Exercise 12 - Solution

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.

Solution 1. 1. We start by applying the definitions and obtain

$$\mathbb{P}(\boldsymbol{X} \in A) = \mathbb{P}(\boldsymbol{\delta}^T(\boldsymbol{X} - \boldsymbol{x}) \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})).$$

Next we replace $\boldsymbol{X} - \boldsymbol{x} \sim \mathcal{N}(0, \sigma^2 I)$ by $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I)$ and get

$$\mathbb{P}(\boldsymbol{X} \in A) = \mathbb{P}(\boldsymbol{\delta}^T \boldsymbol{Z} \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})).$$

Here, $\boldsymbol{\delta}^T \boldsymbol{Z} = \sum_i \boldsymbol{\delta}_i z_i$ where $z_i \sim \mathcal{N}(0, \sigma^2)$. Using the standard rule for multiplying the normally distributed gaussian variable z_i by a constant $\boldsymbol{\delta}_i$, we get $\boldsymbol{\delta}_i z_i \sim \mathcal{N}(0, \boldsymbol{\delta}_i^2 \sigma^2)$. By applying the sum rule to $\boldsymbol{\delta}_i z_i$ we get $\sum_i \boldsymbol{\delta}_i z_i \sim \mathcal{N}(0, \sigma^2 \sum_i \boldsymbol{\delta}_i^2)$. Further, with $\mathcal{N}(0, \sigma^2) = \sigma\mathcal{N}(0, 1)$ and $\sum_i \boldsymbol{\delta}_i^2 = \|\boldsymbol{\delta}\|^2$ we get

$$\begin{aligned}
\mathbb{P}(\boldsymbol{X} \in A) &= \mathbb{P}(\sigma\|\boldsymbol{\delta}\| z \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})) && (z \sim \mathcal{N}(0, 1)) \\
&= \mathbb{P}(z \leq \Phi^{-1}(\underline{p_A})) \\
&= \Phi(\Phi^{-1}(\underline{p_A})) \\
&= \underline{p_A}.
\end{aligned}$$

Similarly for $\mathbb{P}(\boldsymbol{X} \in B) = \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\boldsymbol{\delta}\|}{\sigma}\right)$.

2. Using the definitions for $\boldsymbol{X}$ and $\boldsymbol{Y}$ together with (Eq. (1)), we see that

$$\mathbb{P}(f(\boldsymbol{X}) = c_A) \geq \underline{p_A} \quad \text{and} \quad \mathbb{P}(f(\boldsymbol{X}) = c_B) \leq \overline{p_B}.$$

Applying Lemma 1 directly yields

$$\begin{aligned}
\mathbb{P}(f(\boldsymbol{Y}) = c_A) &\geq \mathbb{P}(\boldsymbol{Y} \in A), \\
\mathbb{P}(f(\boldsymbol{Y}) = c_B) &\leq \mathbb{P}(\boldsymbol{Y} \in B).
\end{aligned}$$

3. The calculation is similar to the one executed in sub-task 1:

$$\mathbb{P}(\boldsymbol{Y} \in A) = \mathbb{P}(\boldsymbol{\delta}^T(\boldsymbol{Y} - \boldsymbol{x}) \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})).$$

Here we replace $\boldsymbol{Y} - \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\delta}, \sigma^2 I)$ by $\boldsymbol{Z} + \boldsymbol{\delta}$, where $\boldsymbol{Z} \sim \mathcal{N}(0, \sigma^2 I)$. Thus $\boldsymbol{\delta}^T(\boldsymbol{Y} - \boldsymbol{x}) = \boldsymbol{\delta}^T Z + \|\delta\|^2$:

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y} \in A) &= \mathbb{P}(\boldsymbol{\delta}^T\boldsymbol{Z} + \|\boldsymbol{\delta}\|^2 \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A})) \\ &= \mathbb{P}(\boldsymbol{\delta}^T\boldsymbol{Z} \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A}) - \|\boldsymbol{\delta}\|^2).\end{aligned}$$

Following the same steps for the term $\boldsymbol{\delta}^T\boldsymbol{Z}$ as in sub-task 1, we get

$$\begin{aligned}\mathbb{P}(\boldsymbol{Y} \in A) &= \mathbb{P}(\sigma\|\boldsymbol{\delta}\|z \leq \sigma\|\boldsymbol{\delta}\|\Phi^{-1}(\underline{p_A}) - \|\boldsymbol{\delta}\|^2) && (z \sim \mathcal{N}(0,1)) \\ &= \mathbb{P}\left(z \leq \Phi^{-1}(\underline{p_A}) - \frac{\|\boldsymbol{\delta}\|}{\sigma}\right) \\ &= \Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\boldsymbol{\delta}\|}{\sigma}\right).\end{aligned}$$

Similarly for $\mathbb{P}(\boldsymbol{Y} \in B) = \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\boldsymbol{\delta}\|}{\sigma}\right)$.

