Exercise 08

Certified Defenses

Reliable and Interpretable Artificial Intelligence
ETH Zurich

Problem 1 (COLT Projections)**.**

1. Consider the zonotope below, shown in Fig. 1:

$$x = 2e_1 - e_2$$
$$y = e_1 + e_2$$

   Construct the perpendicular projection of the point $P = (-2, 3)$ onto the zonotope.

   Now, use the COLT projection scheme described in the lecture slides to project P onto the zonotope. Is the COLT projection sound? Is the COLT projection optimal?

2. Consider the zonotope below:

$$x = 2e_1 - e_2 + e_3$$
$$y = e_1 + e_2 + e_3$$

   Can you construct the COLT projection of P. Why or why not? How does COLT solve the problem?

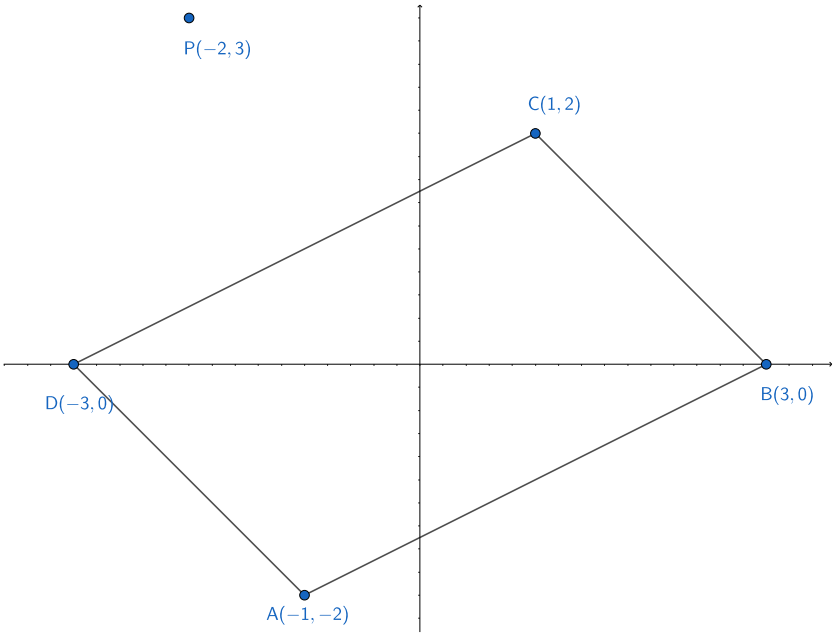

Figure 1: Zonotope for Problem 1.1

Problem 2 (Abstract Transformers of NN Loss Functions)**.**
Consider a classification neural network with three logit outputs z_1, z_2, z_3. Let the zonotope resulting from pushing an input region through the network is given by the zonotope:

$$z_1 = 0.5e_1 + e_2$$
$$z_2 = -0.5e_1 + 3e_2$$
$$z_3 = -1.5e_1 + 2e_2$$

The corresponding target label of this region is z_3.

1. Compute the abstract transformer of the max loss, as described in the lectures. Is there a concrete point in the zonotope for which the maximal loss is achieved? If so, which one?

2. Compute the abstract transformer of the cross entropy loss, as described in the lectures. Is there a concrete point in the zonotope for which the maximal loss is achieved? If so, which one?