Overview

Two topics:

1. Regularization methods
2. Randomized smoothing

Overview

Two topics:

1. Regularization methods
   Goal: submit a journal paper ASAP (hopefully before April) for the partial isometry regularization (ISO) or other variants
   - Need some practical improvements.
   - Need more experiments.
2. Randomized smoothing

Overview

Two topics:

1. Regularization methods
2. Randomized smoothing
   Goal: submit a paper before August
   - Need to provide a synthesis of the current literature on the topic.
   - Hopefully reformulate the method with “information geometry" and highlight interesting new results.

Overview

Two topics:

1. Regularization methods
2. Randomized smoothing
   Goal: submit a paper before August
   - Need to provide a synthesis of the current literature on the topic.
   - Hopefully reformulate the method with “information geometry" and highlight interesting new results.

   Other goal: an application paper to trajectory prediction robustness in ATM, hopefully more or less done by August. Either apply ISO or randomized smoothing (RS). If no new results on RS, we can still apply current RS methods.

Overview

Two topics:

1. Regularization methods
2. Randomized smoothing
3. (Bonus topic if time: ramble about robustness)

Regularization methods

Quick reminder of ISO

We consider **multiclass classification**.
A model (e.g., a neural network) is seen as a **smooth map**

$$f : (\mathbb{R}^d, \overline{g}) \to (\Delta^c, g)$$

from the **Euclidean space** $\mathbb{R}^d$ endowed with the **Euclidean metric** $\overline{g}$ to the **probability simplex** Δ^c endowed with the **Fisher information metric** g (FIM).

Regularization methods

Quick reminder of ISO

We consider **multiclass classification**.
A model (e.g., a neural network) is seen as a **smooth map**

$$f : (\mathbb{R}^d, \overline{g}) \to (\Delta^c, g)$$

from the **Euclidean space** $\mathbb{R}^d$ endowed with the **Euclidean metric** $\overline{g}$ to the **probability simplex** Δ^c endowed with the **Fisher information metric** g (FIM).
Let $x \in \mathbb{R}^d$ be a **training point**. In coordinates, the **differential** of f at x denoted $f_{*,x}$ is:

$$\begin{aligned} f_{*,x} : T_x\mathbb{R}^d &\to T_{f(x)}\Delta^c \\ v &\mapsto J_x v \end{aligned}$$

where J_x is the Jacobian matrix of f at x.

Regularization methods

Quick reminder of ISO

The ISO method consists in constraining $f_{*,x}$ to be as close as possible to a **partial isometry**, i.e.,:

$$\overline{g}(v, w) = K.g(J_x v, J_x w) \text{ for all } v, w \in T_x\mathbb{R}^d \text{ such that } v, w \notin \ker J_x,$$

where K is some constant factor to take into account the chosen budget.

Regularization methods

Quick reminder of ISO

The ISO method consists in constraining $f_{*,x}$ to be as close as possible to a **partial isometry**, i.e.,:

$$\overline{g}(v, w) = K.g(J_x v, J_x w) \text{ for all } v, w \in T_x\mathbb{R}^d \text{ such that } v, w \notin \ker J_x,$$

where K is some constant factor to take into account the chosen budget. We obtain the following **regularization term**:

$$\alpha(x, f) = \epsilon^2 \left\| \tilde{J}_x \tilde{J}_x^T - \frac{\delta(x)^2}{\rho(x)^2 \epsilon^2} I_{c-1} \right\|_F,$$

where $\tilde{J}_x$ is the Jacobian matrix of f at x in another coordinate system on Δ^c.