4. Finally, algebra shows that $\mathbb{P}(\boldsymbol{Y} \in A) > \mathbb{P}(\boldsymbol{Y} \in B)$ if and only if:

$$\begin{aligned}&\mathbb{P}(\boldsymbol{Y} \in A) > \mathbb{P}(\boldsymbol{Y} \in B) \\ \iff &\Phi\left(\Phi^{-1}(\underline{p_A}) - \frac{\|\boldsymbol{\delta}\|}{\sigma}\right) > \Phi\left(\Phi^{-1}(\overline{p_B}) + \frac{\|\boldsymbol{\delta}\|}{\sigma}\right) \\ \iff &\Phi^{-1}(\underline{p_A}) - \frac{\|\boldsymbol{\delta}\|}{\sigma} > \Phi^{-1}(\overline{p_B}) + \frac{\|\boldsymbol{\delta}\|}{\sigma} \\ \iff &\Phi^{-1}(\underline{p_A}) > \Phi^{-1}(\overline{p_B}) + \frac{2\|\boldsymbol{\delta}\|}{\sigma} \\ \iff &\|\boldsymbol{\delta}\| < \frac{\sigma}{2}(\Phi^{-1}(\underline{p_A}) - \Phi^{-1}(\overline{p_B}))\end{aligned}$$

which recovers the theorem statement.

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?

Solution 2.

1. Without loss of generality we can assume $\boldsymbol{\mu} = 0$, to simplify the notation.

$$
\begin{aligned}
m(\boldsymbol{z}) &:= \frac{g_{\boldsymbol{Y}}(\boldsymbol{z})}{g_{\boldsymbol{X}}(\boldsymbol{z})} \\
&= \frac{\frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{z}-(\boldsymbol{\mu}+\boldsymbol{\delta}))^T(\boldsymbol{z}-(\boldsymbol{\mu}+\boldsymbol{\delta}))\right)}{\frac{1}{\sqrt{(2\pi\sigma^2)^d}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{z}-\boldsymbol{\mu})^T(\boldsymbol{z}-\boldsymbol{\mu})\right)} \\
&= \frac{\exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{z}-\boldsymbol{\delta})^T(\boldsymbol{z}-\boldsymbol{\delta})\right)}{\exp\left(-\frac{1}{2\sigma^2}\boldsymbol{z}^T\boldsymbol{z}\right)} \\
&= \exp\left(-\frac{1}{2\sigma^2}\left((\boldsymbol{z}-\boldsymbol{\delta})^T(\boldsymbol{z}-\boldsymbol{\delta})-\boldsymbol{z}^T\boldsymbol{z}\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(\boldsymbol{z}^T\boldsymbol{z}-2\boldsymbol{z}^T\boldsymbol{\delta}+\boldsymbol{\delta}^T\boldsymbol{\delta}-\boldsymbol{z}^T\boldsymbol{z}\right)\right) \\
&= \exp\left(-\frac{1}{2\sigma^2}\left(-2\boldsymbol{z}^T\boldsymbol{\delta}+\boldsymbol{\delta}^T\boldsymbol{\delta}\right)\right) \\
&= \exp\left(\frac{1}{\sigma^2}\boldsymbol{z}^T\boldsymbol{\delta}-\frac{\boldsymbol{\delta}^T\boldsymbol{\delta}}{2\sigma^2}\right) \qquad (4)
\end{aligned}
$$

2. We define $a := \frac{1}{\sigma^2}$ and $b := \frac{-(\boldsymbol{\delta}^T\boldsymbol{\delta})}{2\sigma^2}$ and thus can rewrite Eq. (4) as $\exp\left(a\boldsymbol{\delta}^T z + b\right)$. Thus if we know $\boldsymbol{\delta}^T\boldsymbol{z} \leq t$ we can write $m(\boldsymbol{z}) = \exp\left(a\boldsymbol{\delta}^T\boldsymbol{z}+b\right) \leq \exp\left(at+b\right) =: t'$. Thus we obtain $t' = \exp\left(at+b\right)$.

