THE UNIVERSITY OF MELBOURNE

Double Bubble, Toil and Trouble: Certified Robustness through Transitivity

Andrew Cullen (UoM) - **andrew.cullen@unimelb.edu.au**

Paul Montague (DST Group, Adelaide)

Shijie Liu (UoM)

Sarah M. Erfani (UoM)

Benjamin I.P. Rubinstein (UoM)

Supported by: Department of Defence Next Generation Technologies Fund, as well as a DECRA and LIEF grant

Certification for Classifiers

How close is the **nearest possible adversarial example**?

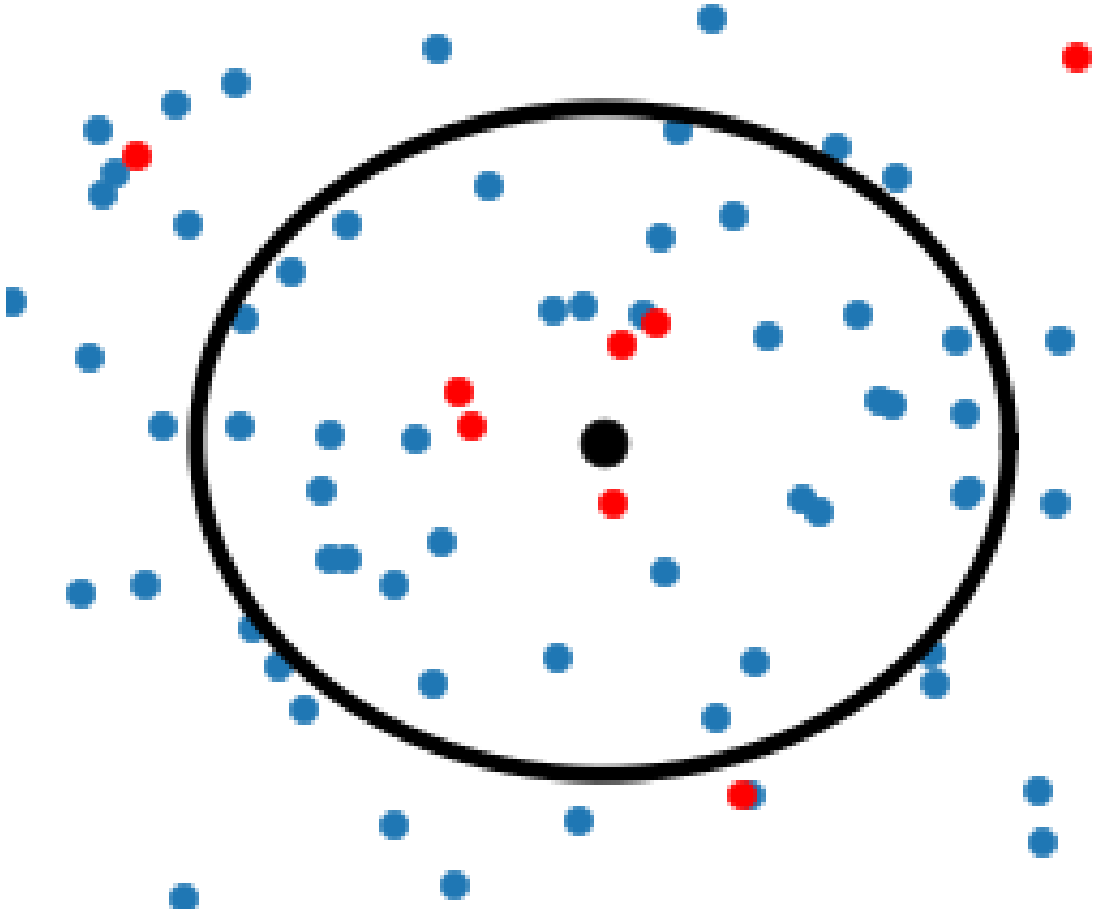
Provides a way of ranking samples based upon adversarial risk.