Regularization methods

Quick reminder of ISO

The ISO method consists in constraining $f_{*,x}$ to be as close as possible to a **partial isometry**, i.e.,:

$$\overline{g}(v, w) = K.g(J_x v, J_x w) \text{ for all } v, w \in T_x\mathbb{R}^d \text{ such that } v, w \notin \ker J_x,$$

where K is some constant factor to take into account the chosen budget.
We obtain the following **regularization term**:

$$\alpha(x, f) = \epsilon^2 \left\| \tilde{J}_x \tilde{J}_x^T - \frac{\delta(x)^2}{\rho(x)^2 \epsilon^2} I_{c-1} \right\|_F,$$

where $\tilde{J}_x$ is the Jacobian matrix of f at x in another coordinate system on Δ^c.
The **regularized loss function** is given as:

$$\mathcal{L}(\hat{\theta}, x, f) = (1 - \eta) H(\hat{\theta} || f(x)) + \eta \alpha(x, f),$$

Regularization methods

Practical issues

- High **variability** form run to run (see next slides).

Regularization methods

Practical issues

- High **variability** form run to run (see next slides).
- High **computational cost** of the regularization term.
  - Maybe using RandNLA. Is it possible to *not* use backpropagation?

Regularization methods

Practical issues

- High **variability** form run to run (see next slides).
- High **computational cost** of the regularization term.
  - Maybe using RandNLA. Is it possible to *not* use backpropagation?
- Drop in benign accuracy.

Regularization methods

Practical issues

- High **variability** form run to run (see next slides).
- High **computational cost** of the regularization term.
  - Maybe using RandNLA. Is it possible to *not* use backpropagation?
- Drop in benign accuracy.
- Find a metric to beat adversarial training.
  - We must provide an experimental validation for why ISO is “better" than adversarial training $\rightarrow$ measure the robustness differently (maybe with Gaussian noise?).

Regularization methods

High variability

At first, I was thinking that ISO was very sensitive to the hyperparameter η.

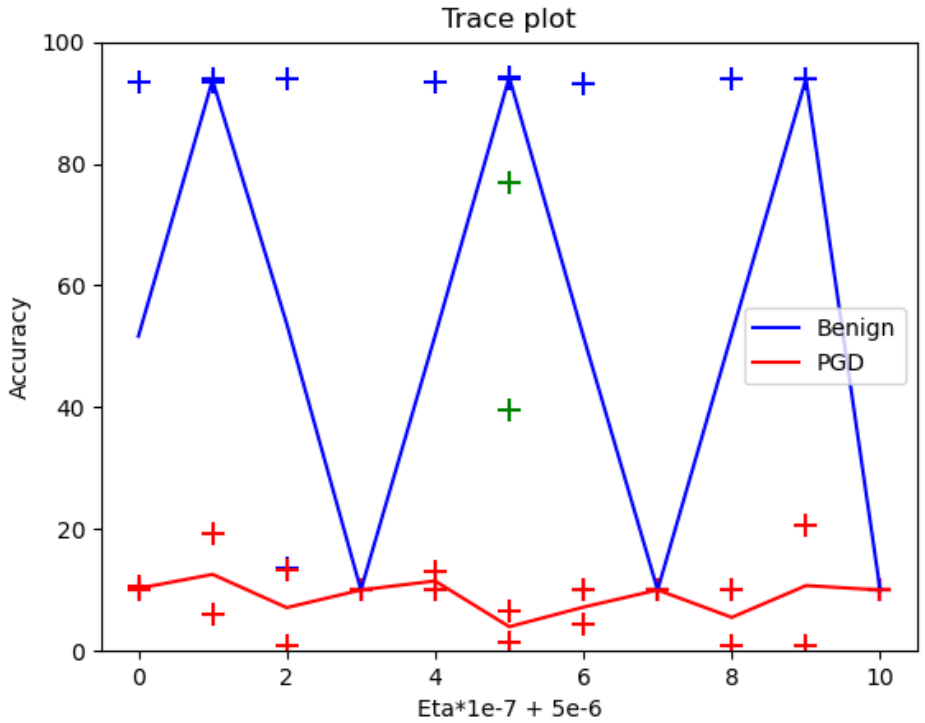

Figure 1: Trace plot for $\eta \in [5.10^{-6}, 6.10^{-6}]$. Two runs per values. In green, the accuracy of a robust model.

But it seems that the performance of the model is independent of η (for the small range considered here).

Regularization methods

High variability

In fact, ISO seems to be very sensitive to the **initialization**.

