A Formal Study of Model Inversion Attacks

Xi Wu

xiwu@cs.wisc.edu

Joint work with Matt Fredrikson, Somesh Jha
and Jeffrey F. Naughton

November 9, 2016

Theme of the Talk

- Model Inversion Attacks

Theme of the Talk

- Model Inversion Attacks
  - A kind of privacy attacks which try to "back out" sensitive data.

Theme of the Talk

- Model Inversion Attacks
  - A kind of privacy attacks which try to "back out" sensitive data.
- Main Results to Discuss

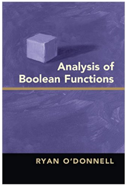
Theme of the Talk

- Model Inversion Attacks
  - A kind of privacy attacks which try to "back out" sensitive data.

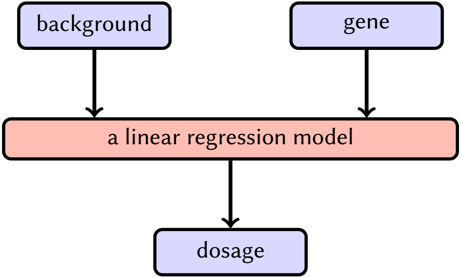
- Main Results to Discuss
  - The connection between model inversion and Boolean analysis.

Theme of the Talk

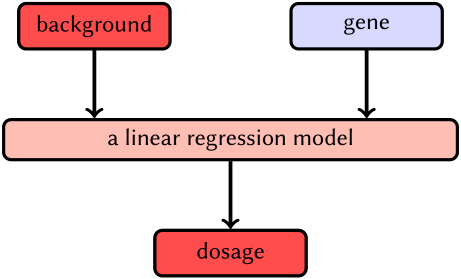
- Model Inversion Attacks
  - A kind of privacy attacks which try to "back out" sensitive data.
- Main Results to Discuss
  - The connection between model inversion and Boolean analysis.
  - Found major applications in complexity theory.



Model Inversion Attack 1 (Fredrikson et al., USENIX 2014)

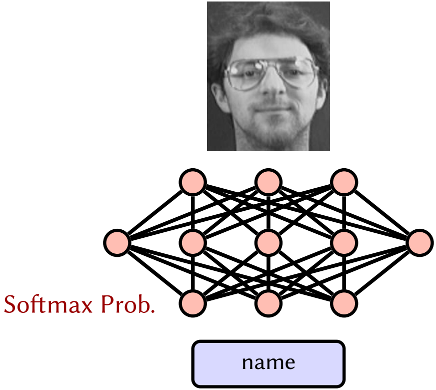


Model Inversion Attack 1 (Fredrikson et al., USENIX 2014)



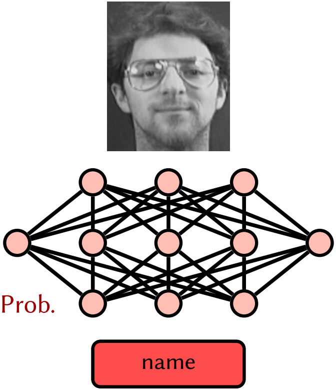
- Going from dosage and background to the genetic marker.

Model Inversion Attack 2 (Fredrikson et al., CCS 2015)



- Recover image from name:

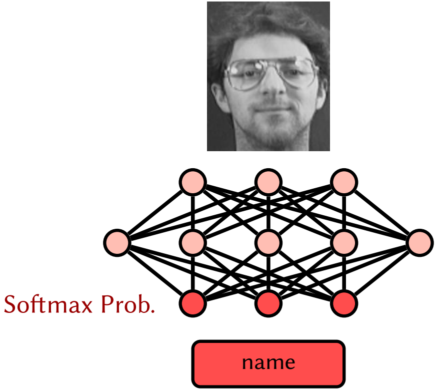
Model Inversion Attack 2 (Fredrikson et al., CCS 2015)



Softmax Prob.

