Robustness of classifiers: from adversarial to random noise

Alhussein Fawzi Seyed-Mohsen Moosavi-Dezfooli Pascal Frossard

Ecole Polytechnique Federale de Lausanne

NIPS, 2016

Presenter: Tianlu Wang

Outline

Introduction

- Classifiers are vulnerable to worst-case (adversarial) perturbations of the datapoints.
- Classifiers are relatively robust to random noise.
- Worst-case noise is only a specific type of noise. This paper focuses on a semi-random noise regime that generalizes both the random and worst-case noise regime, provides theoretical bounds on the robustness of classifiers in this general regime.

Background

- Random noise regime: datapoints are perturbed by noise with random direction in the input space.
- Semi-random regime: random subspaces of arbitrary dimension, where a worst-case perturbation is sought within the subspace.
- Well-sought perturbations of the data can easily cause misclassification, because data points lie very close to the decision boundary.

Contribution

- Robustness of classifiers depends on the curvature of the decision boundary: (d denotes the dimension of the data, l is the distance from the datapoint to the classification boundary)
  - Random regime: $robustness = \sqrt{d} \times l$, when curvature is sufficiently small. $\Rightarrow$ In high dimensional classification problems, robustness to random noise can be achieved, even when datapoints are very close to decision boundary.
  - Semi-random regime: $robustness = \sqrt{d/m} \times l$, m is the dimension of the subspace. Even when m is chosen as a small fraction of the dimension d, it is still possible to find small perturbations that cause data misclassification.

Definitions and notations

- An L-class classifier: $f : \mathbb{R}^d \rightarrow \mathbb{R}^L$
- Given an datapoint $x_0 \in \mathbb{R}^d$, $\hat{k}(x_0) = argmax_k f_k(x_0)$
- S is an arbitrary subspace of $\mathbb{R}^d$ of dimension m, r_s^* is the perturbation in S of minimal norm that is required to change the estimated label of f at x_0:

$$r_S^*(x_0) = \underset{r \in S}{\text{argmin}} ||r||_2 s.t. \hat{k}(x_0 + r) \neq \hat{k}(x_0) \quad (1)$$

$$r_S^*(x_0) = \underset{r \in S}{\text{argmin}} ||r||_2 s.t. \exists k \neq \hat{k}(x_0) : f_k(x_0 + r) \geq f_{\hat{k}(x_0)}(x_0 + r) \quad (2)$$

- When $S = \mathbb{R}^d$, $r^*(x_0) := r_{\mathbb{R}^d}^*(x_0)$ is the adversarial perturbation.
- $||r^*(x_0)||$ is the minimal distance from x_0 to the classifier boundary

Definitions and notations

- Random noise regime: S is a one-dimensional subspace ($m = 1$) with direction v, where v is a random vector sampled uniformly from the unit sphere $\mathbb{S}^{d-1}$
- Semi-random noise regime: S is a random space, the span of m independent vectors drawn uniformly at random from $\mathbb{S}^{d-1}$
- In the rest of slides, fix x_0, use r_S^* instead of $r_S^*(x_0)$ and $\hat{k}$ instead of $\hat{k}(x_0)$

Robustness of affine classifiers

Theorem (1)

Let $\delta > 0$ and S be a random m-dimensional subspace of $\mathbb{R}^d$, and f be a L-class affine classifier. Let

$$\zeta_1(m,\delta) = \left(1 + 2\sqrt{\frac{\ln(1/\delta)}{m}} + \frac{2\ln(1/\delta)}{m}\right)^{-1} \tag{3}$$

$$\zeta_2(m,\delta) = \left(\max\left((1/e)\delta^{2/m}, 1 - \sqrt{2(1-\delta^{2/m})}\right)\right)^{-1} \tag{4}$$

The following inequalities hold between the robustness to semi-random noise $||r_S^*||_2$, and the robustness to adversarial perturbations $||r^*||_2$:

$$\sqrt{\zeta_1(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2 \leq ||r_S^*||_2 \leq \sqrt{\zeta_2(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2 \tag{5}$$

with probability exceeding $1 - 2(L+1)\delta$.