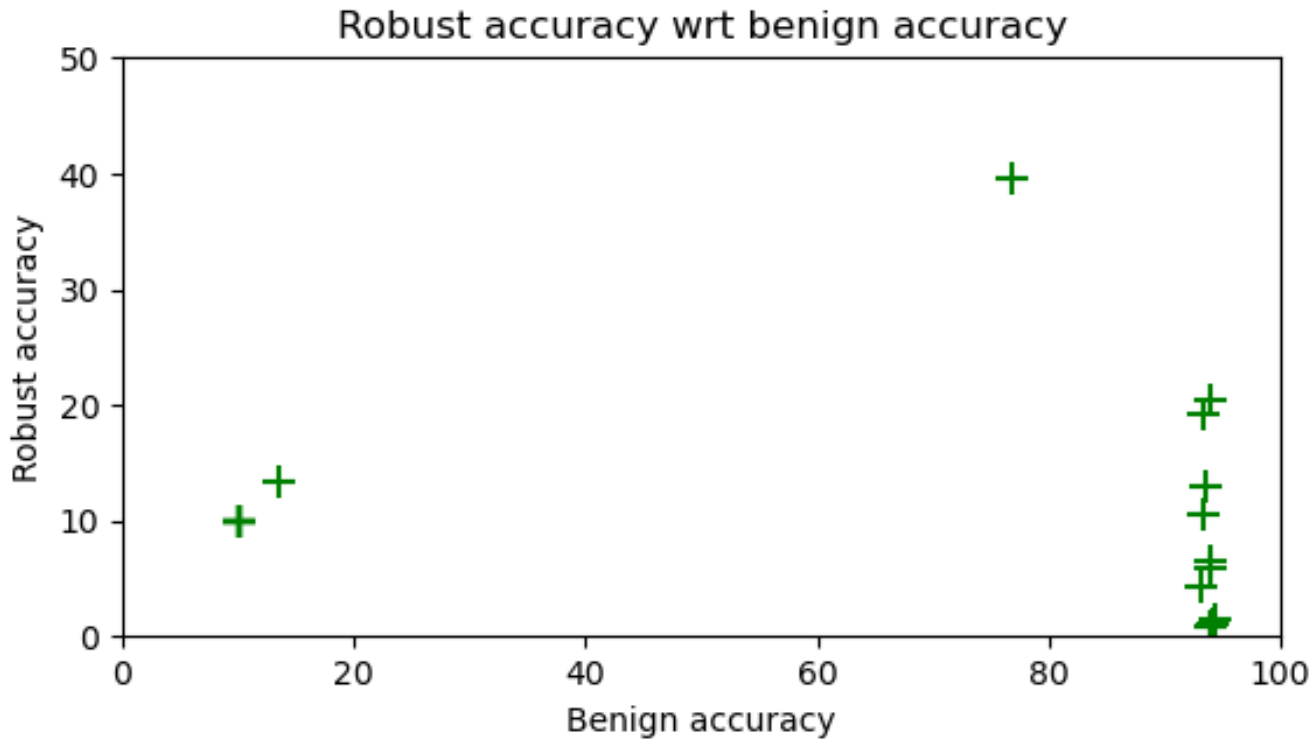

Figure 2: Robust accuracy wrt benign accuracy for the same runs as in Figure 1.

There are 3 clusters. It may correspond to 3 different **local minima** of the regularized loss (?)

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.
  - More importantly, Parseval does not suffer from the high variability issue (and doesn't have any hyperparameter).

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.
  - More importantly, Parseval does not suffer from the high variability issue (and doesn't have any hyperparameter).
  - Parseval seems similar to ISO, but on a layer-per-layer basis. *Can we do the same with ISO?*

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.
  - More importantly, Parseval does not suffer from the high variability issue (and doesn't have any hyperparameter).
  - Parseval seems similar to ISO, but on a layer-per-layer basis. *Can we do the same with ISO?*
  - Parseval separates the learning from the regularization: 1 step of gradient wrt cross-entropy, then 1 step of gradient wrt regularizer.