Certifications can be constructed by **aggregating multiple draws of normally distributed noise** about our sample point to construct a radius.

$$r = \frac{\sigma}{2}\left(\Phi^{-1}(E_{\text{Blue}}) - \Phi^{-1}(E_{\text{Red}})\right)$$

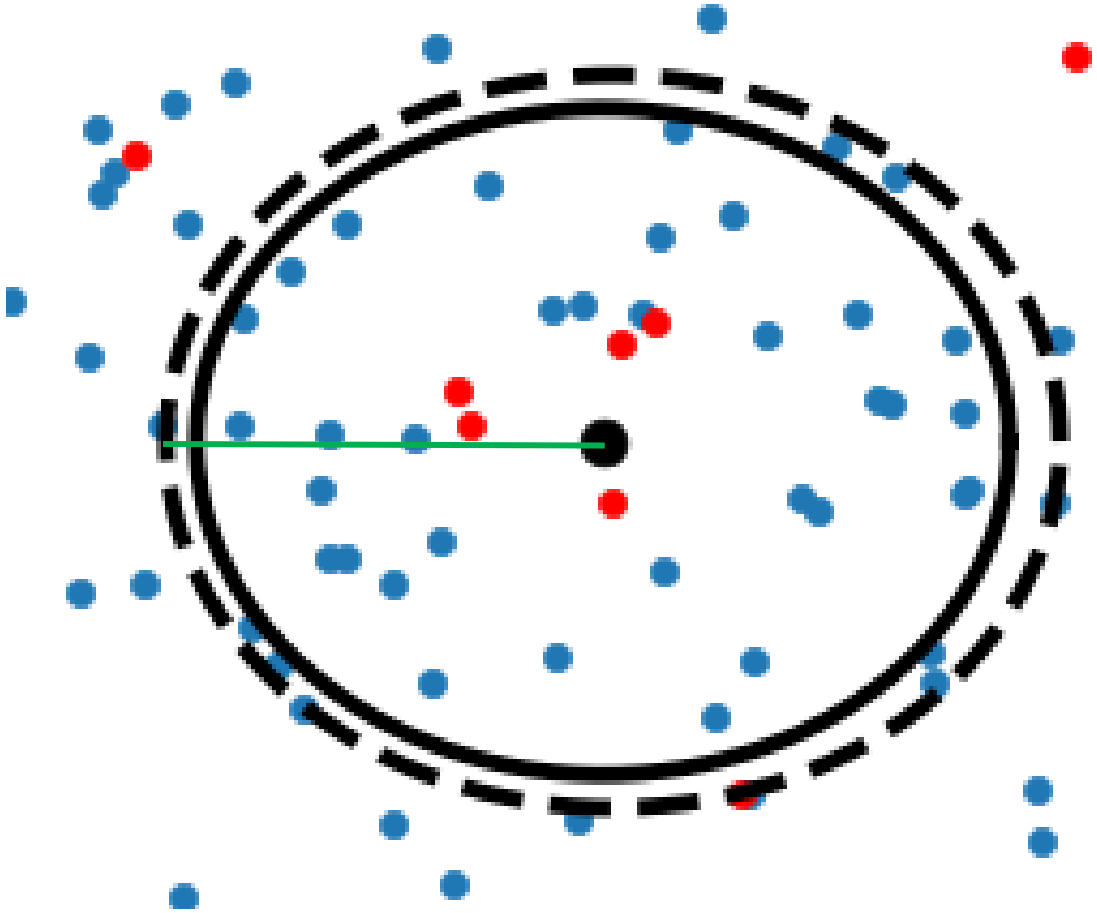
From Cohen et. al (2019)

The Cohen radius is provably the largest possible certification. Or is it?