Robustness of affine classifiers

- $\zeta_1(m, \delta)$ and $\zeta_2(m, \delta)$ are independent of the data dimension d
- For sufficiently large m, $\zeta_1(m, \delta)$ and $\zeta_2(m, \delta)$ are very close to 1 but the difference is much larger when $m = 1$

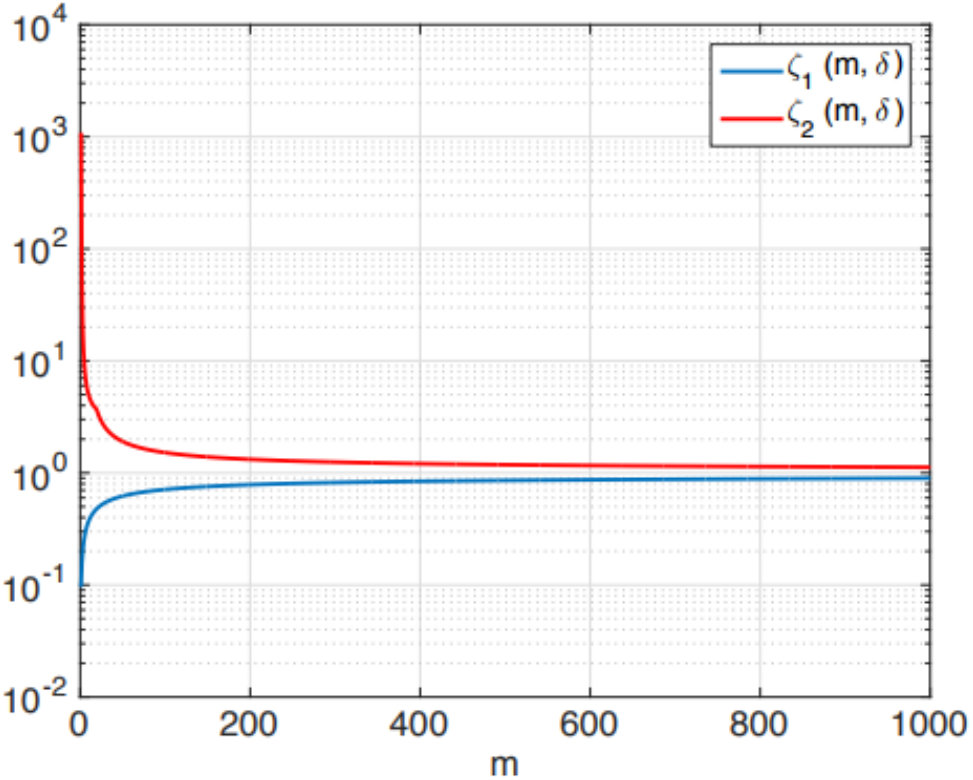


Figure 1: $\zeta_1(m, \delta)$ and $\zeta_2(m, \delta)$ in function of m $[\delta = 0.05]$.

Robustness of affine classifiers

- $\sqrt{\zeta_1(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2 \leq ||r_S^*||_2 \leq \sqrt{\zeta_2(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2$
- Random noise regime($m = 1$): $\Theta(\sqrt{d}||r^*||_2)$, in high dimensional classification, affine classifiers can be robust to random noise
- Semi-random regime: when m is sufficiently large, $||r_S^*||_2 \approx \sqrt{d/m}||r^*||_2$(because $\zeta_1(m,\theta) \approx \zeta_2(m,\theta) \approx 1$). Semi-random robustness can remain small even m is chosen to be a very small faction of d.
- Conclusion: for semi-random noise that is mostly random and mildly adversarial, affine classifiers remain vulnerable to such noise. (When $m = 0.01d$, semi-random robustness is only $10||r^*||_2$ with high probability)

