Fast Geometric Projections for Local Robustness Certification

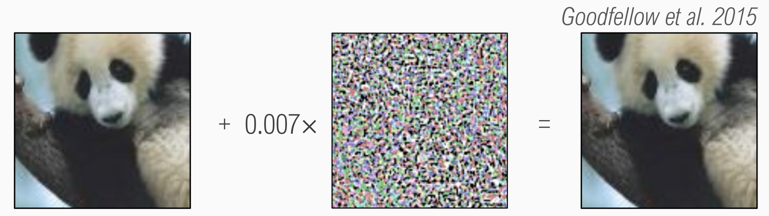
Aymeric Fromherz[1], Klas Leino1, Matt Fredrikson, Bryan Parno, Corina Păsăreanu

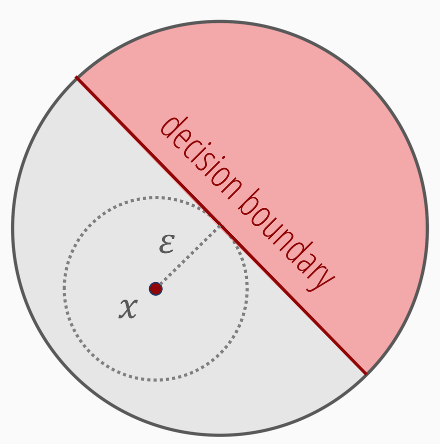
Carnegie Mellon University

Adversarial Examples & Local Robustness

Deep networks have extensively been shown to be vulnerable to *adversarial examples*, wherein inconspicuous perturbations are chosen to cause arbitrary misclassifications.

Goodfellow et al. 2015



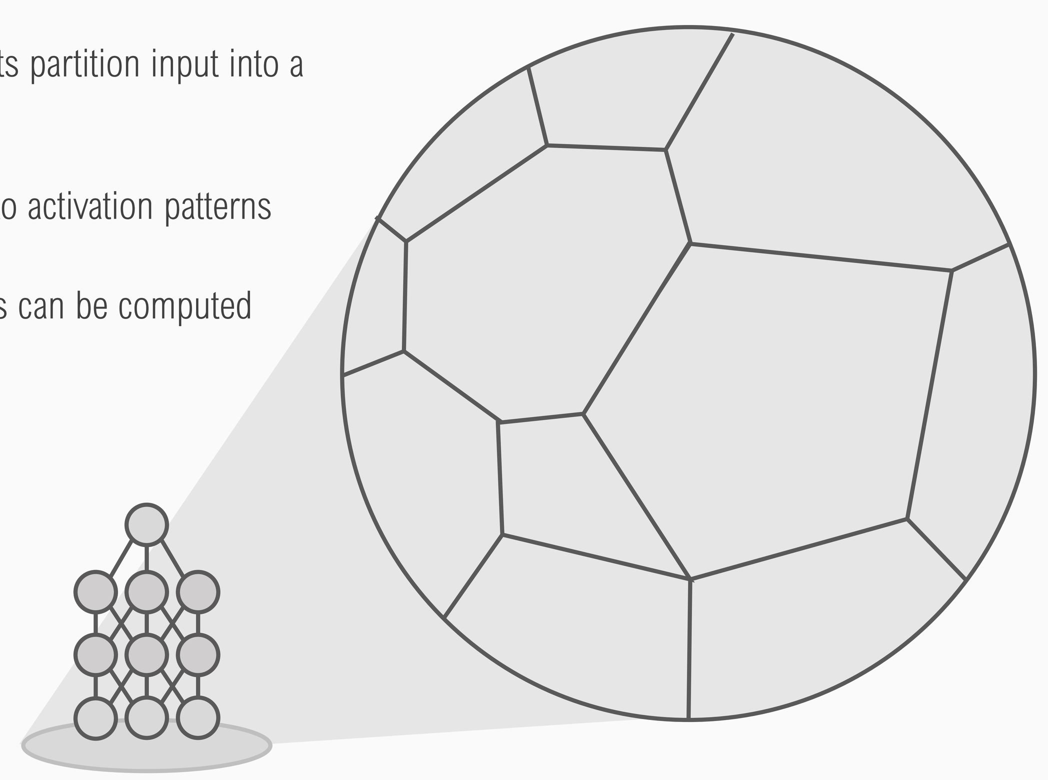


local robustness — a model is *ε-locally-robust* at a point, *x*, if it classifies all points in the ε-ball centered at *x* consistently; i.e., there are no decision boundaries within ε from *x*

certification — we would like to **prove** that a model satisfies local robustness at a given point; this precludes small-norm adversarial examples