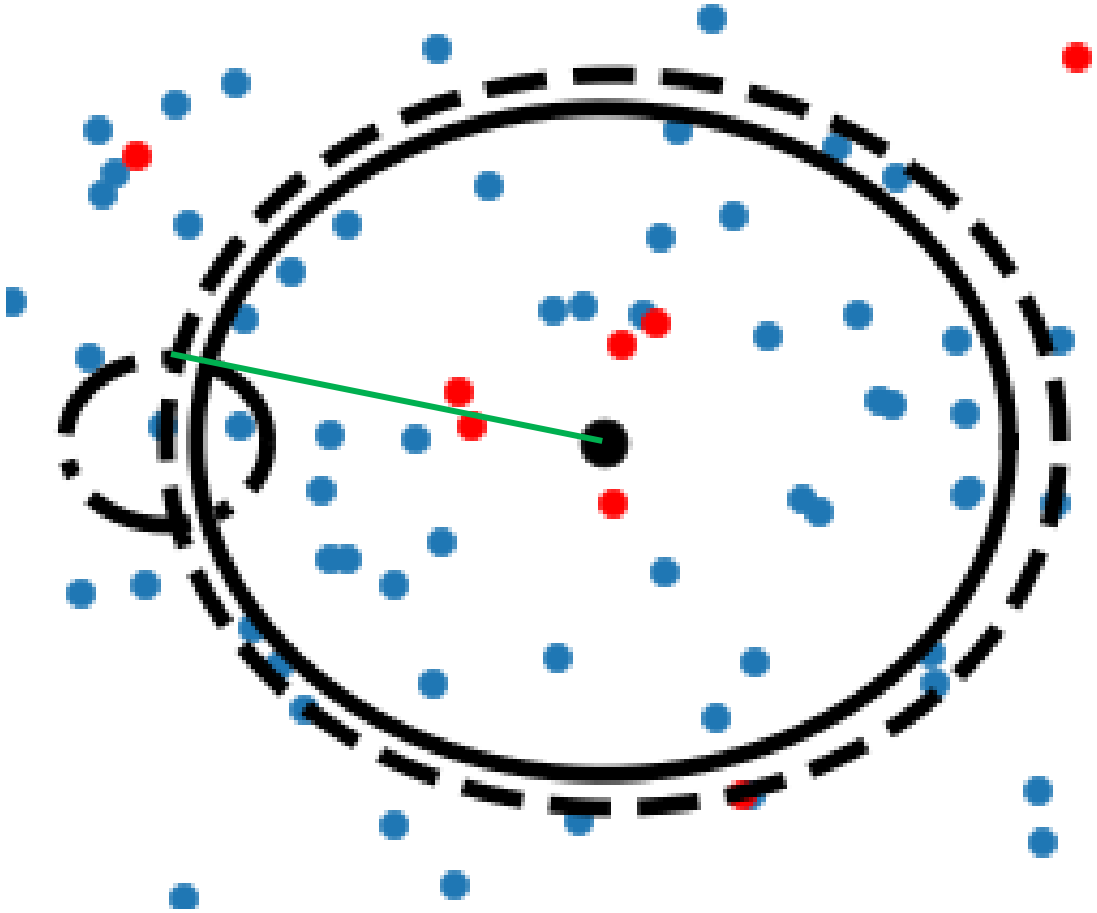
Improving on the unimprovable?

- Every point in space has its own certification
- Some of those certifications will completely enclose our original region of certification
- New certification radius is now the distance to the nearest point on the new certification hypersphere
- How do we find this new hypersphere? Gradient based search