3. By definition of $m(\boldsymbol{z})$ and S we know that $g_{\boldsymbol{Y}}(\boldsymbol{z}) \leq t' g_{\boldsymbol{X}}(\boldsymbol{z}) \quad \forall \boldsymbol{z} \in S$ and therefore $\int_S g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} \leq t' \int_S g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}$. Similarly $g_{\boldsymbol{Y}}(\boldsymbol{z}) > t' g_{\boldsymbol{X}}(\boldsymbol{z}) \quad \forall \boldsymbol{z} \in S^C$ and thus $\int_{S_C} g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} > t' \int_{S_C} g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}$. Combining both facts accounting for the sign the statement follows immediately.

   By denoting $F := \{\boldsymbol{z} \in \mathbb{R}^d : [f(\boldsymbol{z}) = c]\}$ we can analogously conclude

$$
\begin{aligned}
\int_S [f(\boldsymbol{z}) = c] \cdot g_{\boldsymbol{Y}}(\boldsymbol{z}) &= \int_{S\cap F} g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} \\
&\leq t' \int_{S\cap F} g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z} \\
&= \int_S [f(\boldsymbol{z}) = c] \cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}.
\end{aligned}
$$

   Likewise we can show $\int_S [f(\boldsymbol{z}) \neq c] \cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z} > t' \int_S [f(\boldsymbol{z}) \neq c] \cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}$ and finally

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

4.

$$\begin{aligned}
&\mathbb{P}(f(\boldsymbol{Y})=c)-\mathbb{P}(\boldsymbol{Y}\in S)\\
&\overset{\text{definition}}{=}\int_{\mathbb{R}^d}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}-\int_{S}g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\\
&\overset{\text{by Eq. (3)}}{=}\left[\int_{S^C}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}+\int_{S}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\right]-\int_{S}g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\\
&=\quad\left[\int_{S^C}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}+\int_{S}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{Y}}(\boldsymbol{z})d\boldsymbol{z}\right]\\
&\qquad-\left[\int_{S}[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]\\
&=\quad\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}\\
&\overset{\text{sub-task 3}}{\geq}\quad 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]\\
&\overset{\text{add 0}}{=}\quad t'\Bigg[\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}\\
&\qquad+\int_{S}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}-\int_{S}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}\Bigg]\\
&=\quad t'\left[\int_{\mathbb{R}^d}[f(\boldsymbol{z})=c]\cdot g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}-\int_{S}g_{\boldsymbol{X}}(\boldsymbol{z})d\boldsymbol{z}\right]\\
&\overset{\text{definition}}{=}\quad t'\left[\mathbb{P}(f(\boldsymbol{X})=c)-\mathbb{P}(\boldsymbol{X}\in S)\right]\\
&\overset{\text{assumption}}{\geq}\quad 0
\end{aligned}$$

5. The generalization for the second part is straight-forward and only requires to change the direction of some inequalities.

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 $\subseteq$.

Solution 3.

1. $\|\boldsymbol{x}\|_\infty = \max_{i \in \{1,\dots,d\}} |x_i| \leq \sum_{i=1}^{d} |x_i| = \|\boldsymbol{x}\|_1$.
   Thus $c_1 = 1$.
   Example for tightness: $\|\binom{1}{0}\|_\infty = 1 = \|\binom{1}{0}\|_1$.

2. $\|\boldsymbol{x}\|_1 = \sum_{i=1}^{d} |x_i| \leq \sum_{i=1}^{d} \max_{i \in \{1,\dots,d\}} |x_i| = d \cdot \max_{i \in \{1,\dots,d\}} |x_i| = d\|\boldsymbol{x}\|_\infty$.
   Thus $c_2 = d$.
   Example for tightness: $\|\binom{1}{1}\|_1 = 2 = 2 \cdot 1 = 2 \cdot \|\binom{1}{1}\|_\infty$.

