},$$

the log-moment generating function of $\frac{q^*(X)}{p(X)}$ under H_0, i.e., when $X \sim p$, and its convex dual

$$\Lambda_0^*(x) = \sup_{\lambda \in \mathbb{R}} \{\lambda x - \Lambda_0(\lambda)\},\ x \in \mathbb{R}.$$

Theorem

Assume (A1)-(A4). Then,

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

Remarks

- Lower bound: let the attacker play the strategy q^*.

Remarks

- Lower bound: let the attacker play the strategy q^*.
- Upper bound: let the defender play a fixed decision rule and make use of the concentration properties of NE.

Remarks

- Lower bound: let the attacker play the strategy q^*.
- Upper bound: let the defender play a fixed decision rule and make use of the concentration properties of NE.
- The result holds for all NE $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.

Remarks

- Lower bound: let the attacker play the strategy q^*.
- Upper bound: let the defender play a fixed decision rule and make use of the concentration properties of NE.
- The result holds for all NE $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- We obtain this result without explicitly computing the structure of NE.

Remarks

- Lower bound: let the attacker play the strategy q^*.
- Upper bound: let the defender play a fixed decision rule and make use of the concentration properties of NE.
- The result holds for all NE $(\hat{\sigma}_n^A, \hat{\sigma}_n^D)$.
- We obtain this result without explicitly computing the structure of NE.
- The error exponent is the same as that of classical binary hypothesis testing between p and q^*.

Proof sketch

- Lower bound:
- Define

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

- $\mathcal{G}^B(d,n)$ is equivalent to the zerosum game with utility e_n^{eq}.
- $u_n^{eq}(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \geq u_n^{eq}(q^*, \hat{\sigma}_n^D)$.
- Can show (using the uniqueness of q^*)

$$e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \geq \sum_{\mathbf{x}^n}\left((1-\varphi_n^*(\mathbf{x}^n))q^*(\mathbf{x}^n) + \gamma\varphi_n^*(\mathbf{x}^n)p(\mathbf{x}^n)\right).$$

- Use the error exponent for classical testing of p versus q^*:

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

Proof sketch

- Upper bound:

$$\varphi'_n(\mathbf{x}^n) = \begin{cases} 1, & \text{if } \frac{q^*(\mathbf{x}^n)}{p(\mathbf{x}^n)} \geq 1, \\ 0, & \text{otherwise.} \end{cases}$$

- Decision region of φ'_n (in terms of empirical distribution):

$$\Gamma' = \{\nu \in M_1(\mathcal{X}) : D(\nu||q^*) - D(\nu||p) > 0\}.$$

- $e_n(\hat{\sigma}_n^A, \hat{\sigma}_n^D) \leq e_n(\hat{\sigma}_n^A, \varphi'_n)$.
- Easy to check:

$$e_n(\hat{\sigma}_n^A, \varphi'_n) = \int q(\mathcal{P}_{\mathbf{x}^n} \in \Gamma')\hat{\sigma}_n^A(dq) + p(\mathcal{P}_{\mathbf{x}^n} \in (\Gamma')^c).$$

- We can show that

$$q(\mathcal{P}_{\mathbf{x}^n} \in \Gamma') \leq (n+1)^d e^{-n \inf_{\nu \in \Gamma'} D(\nu||q)}.$$

Proof sketch

- Using the concentration of $\hat{\sigma}_n^A$, we have, for any $\varepsilon > 0$,

$$D(\nu \| q) \geq D(\nu \| q^*) - \varepsilon \text{ for all } q \in \operatorname{supp}(\hat{\sigma}_n^A).$$

  for sufficiently large n.
- Thus, $q(\mathcal{P}_{\mathbf{x}^n} \in \Gamma') \leq (n+1)^d e^{-n(\inf_{\nu \in \Gamma'} D(\nu \| q^*) - \varepsilon)}$
- Similarly,

$$p(\mathcal{P}_{\mathbf{x}^n} \in (\Gamma')^c) \leq (n+1)^d e^{-n(\inf_{\nu \notin \Gamma'} D(\nu \| p) - \varepsilon)}$$