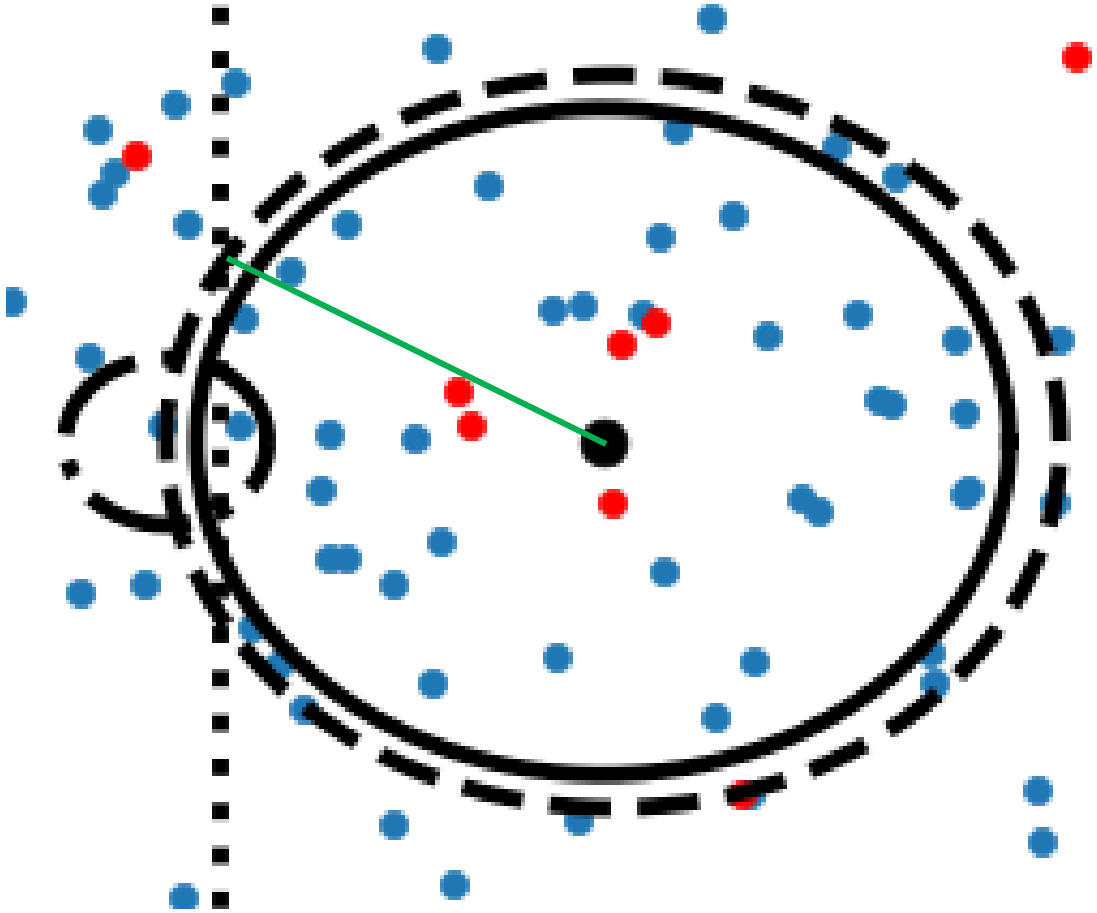
But wait, there's more!

- **Why stop at one additional hypersphere? Why not two? Or more?**
- More is highly problematic, but two allows us to further increase the radius of certification, which is now the **distance to the closest point on the surface of intersection between the two additional hyperspheres**.