Viewing ReLU Networks as a Polyhedral Complex

- ReLU networks are piecewise-linear
- Piecewise components partition input into a polyhedral complex
- Regions correspond to activation patterns
- Boundaries to regions can be computed using gradients

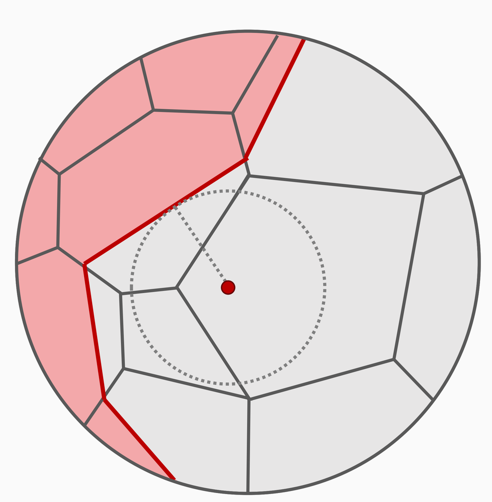


Constraint-Solving for Robustness Certification

A systematic search for decision boundaries within each of the linear regions enables certification.

This can be done using constraint-solving, e.g., GeoCert, MIP; however this is expensive, particularly in Euclidean space.

GeoCert: Jordan et al. 2019; MIP: Tjeng & Tedrake, 2017



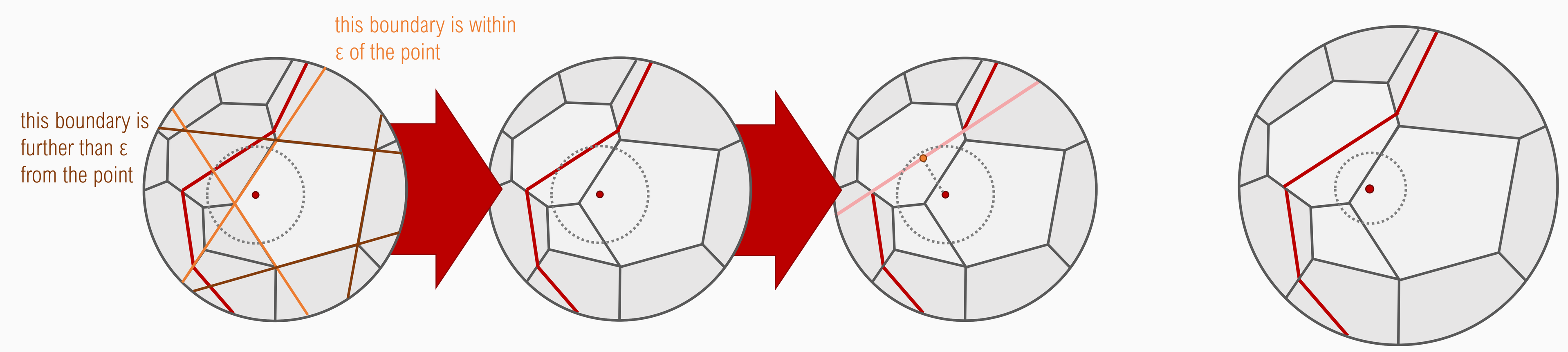
Local Robustness Certification via Projections

We present the **F**ast **G**eometric **P**rojections (FGP) algorithm for certifying local robustness. FGP relies on *projections* rather than constraint-solving to search for a decision boundary in the polyhedral complex defined by the network.

key properties

- replacing constraint-solving with projections makes FGP much faster than comparable techniques
- FGP is an *overapproximation*, but often gives exact results in practice
- when a point is deemed not to be robust, FGP provides a concrete adversarial example

Example



begin by exploring the starting region: for each boundary of starting region, check if the boundary is in the ε-ball