Outline

Decision boundary curvature

- Pairwise boundary $\mathcal{B}_{i,j}$ as the boundary of binary classifier where only class i and class j are considered: $\mathcal{B}_{i,j} = \{x \in \mathbb{R}^d : f_i(x) - f_j(x) = 0\}$
- $\mathcal{B}_{i,j}$ seperates two regions of $\mathbb{R}^d$:

$$\mathcal{R}_i = \{x \in \mathbb{R}^d : f_i(x) > f_j(x)\}$$

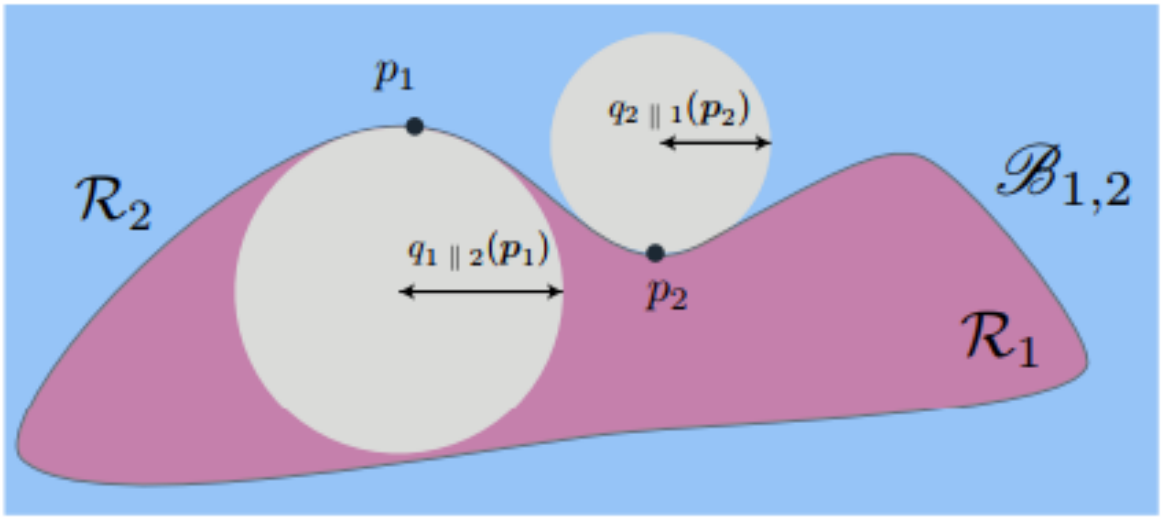
$$\mathcal{R}_j = \{x \in \mathbb{R}^d : f_j(x) > f_i(x)\}$$

Decision boundary curvature

- Curvature: global bending of the decision boundary by inscribing balls in the regions separated by the decision boundary. For a given $\mathbf{p} \in \mathcal{B}_{i,j}$, we define $q_{i||j}(\mathbf{p})$ to be the radius of the largest open ball included in the region $\mathcal{R}_i$ that intersects with $\mathcal{B}_{i,j}$ at $\mathbf{p}$; i.e.,

$$q_{i||j}(\mathbf{p}) = \sup_{\mathbf{z} \in \mathbb{R}^d} \left\{ \|\mathbf{z} - \mathbf{p}\|_2 : \mathbb{B}(\mathbf{z}, \|\mathbf{z} - \mathbf{p}\|_2) \subseteq \mathcal{R}_i \right\}, \tag{6}$$

where $\mathbb{B}(\mathbf{z}, \|\mathbf{z} - \mathbf{p}\|_2)$ is the open ball in $\mathbb{R}^d$ of center $\mathbf{z}$ and radius $\|\mathbf{z} - \mathbf{p}\|_2$.



