Exercise 12

Randomized Smoothing

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1. In this problem, we want to proof the following theorem:

Theorem 1 (From [1])**.** *Let $f : \mathbb{R}^d \to \mathcal{Y}$ be any deterministic or random function. Let $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, \sigma^2 I)$. Let $g(\boldsymbol{x}) = \operatorname{argmax}_c \mathbb{P}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}) = c)$. Suppose that for a specific $\boldsymbol{x} \in \mathbb{R}^d$, there exist $c_A \in \mathcal{Y}$ and $\underline{p_A}, \overline{p_B} \in [0, 1]$ such that:*

$$\mathbb{P}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}) = c_A) \geq \underline{p_A} \geq \overline{p_B} \geq \max_{c \neq c_A} \mathbb{P}(f(\boldsymbol{x}+\boldsymbol{\varepsilon}) = c) \tag{1}$$

Then $g(\boldsymbol{x}+\boldsymbol{\delta}) = c_A$ for all $\|\boldsymbol{\delta}\|_2 < R$, where

$$R = \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B})). \tag{2}$$

The proof is broken down into the Steps (1) - (4). We (1) decompose the input space into two half spaces, A and B, such that the probabilities for the samples from the non-displaced gaussian $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2 I)$ to lie in A or B are $\underline{p_A}$ and $\overline{p_B}$ respectively. This induces a linear separation between the two gaussians (the worst case).

We (2) use Lemma 1 to relate the probabilities of the displaced gaussian $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{x} + \boldsymbol{\delta}, \sigma^2 I)$ to observe class c_A or c_B to the probabilities that $Y \in A$ or $Y \in B$ respectively.

Then in (3), we show how these probabilities relate to $\boldsymbol{\delta}$ and σ. Finally, in (4) we obtain a condition on $\|\boldsymbol{\delta}\|$ such that the classification is robust for all $\boldsymbol{\delta}$ satisfying $\|\boldsymbol{\delta}\| < R$.

1. Show that $\mathbb{P}(\boldsymbol{X} \in A) = \underline{p_A}$ and $\mathbb{P}(\boldsymbol{X} \in B) = \overline{p_B}$, where $\boldsymbol{X} := \boldsymbol{x} + \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{x}, \sigma^2 I)$ and

$$A := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{\delta}^T(\boldsymbol{z} - \boldsymbol{x}) \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})\}$$
$$B := \{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{\delta}^T(\boldsymbol{z} - \boldsymbol{x}) \geq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(1 - \overline{p_B})\}.$$

   Hint: Let $x \sim \mathcal{N}(\mu_x, \sigma_x^2)$ and $y \sim \mathcal{N}(\mu_y, \sigma_y^2)$. Then $x + y \sim \mathcal{N}(\mu_x + \mu_y, \sigma_x^2 + \sigma_y^2)$ and $cx \sim \mathcal{N}(c\mu_x, c^2\sigma_x^2)$.

2. Use Lemma 1 (see the bonus Problem 2), the results from sub-task 1 and the assumptions of the theorem to show

$$\mathbb{P}(f(\boldsymbol{Y}) = c_A) \geq \mathbb{P}(\boldsymbol{Y} \in A) \quad \text{and} \quad \mathbb{P}(f(\boldsymbol{Y}) = c_B) \leq \mathbb{P}(\boldsymbol{Y} \in B),$$

where $\boldsymbol{Y} := (\boldsymbol{x} + \boldsymbol{\delta} + \boldsymbol{\varepsilon}) \sim \mathcal{N}(\boldsymbol{x} + \boldsymbol{\delta}, \sigma^2 I)$.

3. Show that $\mathbb{P}(\boldsymbol{Y} \in A) = \Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\boldsymbol{\delta}\|}{\sigma}\right)$ and $\mathbb{P}(\boldsymbol{Y} \in B) = \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\boldsymbol{\delta}\|}{\sigma}\right)$.

Hint: Let $z \sim \mathcal{N}(0, \sigma^2)$. Then $(z + \mu) \sim \mathcal{N}(\mu, \sigma^2)$.

4. Find the condition for $\boldsymbol{\delta}$ such that $\mathbb{P}(\boldsymbol{Y} \in A) > \mathbb{P}(\boldsymbol{Y} \in B)$ holds.

Problem (opt.) 2. In this task we will prove the following Lemma:

Lemma 1 (Special case of Neyman-Pearson). *Let $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I)$, $\boldsymbol{Y} \sim \mathcal{N}(\boldsymbol{\mu} + \boldsymbol{\delta}, \sigma^2 I)$, $f : \mathbb{R}^d \to \mathcal{Y}$ a deterministic or random function and $c \in \mathcal{Y}$. Then:*

1. If $S = \left\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{\delta}^T \boldsymbol{z} \leq t\right\}$ for some t and $\mathbb{P}(f(\boldsymbol{X}) = c) \geq \mathbb{P}(\boldsymbol{X} \in S)$, then $\mathbb{P}(f(\boldsymbol{Y}) = c) \geq \mathbb{P}(\boldsymbol{Y} \in S)$

2. If $S = \left\{\boldsymbol{z} \in \mathbb{R}^d : \boldsymbol{\delta}^T \boldsymbol{z} \geq t\right\}$ for some t and $\mathbb{P}(f(\boldsymbol{X}) = c) \leq \mathbb{P}(\boldsymbol{X} \in S)$, then $\mathbb{P}(f(\boldsymbol{Y}) = c) \leq \mathbb{P}(\boldsymbol{Y} \in S)$.