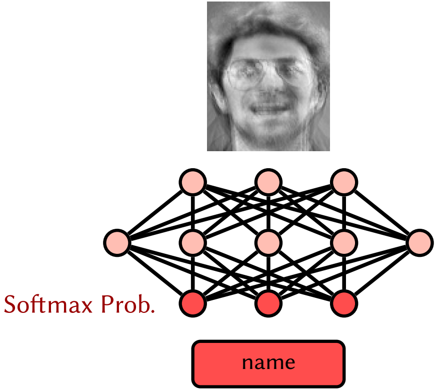
- Recover image from name: Not good if one only knows the name..

Model Inversion Attack 2 (Fredrikson et al., CCS 2015)



- Recovery is sensible if softmax probabilities are known.

Model Inversion Attack 2 (Fredrikson et al., CCS 2015)



- Recovery is sensible if softmax probabilities are known.

Other Attempts in Inverting Models

Actually, many previous attempts. (not necessarily under "data privacy.")

Other Attempts in Inverting Models

Actually, many previous attempts. (not necessarily under "data privacy.")

- *Inverting feedforward neural networks using linear and nonlinear programming.* Lu et al., 1999
- *Image Reconstruction from Bag-of-Visual-Words* Kato and Harada, CVPR 2014.
- *Image reconstruction based on local feature descriptors* Maryam Daneshi, JQ Guo, 2011
- *From Bits to Images: Inversion of Local Binary Descriptors* d'Angelo et al. arXiv 2012

Other Attempts in Inverting Models

Actually, many previous attempts. (not necessarily under "data privacy.")

- *Inverting feedforward neural networks using linear and nonlinear programming.* Lu et al., 1999
- *Image Reconstruction from Bag-of-Visual-Words* Kato and Harada, CVPR 2014.
- *Image reconstruction based on local feature descriptors* Maryam Daneshi, JQ Guo, 2011
- *From Bits to Images: Inversion of Local Binary Descriptors* d'Angelo et al. arXiv 2012

- Essence: Sensible recovery from highly compressed information.

In the Rest of the Talk

We study:

In the Rest of the Talk

We study:

- Black-box model inversion attacks for Boolean models.

In the Rest of the Talk

We study:

- Black-box model inversion attacks for Boolean models.
- Formulate noiseless and noisy models and study their "invertibility."

In the Rest of the Talk

We study:

- Black-box model inversion attacks for Boolean models.
- Formulate noiseless and noisy models and study their "invertibility."
- Connect invertibility to notions in Boolean analysis.

Things We Skip

Things We Skip

- The general framework to study model inversion attacks.
  - E.g. framework to model white-box attacks.
  - Section 2.

Things We Skip

- The general framework to study model inversion attacks.
  - E.g. framework to model white-box attacks.
  - Section 2.

- Special structure of machine learning models in white-box attacks.
  - Sequential compositions in a model as "communication games."
  - Section 5.A.

Things We Skip

- The general framework to study model inversion attacks.
  - E.g. framework to model white-box attacks.
  - Section 2.

- Special structure of machine learning models in white-box attacks.
  - Sequential compositions in a model as "communication games."
  - Section 5.A.

- Computational power of restricted communication games.
  - Very limited communication channel can leak "everything."
  - Section 5.B.

Things We Skip

- The general framework to study model inversion attacks.
  - E.g. framework to model white-box attacks.
  - Section 2.
- Special structure of machine learning models in white-box attacks.
  - Sequential compositions in a model as "communication games."
  - Section 5.A.
- Computational power of restricted communication games.
  - Very limited communication channel can leak "everything."
  - Section 5.B.

Please refer to the paper.

Boolean Analysis (1/2)

- Studies Boolean functions $f : \{\text{-1, 1}\}^n \mapsto \{\text{-1, 1}\}$.
  - $b \in \{0, 1\} \mapsto (-1)^b$.
  - Found many applications in theoretical computer science (circuit complexity, learning theory, cryptography, ...).

Boolean Analysis (1/2)

- Studies Boolean functions $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.
  - $b \in \{0, 1\} \mapsto (-1)^b$.
  - Found many applications in theoretical computer science (circuit complexity, learning theory, cryptography, ...).