Decision boundary curvature

- $q_{i||j}(\mathbf{p}) \neq q_{j||i}(\mathbf{p})$ as the radius of the largest ball one can inscribe in both regions need not be equal. So define a symmetric quantity $q_{i,j}(\mathbf{p}) = min(q_{i||j}(\mathbf{p}), q_{j||i}(\mathbf{p}))$
- To measuer the global curvature, the worst-case radius is taken over all points on the decision boundary, i.e.,

$$q(\mathcal{B}_{i,j}) = \inf_{\mathbf{p} \in \mathcal{B}_{i,j}} q_{i,j}(\mathbf{p}), \tag{7}$$

$$\kappa(\mathcal{B}_{i,j}) = \frac{1}{q(\mathcal{B}_{i,j})}. \tag{8}$$

- The curvature $\kappa(\mathcal{B}_{i,j})$ is simply defined as the inverse of the worst-case radius over all points $\mathbf{p}$ on the decision boundary

Decision boundary curvature

- Affine classifiers: $\kappa(\mathcal{B}_{i,j}) = 0$, as it is possible to inscribe balls of infinite radius inside each region of the space
- In general, the quantity $\kappa(\mathcal{B}_{i,j})$ provides an intuitive way of describing the nonlinearity of the decision boundary by fitting balls inside the classification regions.
- A precise characterization of the robustness to semi-random and random noise of nonlinear classifiers in terms of the curvature of the decision boundaries $\kappa(\mathcal{B}_{i,j})$.

Outline

Binary classification

- First study binary classification problem, where only classes $\hat{k}$ and $k \in \{1, \ldots, L\} \backslash \{\hat{k}\}$ are considered
- $\mathcal{B}_k := \mathcal{B}_{k,\hat{k}}$ is the decision boundary between class k and $\hat{k}$,

$$r_S^k = \underset{r \in S}{\operatorname{argmin}} ||r||_2 \text{ s.t. } f_k(x_0 + r) \geq f_{\hat{k}}(x_0 + r), \tag{9}$$

$$r^k = \underset{r}{\operatorname{argmin}} ||r||_2 \text{ s.t. } f_k(x_0 + r) \geq f_{\hat{k}}(x_0 + r). \tag{10}$$

- The global quatities r_s^* and r^* are obtained from r_s^k and r^k by taking the vectors with minimum norm over all classes k

Binary classfification

Theorem (2)

Let $\mathcal{S}$ be a random m-dimensional subspace of $\mathbb{R}^d$. Let $\kappa := \kappa(\mathcal{B}_k)$. Assuming that the curvature satisfies

$$\kappa \leq \frac{C}{\zeta_2(m,\delta)||r^k||_2}\frac{m}{d},$$

the following inequality holds between the semi-random robustness $||r^k_{\mathcal{S}}||_2$ and the adversarial robustness $||r^k||_2$:

$$\left(1 - C_1||r^k||_2\kappa\zeta_2(m,\delta)\frac{d}{m}\right)\sqrt{\zeta_1(m,\delta)}\sqrt{\frac{d}{m}} \leq \frac{||r^k_{\mathcal{S}}||_2}{||r^k||_2} \leq \left(1 + C_2||r^k||_2\kappa\zeta_2(m,\delta)\frac{d}{m}\right)\sqrt{\zeta_2(m,\delta)}\sqrt{\frac{d}{m}} \tag{11}$$

with probability larger than $1 - 4\delta$. The constants are $C = 0.2, C_1 = 0.625, C_2 = 2.25$.

Binary classification

- Bounds relating the robustness to random and semi-random noise to the worst-case robustness can be extended to nonlinear classifiers, provided the curvature of the boundary $\kappa(\mathcal{B}_k)$ is sufficiently small.
- In the case of linear classifiers, we have $\kappa(\mathcal{B}_k) = 0$, and we recover the result for affine classifiers from Theorem 1.

Multi-class classification

- To extend this result to multi-class classification,if k denotes a class that has no boundary with class $\hat{k}$, we have $||r^k||_2 = \infty$, and the previous curvature condition cannot be satisfied.
- It is therefore crucial to *exclude* such classes that have no boundary in common with class $\hat{k}$, or more generally, boundaries that are far from class *lab*. We define the set A of excluded classes k where $||r^k||_2$ is large

$$A = \{k : ||r^k||_2 \geq 1.45\sqrt{\zeta_2(m,\theta)}\sqrt{\frac{d}{m}}||r^*||_2\}. \tag{12}$$

Note that A is independent of $\mathcal{S}$, and depends only on d, m and δ.