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.
  - More importantly, Parseval does not suffer from the high variability issue (and doesn't have any hyperparameter).
  - Parseval seems similar to ISO, but on a layer-per-layer basis. *Can we do the same with ISO?*
  - Parseval separates the learning from the regularization: 1 step of gradient wrt cross-entropy, then 1 step of gradient wrt regularizer.
  - Parseval enforces robustness on every inputs, while ISO enforces robustness only on the training examples.

Regularization methods

Other issues

- Comparison with **Parseval** networks.
  - The experiments on MNIST have shown that Parseval is a little better than ISO.
  - More importantly, Parseval does not suffer from the high variability issue (and doesn't have any hyperparameter).
  - Parseval seems similar to ISO, but on a layer-per-layer basis. *Can we do the same with ISO?*
  - Parseval separates the learning from the regularization: 1 step of gradient wrt cross-entropy, then 1 step of gradient wrt regularizer.
  - Parseval enforces robustness on every inputs, while ISO enforces robustness only on the training examples.
- Why is the **Jacobian regularization** not working?
  - ISO $\approx$ enforcing local Lipschitz constant close to 1 works.
  - JAC $\approx$ enforcing local Lipschitz constant *smaller than 1*.
  - JAC should achieve the same robustness with less constraints.

Regularization methods

Other issues

- What are the benefits of the **FIM**, *really?*

Regularization methods

Other issues

- What are the benefits of the **FIM**, *really?*
  - We don't rely on the statistical "meaning" of Δ^c.

Regularization methods

Other issues

- What are the benefits of the **FIM**, *really?*
  - We don't rely on the statistical "meaning" of Δ^c.
  - With the FIM, Δ^c is a portion of a sphere. With the Euclidean metric, Δ^c is a portion of a plane. Is the difference really significant?

Regularization methods

Other issues

- What are the benefits of the **FIM**, *really?*
  - We don't rely on the statistical "meaning" of Δ^c.
  - With the FIM, Δ^c is a portion of a sphere. With the Euclidean metric, Δ^c is a portion of a plane. Is the difference really significant?
  - However, if we use the Euclidean metric, then ISO is very similar to Parseval, so there is no novelty.

Regularization methods

Other issues

- What are the benefits of the **FIM**, *really?*
  - We don't rely on the statistical "meaning" of Δ^c.
  - With the FIM, Δ^c is a portion of a sphere. With the Euclidean metric, Δ^c is a portion of a plane. Is the difference really significant?
  - However, if we use the Euclidean metric, then ISO is very similar to Parseval, so there is no novelty.
- ISO is not a **certified defense**. Are **empirical defenses** still relevant in the current state of the literature?

Robustness

Ramble n°1

The real challenge for machine learning robustness is to **design a defense** that:

1. is **certifiable**.
2. is **computationally efficient**, i.e., scalable to high-dimensional datasets, and large networks.
3. certifies against any "meaningful" attack, not only l_p but also "**semantic attacks**" (or spatial attacks, unrestricted attacks, Wasserstein attacks etc.).
4. certifies with "large enough" radii.
5. also addresses out-of-distribution examples, distribution shift etc. I feel that all these problems are closely related. It is impossible to understand and "solve" robustness alone.

Or to prove that such method cannot exist.

Randomized smoothing

Motivations

- Randomized smoothing (RS) is a **certification** method.

Randomized smoothing

Motivations

- Randomized smoothing (RS) is a **certification** method.
- RS is claimed to be the only certified method that **scales to large networks and high-dimensional datasets** (i.e., ImageNet).
  However:
  - I've seen papers claiming to achieve certified robust accuracy on ImageNet with other methods, but I don't know if it is comparable.
  - Several papers exhibit limits of RS: computational cost of Monte Carlo sampling, certified radius tends to shrink in high dimension, or to be biased against particular classes.