3. Let $i \in \operatorname{argmax}_{i \in \{1,\dots,d\}} |x_i|^2$. Then

$$\|\boldsymbol{x}\|_\infty^2 = |x_i|^2 \leq |x_i|^2 + \sum_{\substack{j=1 \\ j \neq i}}^{d} |x_j|^2 = \sum_{j=1}^{d} |x_j|^2 = \|\boldsymbol{x}\|_2^2.$$

The taking the square root yields $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2$.
Thus $c_3 = 1$.
Example for tightness: $\|\binom{1}{0}\|_\infty = 1 = \|\binom{1}{0}\|_2$.

4. $\|\boldsymbol{x}\|_2^2 = \sum_{i=1}^d |x_i|^2 \leq d \cdot \max_{i\in\{1,\dots,d\}} |x_i|^2 = d\|\boldsymbol{x}\|_\infty^2$. Taking the square root of both sides we obtain $\|\boldsymbol{x}\|_2 \leq \sqrt{d}\|\boldsymbol{x}\|_\infty$.
Thus $c_4 = \sqrt{d}$.
Example for tightness: $\|\binom{1}{1}\|_2 = \sqrt{2} = \sqrt{2}\cdot 1 = \sqrt{2}\cdot\|\binom{1}{1}\|_\infty$.

5. $\|\boldsymbol{x}\|_2 = \sqrt{\sum_{i=1}^d |x_i|^2} \leq \sum_{i=1}^d \sqrt{|x_i|^2} = \sum_{i=1}^d |x_i| = \|\boldsymbol{x}\|_1$ where the inequality holds due to the subadditivity of $\sqrt{\cdot}$ as

$$x_1 + x_2 \leq x_1 + 2\sqrt{x_1 x_2} + x_2 = \left(\sqrt{x_1} + \sqrt{x_2}\right)^2 \implies \sqrt{x_1 + x_2} \leq \sqrt{x_1} + \sqrt{x_2}$$

which by induction holds for arbitrarily many $x_i \geq 0$.
Thus $c_5 = 1$.
Example for tightness: $\|\binom{1}{0}\|_2 = 1 = \|\binom{1}{0}\|_1$.

6. Let $\boldsymbol{z} \in \mathbb{R}^d; z_i = \operatorname{sgn}(x_i)$. $\|\boldsymbol{x}\|_1 = \sum_{i=1}^d \boldsymbol{x}_i\boldsymbol{z}_i = \langle \boldsymbol{x}, \boldsymbol{z}\rangle \leq \|\boldsymbol{x}\|_2\|\boldsymbol{z}\|_2 = \sqrt{d}\|\boldsymbol{x}\|_2$.
Where the inequality holds due to the Cauchy-Schwarz inequality:
$|\langle \boldsymbol{a}, \boldsymbol{b}\rangle|^2 \leq \langle \boldsymbol{a}, \boldsymbol{a}\rangle\langle \boldsymbol{b}, \boldsymbol{b}\rangle$, which implies $|\langle \boldsymbol{a}, \boldsymbol{b}\rangle| \leq \|\boldsymbol{a}\|_2\|\boldsymbol{b}\|_2$
Thus $c_6 = \sqrt{d}$.
Example for tightness: $\|\binom{1}{1}\|_1 = 2 = \sqrt{2}\cdot\sqrt{2} = \sqrt{2}\cdot\|\binom{1}{1}\|_2$.

7. **Claim $\mathbb{B}_\epsilon^\infty \subseteq \mathbb{B}^2_{\epsilon\cdot\sqrt{d}}$.**
We can show the claim by taking the result from sub-task 4 and rearrange it to show $\frac{1}{\sqrt{d}}\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_\infty$. and we know that $\forall \boldsymbol{x} \in \mathbb{B}_\epsilon^\infty. \|\boldsymbol{x}\| \leq \epsilon$. Combining these two insights we obtain:

$$\begin{aligned} &\forall \boldsymbol{x} \in \mathbb{B}_\epsilon^\infty. \quad \|\boldsymbol{x}\|_\infty \leq \epsilon \\ \implies &\forall \boldsymbol{x} \in \mathbb{B}_\epsilon^\infty. \quad \frac{1}{\sqrt{d}}\|\boldsymbol{x}\|_2 \leq \|\boldsymbol{x}\|_\infty \leq \epsilon \\ \implies &\forall \boldsymbol{x} \in \mathbb{B}_\epsilon^\infty. \quad \|\boldsymbol{x}\|_2 \leq \epsilon\cdot\sqrt{d} \end{aligned}$$

Note: It is crucial here to use the result of sub-task 4 ($\|\boldsymbol{x}\|_2 \leq \sqrt{d}\|\boldsymbol{x}\|_\infty$) rather than the more intuitive seeming result from sub-task 3 ($\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2$) as these are statements about individual vectors. So for example $\|\boldsymbol{x}\|_\infty \leq \|\boldsymbol{x}\|_2$ **does not** imply $\mathbb{B}_1^\infty \subseteq \mathbb{B}_1^2$ as $\|\binom{1}{1}\|_\infty = 1 \leq \sqrt{2} = \|\binom{1}{1}\|_2$.