Let

$$g(\boldsymbol{x}; \boldsymbol{\mu}, \sigma^2 I) := \frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{x} - \boldsymbol{\mu})^T(\boldsymbol{x} - \boldsymbol{\mu})\right)$$

denote the Gaussian probability density function (for mean $\boldsymbol{\mu}$ and co-variance matrix $\sigma^2 I$) evaluate at $\boldsymbol{x}$. For convenience we write $g_{\boldsymbol{X}}(\boldsymbol{x}) := g(\boldsymbol{x}; \boldsymbol{\mu}_{\boldsymbol{X}}, \sigma^2_{\boldsymbol{X}} I)$ for Gaussian Random Variables $\boldsymbol{X} \sim \mathcal{N}(\mu_{\boldsymbol{X}}, \sigma^2_{\boldsymbol{X}} I)$. Further, we let A^C denote the complement of a set A over $\mathbb{R}^d$, $A^C := \mathbb{R}^d \setminus A$. Then we can trivially decompose

$$\mathbb{R}^d = A \cup A^C. \tag{3}$$

1. Compute and simplify $m(\boldsymbol{z}) := \frac{g_{\boldsymbol{Y}}(\boldsymbol{z})}{g_{\boldsymbol{X}}(\boldsymbol{z})}$.

2. Show that for any t there exists a $t' > 0$ such that $\left\{z \in \mathbb{R}^d : \boldsymbol{\delta}^T \boldsymbol{z} \leq t\right\} = \{\boldsymbol{z} \in \mathbb{R}^d : m(\boldsymbol{z}) \leq t'\}$.

3. Given $S := \{\boldsymbol{z} \in \mathbb{R}^d : m(\boldsymbol{z}) \leq t'\}$ show that

$$\left(\int_{S^C} g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} - \int_S g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\right) \geq t' \left(\int_{S^C} g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z} - \int_S g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}\right).$$

Extend this to $\left(\int_{S^C}[f(\boldsymbol{z}) = c] \cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} - \int_S [f(\boldsymbol{z}) \neq c] \cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\right)$
$\geq t' \left(\int_{S^C}[f(\boldsymbol{z}) = c] \cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z} - \int_S [f(\boldsymbol{z}) \neq c] \cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}\right)$.

4. Let $f : \mathbb{R}^d \to Y$ be a classifier (function) and $c \in Y$ a class. Show that for $S := \{\boldsymbol{z} \in \mathbb{R}^d : m(\boldsymbol{z}) \leq t'\}$ for a given $t' > 0$ and $\mathbb{P}(f(\boldsymbol{X}) = c) \geq \mathbb{P}(\boldsymbol{X} \in S)$, then $\mathbb{P}(f(\boldsymbol{Y}) = c) \geq \mathbb{P}(\boldsymbol{Y} \in S)$. Hint: Show that $\mathbb{P}(f(\boldsymbol{Y}) = c) - \mathbb{P}(\boldsymbol{Y} \in S) \leq 0$ and use the results from the previous tasks.

5. Putting the previous tasks together gives you the proof for the first part of Lemma 1. What changes are required for the second part?

Problem 3. Randomized smoothing currently is usually formulated for the ℓ^1 or ℓ^2-norm. However, well-known equalities can be used to bound norms other than the one guaranteed by the method used; for an example see [2] which uses the l_2-norm to obtain ℓ^∞-bounds. In the following we will show different useful inequalities. Show the inequality and provide the tightest value of c you can find. (You don't need to prove the tightness, although you easily can through an example). Hint: To obtain the tightest bounds you might need to use an additional theorem such as the subadditivity of the square root function or the Cauchy-Schwartz inequality. As a reminder:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{d} |x_i| \qquad \|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^{d} |x_i|^2} \qquad \|\boldsymbol{x}\|_\infty = \max_{i \in \{1,\dots,d\}} |x_i| \qquad \text{for } \boldsymbol{x} \in \mathbb{R}^d$$

1. Show $\|\boldsymbol{x}\|_\infty \leq c_1 \|\boldsymbol{x}\|_1$.

2. Show $\|\boldsymbol{x}\|_1 \leq c_2 \|\boldsymbol{x}\|_\infty$.

3. Show $\|\boldsymbol{x}\|_\infty \leq c_3 \|\boldsymbol{x}\|_2$.

4. Show $\|\boldsymbol{x}\|_2 \leq c_4 \|\boldsymbol{x}\|_\infty$.

5. Show $\|\boldsymbol{x}\|_2 \leq c_5 \|\boldsymbol{x}\|_1$.

6. Show $\|\boldsymbol{x}\|_1 \leq c_6 \|\boldsymbol{x}\|_2$.

7. Let $\mathbb{B}_\epsilon^p := \{x \in \mathbb{R}^d \mid \|x\|_p \leq \epsilon\}$ denote the l_p-norm ball of size e. Order $\mathbb{B}_1^1, \mathbb{B}_1^2, \mathbb{B}_1^\infty, \mathbb{B}_d^1, \mathbb{B}_{\sqrt{d}}^2$ with respect to the inclusion relation $\subseteq$.

References

[1] Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. "Certified Adversarial Robustness via Randomized Smoothing". In: *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*. Ed. by Kamalika Chaudhuri and Ruslan Salakhutdinov. Vol. 97. Proceedings of Machine Learning Research. PMLR, 2019, pp. 1310–1320. URL: http://proceedings.mlr.press/v97/cohen19c.html.

[2] Hadi Salman et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers". In: *Advances in Neural Information Processing Systems 32*. Ed. by H. Wallach et al. Curran Associates, Inc., 2019, pp. 11289–11300. URL: http://papers.nips.cc/paper/9307-provably-robust-deep-learning-via-adversarially-trained-smoothed-classifiers.pdf.