Randomized smoothing

Motivations

- Randomized smoothing (RS) is a **certification** method.
- RS is claimed to be the only certified method that **scales to large networks and high-dimensional datasets** (i.e., ImageNet). However:
  - I've seen papers claiming to achieve certified robust accuracy on ImageNet with other methods, but I don't know if it is comparable.
  - Several papers exhibit limits of RS: computational cost of Monte Carlo sampling, certified radius tends to shrink in high dimension, or to be biased against particular classes.
- RS is **stochastic**. Thus I think it is a more natural application of **information geometry** than studying deterministic networks (and pretending to work with probability distributions).

Randomized smoothing

Motivations

- Randomized smoothing (RS) is a **certification** method.
- RS is claimed to be the only certified method that **scales to large networks and high-dimensional datasets** (i.e., ImageNet). However:
  - I've seen papers claiming to achieve certified robust accuracy on ImageNet with other methods, but I don't know if it is comparable.
  - Several papers exhibit limits of RS: computational cost of Monte Carlo sampling, certified radius tends to shrink in high dimension, or to be biased against particular classes.
- RS is **stochastic**. Thus I think it is a more natural application of **information geometry** than studying deterministic networks (and pretending to work with probability distributions).
- RS is not limited to a specific type of attacks. It can certify against l_p norms, but also semantic attacks.

Randomized smoothing

Motivations

- Randomized smoothing (RS) is a **certification** method.
- RS is claimed to be the only certified method that **scales to large networks and high-dimensional datasets** (i.e., ImageNet). However:
  - I've seen papers claiming to achieve certified robust accuracy on ImageNet with other methods, but I don't know if it is comparable.
  - Several papers exhibit limits of RS: computational cost of Monte Carlo sampling, certified radius tends to shrink in high dimension, or to be biased against particular classes.
- RS is **stochastic**. Thus I think it is a more natural application of **information geometry** than studying deterministic networks (and pretending to work with probability distributions).
- RS is not limited to a specific type of attacks. It can certify against l_p norms, but also semantic attacks.
- RS informs us about the properties that the base classifier should have to enjoy large certified radii (obtained with RS).

Randomized smoothing

Overview

The seminal paper is Cohen et al., 2019.
It is the first paper to provide tight radius for RS, relying on the **Neyman-Pearson lemma** (NPL). Since then, all papers about RS seem to rely on the NPL. Several variants and extensions of RS have been introduced.

Randomized smoothing

Overview

The seminal paper is Cohen et al., 2019.
It is the first paper to provide tight radius for RS, relying on the **Neyman-Pearson lemma** (NPL). Since then, all papers about RS seem to rely on the NPL. Several variants and extensions of RS have been introduced. Thus, I propose to:

1. Review the NPL (and maybe its connection with information geometry, as there is a paper by Eguchi & Copas).

Randomized smoothing

Overview

The seminal paper is Cohen et al., 2019.
It is the first paper to provide tight radius for RS, relying on the **Neyman-Pearson lemma** (NPL). Since then, all papers about RS seem to rely on the NPL. Several variants and extensions of RS have been introduced. Thus, I propose to:

1. Review the NPL (and maybe its connection with information geometry, as there is a paper by Eguchi & Copas).
2. Review the original RS certified radius proof from Cohen et al. 2019.

Randomized smoothing

Overview

The seminal paper is Cohen et al., 2019.
It is the first paper to provide tight radius for RS, relying on the **Neyman-Pearson lemma** (NPL). Since then, all papers about RS seem to rely on the NPL. Several variants and extensions of RS have been introduced. Thus, I propose to:

1. Review the NPL (and maybe its connection with information geometry, as there is a paper by Eguchi & Copas).
2. Review the original RS certified radius proof from Cohen et al. 2019.
3. Review several variants and extensions of the proof in subsequent works.