Multi-class classification

Corollary

Let $\mathcal{S}$ be a random m-dimensional subspace of $\mathbb{R}^d$. Assume that, for all $k \notin A$, we have

$$\kappa(\mathcal{B}_k)||r^k||_2 \leq \frac{0.2}{\zeta_2(m,\delta)}\frac{m}{d} \tag{13}$$

Then, we have

$$0.875\sqrt{\zeta_1(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2 \leq ||r^*_{\mathcal{S}}||_2 \leq 1.45\sqrt{\zeta_2(m,\delta)}\sqrt{\frac{d}{m}}||r^*||_2 \tag{14}$$

with probability larger than $1 - 4(L+2)\delta$.

Multi-class classification

- $||r^*_{\mathcal{S}}||_2$ is precisely related to the adversarial robustness $||r^*||_2$ by a factor of $\sqrt{d/m}$.
- Random regime ($m = 1$): factor $\sqrt{d}$ shows that in high dimensional classification problems, classifiers with sufficiently flat boundaries are much more robust to random noise than to adversarial noise.The addition of a sufficiently small random noise does not change the label of the image, even if the image lies very closely to the decision boundary (i.e., $||r^*||_2$ is small).
- Semi-random regime: an adversarial perturbation is found on a randomly chosen subspace of dimension m, the $\sqrt{d/m}$ factor shows that robustness to semi-random noise might not be achieved even if m is chosen to be a tiny fraction of d (e.g., $m = 0.01d$).If a classifier is highly vulnerable to adversarial perturbations, then it is also vulnerable to noise that is overwhelmingly random and only mildly adversarial (i.e. worst-case noise sought in a random subspace of low dimensionality m).

Outline

Estimation of robustness

- Theoretical results show that the robustness $||r^*_{\mathcal{S}}(x)||_2$ of classifiers satisfying the curvature property precisely behaves as $\sqrt{d/m}||r^*(x)||_2$.
- Define

$$\beta(f; m) = \sqrt{m/d}\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\frac{||r^*_{\mathcal{S}}(x)||_2}{||r^*(x)||_2},$$

where $\mathcal{S}$ is chosen randomly for each sample x and $\mathcal{D}$ denotes the test set. β should ideally be equal to 1 (for sufficiently large m).

	m/d					
Classifier	1	1/4	1/16	1/36	1/64	1/100
LeNet (MNIST)	1.00	1.00 ± 0.06	1.01 ± 0.12	1.03 ± 0.20	1.01 ± 0.26	1.05 ± 0.34
LeNet (CIFAR-10)	1.00	1.01 ± 0.03	1.02 ± 0.07	1.04 ± 0.10	1.06 ± 0.14	1.10 ± 0.19
VGG-F (ImageNet)	1.00	1.00 ± 0.01	1.02 ± 0.02	1.03 ± 0.04	1.03 ± 0.05	1.04 ± 0.06
VGG-19 (ImageNet)	1.00	1.00 ± 0.01	1.02 ± 0.03	1.02 ± 0.05	1.03 ± 0.06	1.04 ± 0.08

Table 1: $\beta(f; m)$ for different classifiers f and different subspace dimensions m. The VGG-F and VGG-19 are respectively introduced in [2, 17].