Definition (Difference Operator)

D_i is a linear operator applied to a Boolean function f such that

$$(\mathrm{D}_i\, f)(x) = \frac{f(x^{i \to 1}) - f(x^{i \to -1})}{2}.$$

Intuition: Discrete "derivative."

Boolean Analysis (1/2)

- Studies Boolean functions $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.
  - $b \in \{0, 1\} \mapsto (-1)^b$.
  - Found many applications in theoretical computer science (circuit complexity, learning theory, cryptography, ...).

Definition (Difference Operator)

D_i is a linear operator applied to a Boolean function f such that

$$(\mathrm{D}_i\, f)(x) = \frac{f(x^{i\to 1}) - f(x^{i\to -1})}{2}.$$

Intuition: Discrete "derivative."

Definition (Influence)

$$\mathbf{Inf}_i[f] = \Pr_{x\sim\{\text{-}1,1\}^n}[f(x^{i\to 1}) \neq f(x^{i\to -1})]$$

Intuition: Fraction of input that x_i has influence.

Boolean Analysis (2/2)

- $N_\rho(x)$. $\widetilde{x} \sim N_\rho(x)$ if

$$\widetilde{x}_j = \begin{cases} x_j & \text{w.p. } \frac{1+\rho}{2} \\ 1 - x_j & \text{w.p. } \frac{1-\rho}{2} \end{cases}$$

Boolean Analysis (2/2)

- $N_\rho(x)$. $\widetilde{x} \sim N_\rho(x)$ if

$$\widetilde{x}_j = \begin{cases} x_j & \text{w.p. } \frac{1+\rho}{2} \\ 1 - x_j & \text{w.p. } \frac{1-\rho}{2} \end{cases}$$

Definition (Noise Stability)

Let $-1 \leq \rho \leq 1$. $\mathbf{Stab}_\rho[f] = \mathbb{E}_{\substack{x \sim \{-1,1\}^n \\ y \sim N_\rho(x)}} \left[f(x)f(y)\right].$

Intuition: Measure the change of f under noise.

Boolean Analysis (2/2)

- $N_\rho(x)$. $\widetilde{x} \sim N_\rho(x)$ if

$$\widetilde{x}_j = \begin{cases} x_j & \text{w.p. } \frac{1+\rho}{2} \\ 1 - x_j & \text{w.p. } \frac{1-\rho}{2} \end{cases}$$

Definition (Noise Stability)

Let $-1 \leq \rho \leq 1$. $\mathbf{Stab}_\rho[f] = \mathbb{E}_{\substack{x \sim \{\text{-}1,1\}^n \\ y \sim N_\rho(x)}} [f(x) f(y)]$.

Intuition: Measure the change of f under noise.

Definition (Stable Influence)

Let $0 \leq \rho \leq 1$. $\mathbf{Inf}_i^{(\rho)}[f] = \mathbf{Stab}_\rho[\mathrm{D}_i\, f] = \mathbb{E}_{\substack{x \sim \{\text{-}1,1\}^n \\ y \sim N_\rho(x)}} [\mathrm{D}_i\, f(x)\, \mathrm{D}_i\, f(y)]$.

Intuition: Measure the change of influence of x_i under noise.
Note: when $\rho = 1$, this reduces to $\mathbf{Inf}_i[f]$.

Noiseless Model

Setup:

Noiseless Model

Setup:

- $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.

Noiseless Model

Setup:

- $f : \{\text{-1}, 1\}^n \mapsto \{\text{-1}, 1\}$.
- $i \in [n]$ be the target feature to invert.