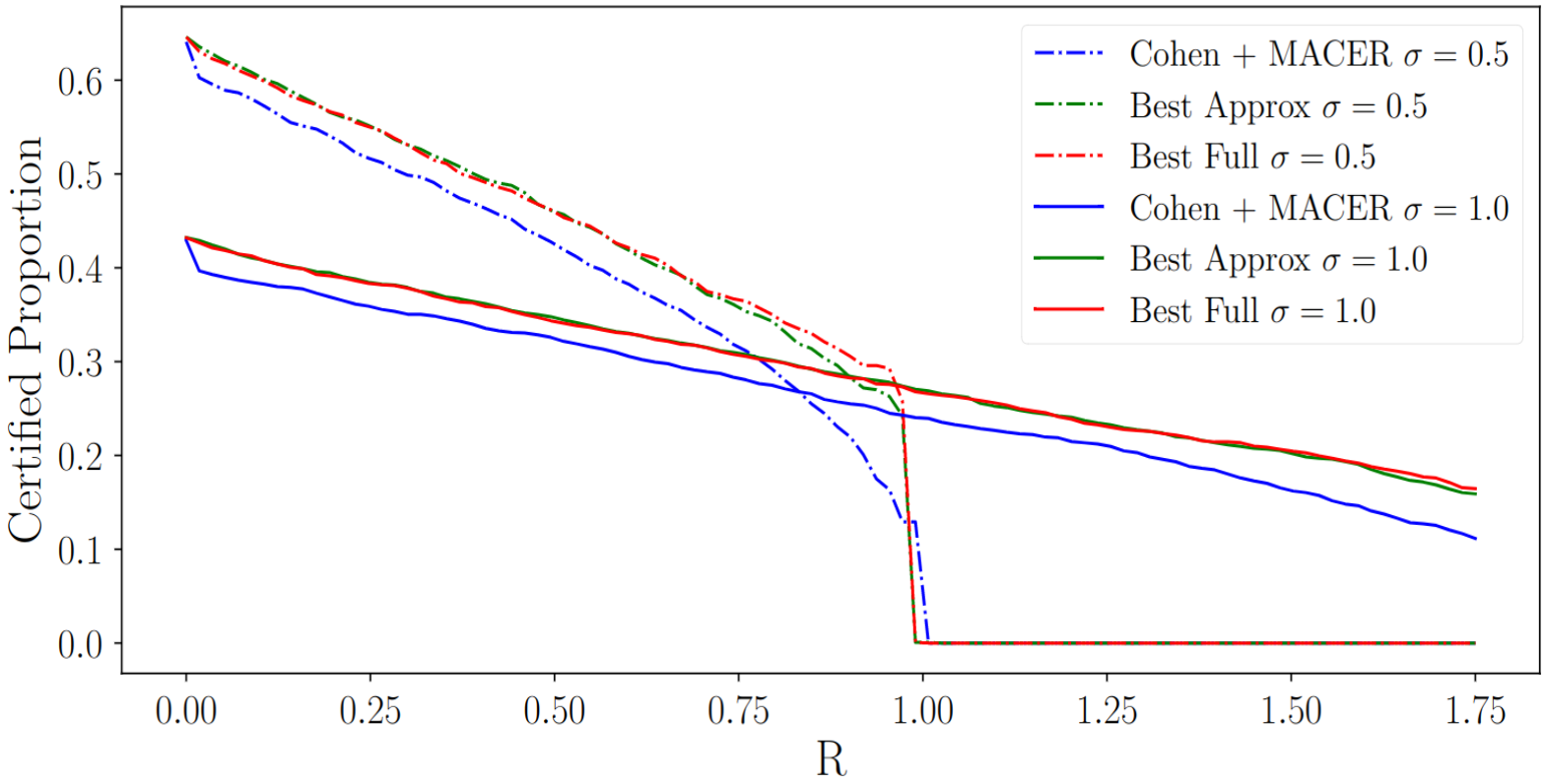


But wait, there's more!

- **Why stop at one additional hypersphere? Why not two? Or more?**
- More is highly problematic, but two allows us to further increase the radius of certification, which is now the **distance to the closest point on the surface of intersection between the two additional hyperspheres**.
- We can also introduce a boundary treatment that takes the bounded geometry of the space into account.
- Gives what we denote as **Geometrically-Informed Certified Robustness**



More samples with Larger Certifications



- The Certified Proportion is the fraction of samples with a certification above a given radius
- Geometrically-Informed Certified Robustness yields a **4-percentage point** improvement over best in class certification approaches.

THE UNIVERSITY OF MELBOURNE

Thank you

On behalf of the authors of "Double Bubble, Toil and Trouble: Enhancing Certified Robustness through Transitivity"

Andrew Cullen (UoM) - andrew.cullen@unimelb.edu.au

Paul Montague (DST Group, Adelaide)

Shijie Liu (UoM)

Sarah M. Erfani (UoM)

Benjamin I.P. Rubinstein (UoM)

Supported by: Department of Defence Next Generation Technologies Fund, as well as a DECRA and LIEF grant