Estimation of robustness

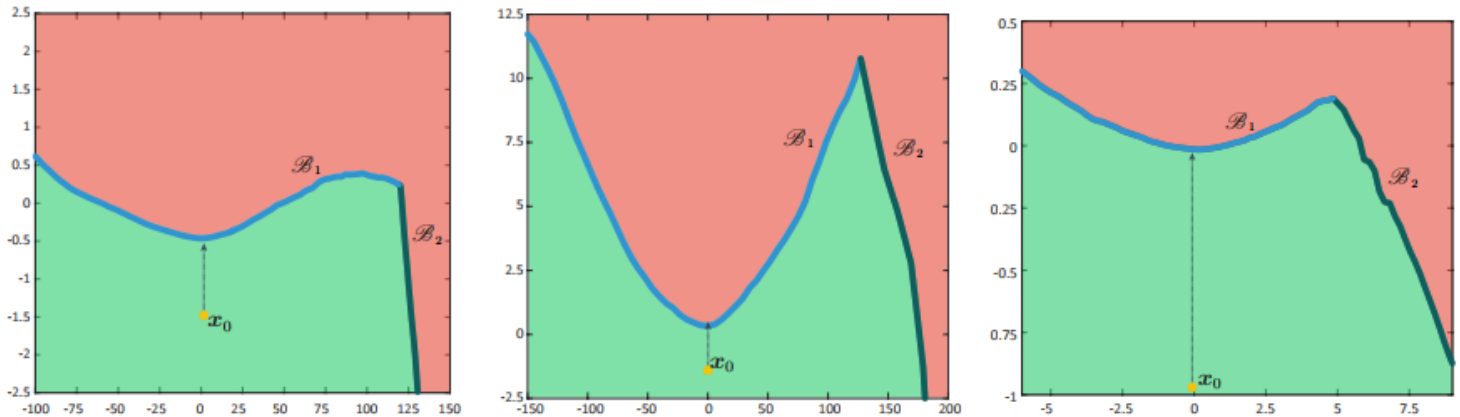
- β is suprisingly close to 1, even when m is a small fraction of d $\Rightarrow$ quantitative analysis provide very accurate estimates of the robustness to semi-random noise.

(a) (b) (c) (d)

Figure 5: (a) Original image classified as "Cauliflower". Fooling perturbations for VGG-F network: (b) Random noise, (c) Semi-random perturbation with $m = 10$, (d) Worst-case perturbation, all wrongly classified as "Artichoke".

Outline

- Table 1 suggests that the decision boundaries of these classifiers have limited curvature $\kappa(\mathcal{B}_k)$, as this is a key assumption of the theoretical findings. So visualize two-dimensional sections of the classifiers' boundary in three different settings.



(a) VGG-F (ImageNet) (b) LeNet (CIFAR) (c) LeNet (MNIST)

Figure 6: Boundaries of three classifiers near randomly chosen samples. Axes are normalized by the corresponding $\|\boldsymbol{r}^*\|_2$ since our assumption in the theoretical bound (Corollary 1) depends on the product of $\|\boldsymbol{r}^*\|_2\kappa$. Note the difference in range between x and y axes. Note also that the range of horizontal axis in (c) is much smaller than the other two, hence the illustrated boundary is more curved.

- Curvature assumption holds in practice, the curvature of such classifiers is very small.

Outline

Demonstration

- $\mathcal{S}$ is the span of random translated and scaled versions of words "NIPS", "SPAIN" and "2016" in an image, such that $\lfloor (d/m) \rfloor = 228$. The resulting perturbations in the subspace are therefore linear combinations of these words with different intensities. Imperceptibly small structured messages can be added to an image causing data misclassification.

(a) Image of a "Potflower"

(b) Structured perturbation containing random placement of words "NIPS", "2016", and "SPAIN"

(c) Classified as "Pineapple"

Figure 7: A fooling hidden message, $\mathcal{S}$ consists of linear combinations of random words.

Summary

- Precisely characterize the robustness of classifiers in a novel semi-random noise regime that generalizes the random noise regime. Bounds depend on the *curvature* of the decision boundary, the data dimension, and the dimension of the subspace to which the perturbation belongs.
- When the decision boundary has a small curvature, classifiers are robust to random noise in high dimensional classification problems (even if the robustness to adversarial perturbations is relatively small). Moreover, for semi-random noise that is mostly random and only mildly adversarial (i.e., the subspace dimension is small), state-of-the-art classifiers remain vulnerable to such perturbations.