Noiseless Model

Setup:

- $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

Noiseless Model

Setup:

- $f : \{\text{-1, 1}\}^n \mapsto \{\text{-1, 1}\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	**The Simulated World**

Noiseless Model

Setup:

- $f : \{\text{-1, 1}\}^n \mapsto \{\text{-1, 1}\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i

Noiseless Model

Setup:

- $f : \{\text{-1, 1}\}^n \mapsto \{\text{-1, 1}\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	**The Simulated World**
Goal: recover x_i Nature samples $(x, b_x) \sim S$	Goal: recover x_i Nature samples $(x, b_x) \sim S$

Noiseless Model

Setup:

- $f : \{\text{-1, 1}\}^n \mapsto \{\text{-1, 1}\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	The Simulated World
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S$	Nature samples $(x, b_x) \sim S$
Nature presents x_{-i}, $\boxed{y = f(x)}$	Nature presents x_{-i}

Noiseless Model

Setup:

- $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S$	Nature samples $(x, b_x) \sim S$
Nature presents x_{-i}, $\boxed{y = f(x)}$	Nature presents x_{-i}
Adversary: $A^f(x_{-i}, y)$	Adversary: $A^*(x_{-i})$

Noiseless Model

Setup:

- $f : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$.
- $i \in [n]$ be the target feature to invert.
- $S \subseteq \{-1, 1\}^n \times \{-1, 1\}$ training set used to learn f.

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S$	Nature samples $(x, b_x) \sim S$
Nature presents x_{-i}, $\boxed{y = f(x)}$	Nature presents x_{-i}
Adversary: $A^f(x_{-i}, y)$	Adversary: $A^*(x_{-i})$

$$Adv(A, A^*) = \Pr_{z \sim S}[A^f(x_{-i}, y) = x_i] - \Pr_{z \sim S}[A^*(x_{-i}) = x_i]$$

Idea: Measure the additional invertibility (advantage) of being able to access the model with model output.

Noiseless is Easy

- As x_i is uniformly random, so $\mathbb{E}_{x\sim\{-1,1\}^n}[A^*(x_{-i}) = x_i] = \frac{1}{2}$.

Noiseless is Easy

- As x_i is uniformly random, so $\mathbb{E}_{x\sim\{\text{-}1,1\}^n}[A^*(x_{-i}) = x_i] = \frac{1}{2}$.
- For $\mathbb{E}_{x\sim\{\text{-}1,1\}^n}[A^f(x_{-i}, y) = x_i]$, consider

Noiseless is Easy

- As x_i is uniformly random, so $\mathbb{E}_{x\sim\{-1,1\}^n}[A^*(x_{-i}) = x_i] = \frac{1}{2}$.
- For $\mathbb{E}_{x\sim\{-1,1\}^n}[A^f(x_{-i}, y) = x_i]$, consider

Algorithm 3 Algorithm $A_\#$

Input: $x_{-i}, y \in \{-1, 1\}$. Oracle access to f.
1: **function** $A_\#(x_{-i}, y)$
2: Compute $y' = f(x_1, \ldots, x_{i-1}, -1, x_{i+1}, \ldots, x_n)$
3: **return** $(-1)^{\mathbb{1}[y'=y]}$

Analysis of $A_{\#}$

Analysis of $A_\#$

- The recovery is correct when $f(x^{i \to 1}) \neq f(x^{i \to -1})$.

Analysis of $A_{\#}$

- The recovery is correct when $f(x^{i\to 1}) \neq f(x^{i\to -1})$.
- Otherwise, no additional information is obtained.

Analysis of $A_\#$

- The recovery is correct when $f(x^{i\to 1}) \neq f(x^{i\to -1})$.
- Otherwise, no additional information is obtained.
- Let $p = \Pr_{x\sim\{-1,1\}^n}[f(x^{i\to 1}) \neq f(x^{i\to -1})]$, then
$$\Pr_{x\sim\{-1,1\}^n}[A_\#^f(x_{-i}, y) = x_i] = (1-p)\cdot\frac{1}{2} + p\cdot 1 = \frac{1}{2} + \frac{p}{2}$$

Analysis of $A_\#$

- The recovery is correct when $f(x^{i\to 1}) \neq f(x^{i\to -1})$.
- Otherwise, no additional information is obtained.
- Let $p = \Pr_{x\sim\{-1,1\}^n}[f(x^{i\to 1}) \neq f(x^{i\to -1})]$, then

$$\Pr_{x\sim\{-1,1\}^n}[A_\#^f(x_{-i}, y) = x_i] = (1-p)\cdot\frac{1}{2} + p\cdot 1 = \frac{1}{2} + \frac{p}{2}$$

Theorem

$(\forall A^*)\ Adv(A_\#, A^*) = \frac{\mathbf{Inf}_i[f]}{2}$.