Similarly we can show $\mathbb{B}_\epsilon^1 \subseteq \mathbb{B}_\epsilon^2$ by sub-task 5, $\mathbb{B}_\epsilon^2 \subseteq \mathbb{B}_\epsilon^\infty$ by sub-task 3, $\mathbb{B}_\epsilon^\infty \subseteq \mathbb{B}^1_{\epsilon\cdot d}$ by sub-task 2 and $\mathbb{B}_\epsilon^2 \subseteq \mathbb{B}^1_{\epsilon\cdot\sqrt{d}}$ by sub-task 6.

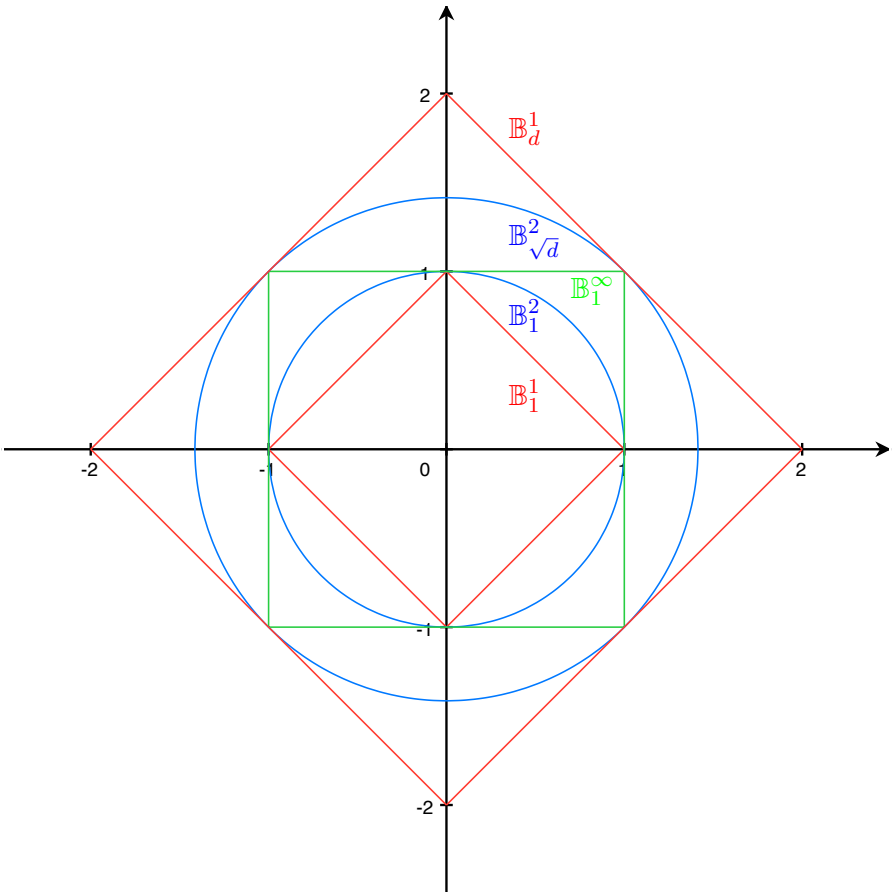

Figure 1: Boundaries of different normballs.

We thus obtain $\mathbb{B}_1^1 \subseteq \mathbb{B}_1^2 \subseteq \mathbb{B}_1^\infty \subseteq \mathbb{B}_{\sqrt{d}}^2 \subseteq \mathbb{B}_d^1$. Figure 1 shows this for $d = 2$.

Take-away message: If we prove a classifier to be save for a l_2-radius of ϵ that implies it is save for a l_∞ radius of $\frac{\epsilon}{\sqrt{d}}$.

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.