- Exercise: $\inf_{\nu \notin \Gamma'} D(\nu \| p) = \inf_{\nu \in \Gamma'} D(\nu \| q^*) = \Lambda_0^*(0)$.
- Thus,

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

Numerical examples: No pure NE

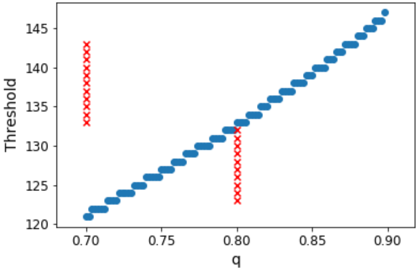

Figure: $Q = [0.7, 0.9], c(q) = |q - 0.8|, n = 200$

Numerical examples: Pure NE

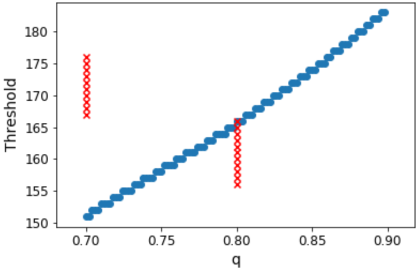

Figure: $Q = [0.7, 0.9], c(q) = |q - 0.8|, n = 250$

- This suggest that pure NE exists for large n.

Numerical examples: Error exponent

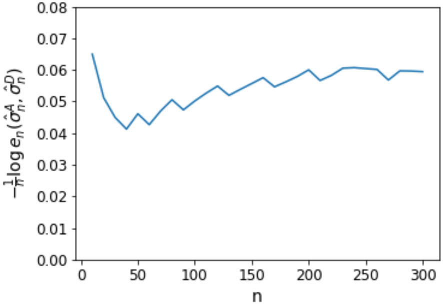

Figure: $Q = [0.7, 0.9], c(q) = |q - 0.8|, \Lambda_0^*(0) \approx 0.054$

Numerical examples: Error exponent

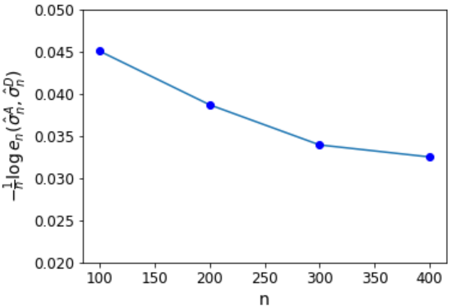

Figure: $Q = [0.6, 0.9], c(q) = 3|q - 0.9|, \Lambda_0^*(0) \approx 0.111$

Summary

- A game-theoretic model to study adversarial classification.

Summary

- A game-theoretic model to study adversarial classification.
- Results:
  - Existence and partial characterisation of mixed NE in these games.
  - Concentration properties of NE.
  - Error exponents associated with classification error.

Summary

- A game-theoretic model to study adversarial classification.
- Results:
  - Existence and partial characterisation of mixed NE in these games.
  - Concentration properties of NE.
  - Error exponents associated with classification error.
- Future work:
  - Characterisation of all NE and algorithms to compute them.
  - Relax assumptions. What if c has multiple minima? A weaker assumption than (A4)?
  - Sequential hypothesis testing game.
  - Conditions of existence of pure NE.

Summary

- A game-theoretic model to study adversarial classification.
- Results:
  - Existence and partial characterisation of mixed NE in these games.
  - Concentration properties of NE.
  - Error exponents associated with classification error.
- Future work:
  - Characterisation of all NE and algorithms to compute them.
  - Relax assumptions. What if c has multiple minima? A weaker assumption than (A4)?
  - Sequential hypothesis testing game.
  - Conditions of existence of pure NE.
- Acknowledgment: Cisco-IISc Research Fellowship Grant.

Thank you