Analysis of $A_\#$

- The recovery is correct when $f(x^{i \to 1}) \neq f(x^{i \to -1})$.
- Otherwise, no additional information is obtained.
- Let $p = \Pr_{x \sim \{-1,1\}^n}[f(x^{i \to 1}) \neq f(x^{i \to -1})]$, then

$$\Pr_{x \sim \{-1,1\}^n}[A_\#^f(x_{-i}, y) = x_i] = (1-p) \cdot \frac{1}{2} + p \cdot 1 = \frac{1}{2} + \frac{p}{2}$$

Theorem

$(\forall A^*)\ Adv(A_\#, A^*) = \frac{\mathbf{Inf}_i[f]}{2}$.

- This is in fact *optimal* given the information the adversary has.

Theorem

$(\forall A, \forall A^*)\ Adv(A, A^*) \leq \frac{\mathbf{Inf}_i[f]}{2}$.

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	The Simulated World

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$	Goal: recover x_i Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$	Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$
Nature presents $\widetilde{x}_{-i}$, $\boxed{y = f(x)}$	Nature presents $\widetilde{x}_{-i}$

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$	Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$
Nature presents $\widetilde{x}_{-i}$, $\boxed{y = f(x)}$	Nature presents $\widetilde{x}_{-i}$
Adversary: $A^f(\widetilde{x}_{-i}, y)$	Adversary: $A^*(\widetilde{x}_{-i})$

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$	Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$
Nature presents $\widetilde{x}_{-i}$, $\boxed{y = f(x)}$	Nature presents $\widetilde{x}_{-i}$
Adversary: $A^f(\widetilde{x}_{-i}, y)$	Adversary: $A^*(\widetilde{x}_{-i})$

Key: The auxiliary information is noisy – the adversary gets $\widetilde{x}_{-i}$.

Noisy Case: ρ-Independent Perturbation Model

The MI-Attack World	**The Simulated World**
Goal: recover x_i	Goal: recover x_i
Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$	Nature samples $(x, b_x) \sim S, \widetilde{x} \sim N_\rho(x)$
Nature presents $\widetilde{x}_{-i}$, $\boxed{y = f(x)}$	Nature presents $\widetilde{x}_{-i}$
Adversary: $A^f(\widetilde{x}_{-i}, y)$	Adversary: $A^*(\widetilde{x}_{-i})$

Key: The auxiliary information is noisy – the adversary gets $\widetilde{x}_{-i}$.

What is model invertibility then?

$A_\#$ Again

- Consider the same algorithm $A_\#$ again

Algorithm 4 Algorithm $A_\#$

Input: $\widetilde{x}_{-i}, y \in \{-1, 1\}$. Oracle access to f.

1: **function** $A^\#(\widetilde{x}_{-i}, y)$
2: Compute $y' = f(\widetilde{x}_1, \ldots, \widetilde{x}_{i-1}, -1, \widetilde{x}_{i+1}, \ldots, \widetilde{x}_n)$
3: **return** $(-1)^{\mathbb{1}[y'=y]}$

Instead of receiving x_{-i}, it gets now $\widetilde{x}_{-i}$.

Invertibility of $A_\#$

- Invertibility becomes "stable influence."

Invertibility of $A_\#$

- Invertibility becomes "stable influence."
- Recall that

Definition (Stable Influence)

Let $0 \leq \rho \leq 1$. The ρ-stable influence of f at i, denoted as $\mathbf{Inf}_i^{(\rho)}[f]$, is defined to be $\mathbf{Inf}_i^{(\rho)}[f] = \mathbf{Stab}_\rho[\mathrm{D}_i\, f] = \mathbb{E}_{\substack{x\sim\{-1,1\}^n \\ y\sim N_\rho(x)}}\left[\mathrm{D}_i\, f(x)\, \mathrm{D}_i\, f(y)\right]$.