Randomized smoothing

Overview

The seminal paper is Cohen et al., 2019.
It is the first paper to provide tight radius for RS, relying on the **Neyman-Pearson lemma** (NPL). Since then, all papers about RS seem to rely on the NPL. Several variants and extensions of RS have been introduced. Thus, I propose to:

1. Review the NPL (and maybe its connection with information geometry, as there is a paper by Eguchi & Copas).
2. Review the original RS certified radius proof from Cohen et al. 2019.
3. Review several variants and extensions of the proof in subsequent works.
4. Synthesize all these variants into a unified framework, then derive new results:
   - How to train the base classifier to achieve large certified radii.
   - How to do efficient Monte Carlo sampling for prediction and certification.
   - How to certify against a large class of attacks including semantic attacks.

Neyman-Pearson Lemma

Let X, Y be two random variables with densities p_X, p_Y and distributions $\mathbb{P}_X, \mathbb{P}_Y$.
Let $W = \left\{ z : \frac{p_Y(z)}{p_X(z)} \geq t \right\}$. Let $\alpha \in (0, 1)$.

Neyman-Pearson Lemma

Let X, Y be two random variables with densities p_X, p_Y and distributions $\mathbb{P}_X, \mathbb{P}_Y$.
Let $W = \left\{ z : \frac{p_Y(z)}{p_X(z)} \geq t \right\}$. Let $\alpha \in (0, 1)$.
Choose $t > 0$ such that $\mathbb{P}_X(W) \geq \alpha$.

Neyman-Pearson Lemma

Let X, Y be two random variables with densities p_X, p_Y and distributions $\mathbb{P}_X, \mathbb{P}_Y$.
Let $W = \left\{ z : \frac{p_Y(z)}{p_X(z)} \geq t \right\}$. Let $\alpha \in (0, 1)$.
Choose $t > 0$ such that $\mathbb{P}_X(W) \geq \alpha$.
Let W^* such that $\alpha \geq \mathbb{P}_X(W^*)$.

Neyman-Pearson Lemma

Let X, Y be two random variables with densities p_X, p_Y and distributions $\mathbb{P}_X, \mathbb{P}_Y$.
Let $W = \left\{ z : \frac{p_Y(z)}{p_X(z)} \geq t \right\}$. Let $\alpha \in (0, 1)$.
Choose $t > 0$ such that $\mathbb{P}_X(W) \geq \alpha$.
Let W^* such that $\alpha \geq \mathbb{P}_X(W^*)$. Then:

$$\begin{aligned}
\mathbb{P}_Y(W) - \mathbb{P}_Y(W^*) &= \mathbb{P}_Y(W \setminus W^*) - \mathbb{P}_Y(W^* \setminus W) \\
&= \int_{W \setminus W^*} p_Y(z)dz - \int_{W^* \setminus W} p_Y(z)dz \\
&\geq t \left(\int_{W \setminus W^*} p_X(z)dz - \int_{W^* \setminus W} p_X(z)dz \right) \\
&\geq t(\mathbb{P}_X(W) - \mathbb{P}_X(W^*)) \\
&\geq 0.
\end{aligned}$$

Neyman-Pearson Lemma

Let X, Y be two random variables with densities p_X, p_Y and distributions $\mathbb{P}_X, \mathbb{P}_Y$.
Let $W = \left\{ z : \frac{p_Y(z)}{p_X(z)} \geq t \right\}$. Let $\alpha \in (0, 1)$.
Choose $t > 0$ such that $\mathbb{P}_X(W) \geq \alpha$.
Let W^* such that $\alpha \geq \mathbb{P}_X(W^*)$. Then:

$$\begin{aligned}
\mathbb{P}_Y(W) - \mathbb{P}_Y(W^*) &= \mathbb{P}_Y(W \setminus W^*) - \mathbb{P}_Y(W^* \setminus W) \\
&= \int_{W \setminus W^*} p_Y(z)dz - \int_{W^* \setminus W} p_Y(z)dz \\
&\geq t \left(\int_{W \setminus W^*} p_X(z)dz - \int_{W^* \setminus W} p_X(z)dz \right) \\
&\geq t(\mathbb{P}_X(W) - \mathbb{P}_X(W^*)) \\
&\geq 0.
\end{aligned}$$

Still true if we replace all "$\geq$" by "$\leq$".

Certified radius using Gaussian noise against l_2 attack