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

1. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) > 0$: "Good," $A_\#$ infers x_i correctly as before.

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

1. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) > 0$: "Good," $A_\#$ infers x_i correctly as before.
2. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$: "Random guessing," the information is "erased," and $A_\#$ is "essentially" doing random guessing.

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

1. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) > 0$: "Good," $A_\#$ infers x_i correctly as before.
2. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) = 0$: "Random guessing," the information is "erased," and $A_\#$ is "essentially" doing random guessing.
3. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) < 0$: "Bad," the information is "reversed," $A_\#$ always gets it wrong!

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

1. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) > 0$: "Good," $A_\#$ infers x_i correctly as before.
2. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$: "Random guessing," the information is "erased," and $A_\#$ is "essentially" doing random guessing.
3. $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) < 0$: "Bad," the information is "reversed," $A_\#$ always gets it wrong!

Theorem

For the same $A_\#$, $(\forall A^*)\ Adv(A_\#, A^*) \leq \frac{\mathbf{Inf}_i^{(\rho)}[f]}{2}$.

Why Stable Influence is the Answer?

Let $\widetilde{x} \sim N_\rho(x)$. For $A_\#$, intuitively, there are three cases:

1. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) > 0$: "Good," $A_\#$ infers x_i correctly as before.
2. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) = 0$: "Random guessing," the information is "erased," and $A_\#$ is "essentially" doing random guessing.
3. $\mathsf{D}_i f(x)\, \mathsf{D}_i f(\widetilde{x}) < 0$: "Bad," the information is "reversed," $A_\#$ always gets it wrong!

Theorem

For the same $A_\#$, $(\forall A^*)\ Adv(A_\#, A^*) \leq \frac{\mathbf{Inf}_i^{(\rho)}[f]}{2}$.

Is $A_\#$ optimal (as in the noiseless case)?

Answer: No

Answer: No

- Subtlety: If $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$, $A_\#$ essentially does random guessing, but one can do better..

Answer: No

- Subtlety: If $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$, $A_{\#}$ essentially does random guessing, but one can do better..

- For example, consider $\mathrm{OR}_n(x_1, \ldots, x_n) = \bigvee_{i=1}^{n} x_i$.
  - The value is -1 if any input bit is -1.

Answer: No

- Subtlety: If $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$, $A_\#$ essentially does random guessing, but one can do better..

- For example, consider $\mathrm{OR}_n(x_1, \ldots, x_n) = \bigvee_{i=1}^n x_i$.
  - The value is -1 if any input bit is -1.

- If we "see" that $y = 1$, then $x = 1$.

Answer: No

- Subtlety: If $\mathsf{D}_i\, f(x)\, \mathsf{D}_i\, f(\widetilde{x}) = 0$, $A_\#$ essentially does random guessing, but one can do better..

- For example, consider $\mathrm{OR}_n(x_1, \ldots, x_n) = \bigvee_{i=1}^n x_i$.
  - The value is -1 if any input bit is -1.

- If we "see" that $y = 1$, then $x = 1$.

- That is, one can use the structure of the model f to "denoise."

Answer: No

- Subtlety: If $\mathrm{D}_i\, f(x)\, \mathrm{D}_i\, f(\widetilde{x}) = 0$, $A_{\#}$ essentially does random guessing, but one can do better..

- For example, consider $\mathrm{OR}_n(x_1, \ldots, x_n) = \bigvee_{i=1}^{n} x_i$.
  - The value is -1 if any input bit is -1.

- If we "see" that $y = 1$, then $x = 1$.

- That is, one can use the structure of the model f to "denoise."

- In fact for OR_n, in noisy model one can always achieve advantage $\frac{\mathbf{Inf}_i[\mathrm{OR}_n]}{2} = 2^{-n}$, while $\frac{\mathbf{Inf}_i^{(\rho)}[\mathrm{OR}_n]}{2} = \rho^{n-1} 2^{-n}$.

Open Question

Open Question

For any A', A^*,

$$Adv(A', A^*) \leq \ Adv(A_\#, A^*) + o_n(1)\ ?$$

Invertibility Interference (1/3)

	Noiseless	**Noisy Model**
Invertibility	Influence	Stable Influence

Invertibility Interference (1/3)

	Noiseless	**Noisy Model**
Invertibility	Influence	Stable Influence

- As $\rho \to 0$, $\mathbf{Inf}_i^{(\rho)}[\mathrm{OR}_n]$ is exponentially smaller than $\mathbf{Inf}_i[\mathrm{OR}_n]$.

Invertibility Interference (1/3)

	Noiseless	**Noisy Model**
Invertibility	Influence	Stable Influence

- As $\rho \to 0$, $\mathbf{Inf}_i^{(\rho)}[\mathrm{OR}_n]$ is exponentially smaller than $\mathbf{Inf}_i[\mathrm{OR}_n]$.
- But $\mathbf{Inf}_i[\mathrm{OR}_n]$ is "exponentially small:" 2^{1-n}. Not very interesting...

Invertibility Interference (1/3)

	Noiseless	**Noisy Model**
Invertibility	Influence	Stable Influence

- As $\rho \to 0$, $\mathbf{Inf}_i^{(\rho)}[\mathrm{OR}_n]$ is exponentially smaller than $\mathbf{Inf}_i[\mathrm{OR}_n]$.
- But $\mathbf{Inf}_i[\mathrm{OR}_n]$ is "exponentially small:" 2^{1-n}. Not very interesting...
- A more interesting phenomenon termed "invertibility interference."

Invertibility Interference (2/3)

- Consider the parity function $\chi_n(x) = \prod_{i=1}^{n} x_i$.

Invertibility Interference (2/3)

- Consider the parity function $\chi_n(x) = \prod_{i=1}^{n} x_i$.
- $\mathbf{Inf}_i[\chi_n] = 1$ — most “invertible” in the noiseless model.

Invertibility Interference (2/3)

- Consider the parity function $\chi_n(x) = \prod_{i=1}^{n} x_i$.
- $\mathbf{Inf}_i[\chi_n] = 1$ — most "invertible" in the noiseless model.
- $\mathbf{Inf}_i^{(\rho)}[\chi_n] = \rho^{n-1}$ — highly "non-invertible" in the noisy model.

Invertibility Interference (2/3)

- Consider the parity function $\chi_n(x) = \prod_{i=1}^n x_i$.
- $\mathbf{Inf}_i[\chi_n] = 1$ — most "invertible" in the noiseless model.
- $\mathbf{Inf}_i^{(\rho)}[\chi_n] = \rho^{n-1}$ — highly "non-invertible" in the noisy model.
- Why? "Influential" coordinates interfere with each other to render the model "non-invertible" when little noise present.

Invertibility Interference (3/3)

Theorem

Suppose that $h : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$ *has* t *coordinates with influence* 1*. Let* $0 < \rho \leq 1$*, then for any* $i \in [n]$*,* $\mathbf{Inf}_i^{(\rho)}[h] \leq \rho^{t-1}\, \mathbf{Inf}_i[h]$*.*

Invertibility Interference (3/3)

Theorem

Suppose that $h : \{\text{-}1, 1\}^n \mapsto \{\text{-}1, 1\}$ has t coordinates with influence 1. Let $0 < \rho \leq 1$, then for any $i \in [n]$, $\mathbf{Inf}_i^{(\rho)}[h] \leq \rho^{t-1}\, \mathbf{Inf}_i[h]$.

Open Question

If, instead of having coordinates of influence 1, we are only guaranteed that individual influence is lower bounded by $1 - \delta$ for some $\delta > 0$, how fast will the stable influence decay with respect to δ?

Thanks!

?