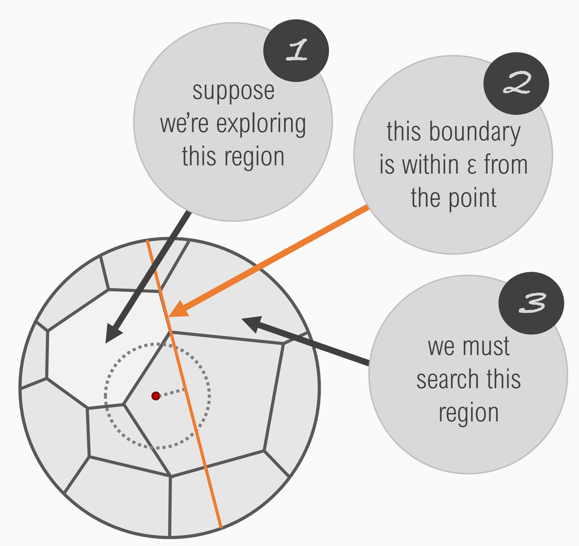
explore each of the neighboring regions whose boundaries were in the ε-ball

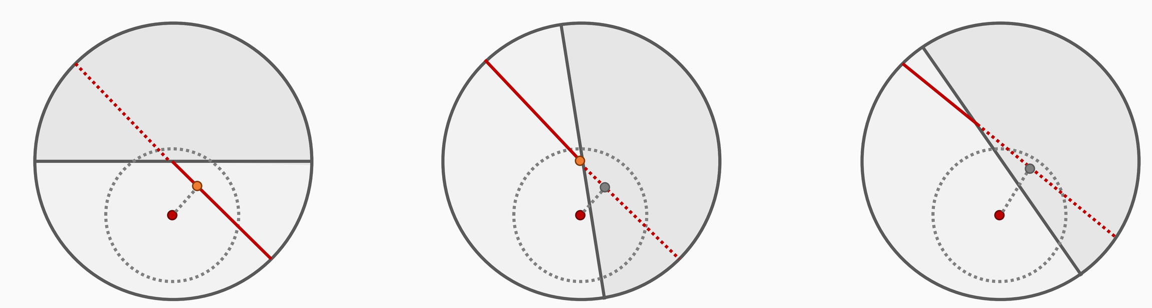
if a decision boundary is found, project onto it to verify that an adversarial example was found

if there are no more regions to explore and no decision boundaries were encountered, return ROBUST

over-aproximation — FGP may search more regions than are necessary to certify robustness

sound but not complete — in some cases where a decision boundary is found, the analysis of FGP may be inconclusive





projection onto decision boundary is in region, adversarial example exists

return NOT_ROBUST

projection onto decision boundary is not in region, but adversarial example exists

projection onto decision boundary is not in region, no adversarial example exists

can't distinguish these two cases

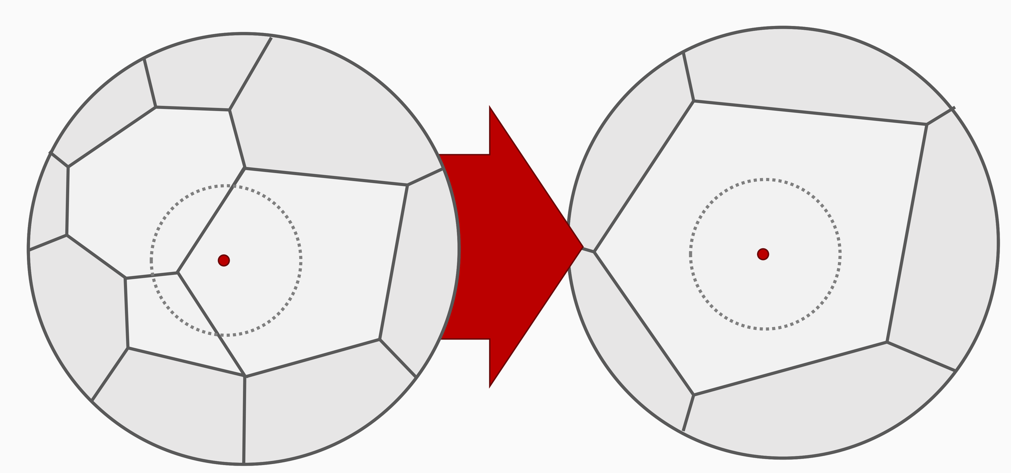
return UNKNOWN

Training for Faster Analysis

By regularizing to decrease the number of regions around any given point, we can significantly increase the speed and scalability of FGP.

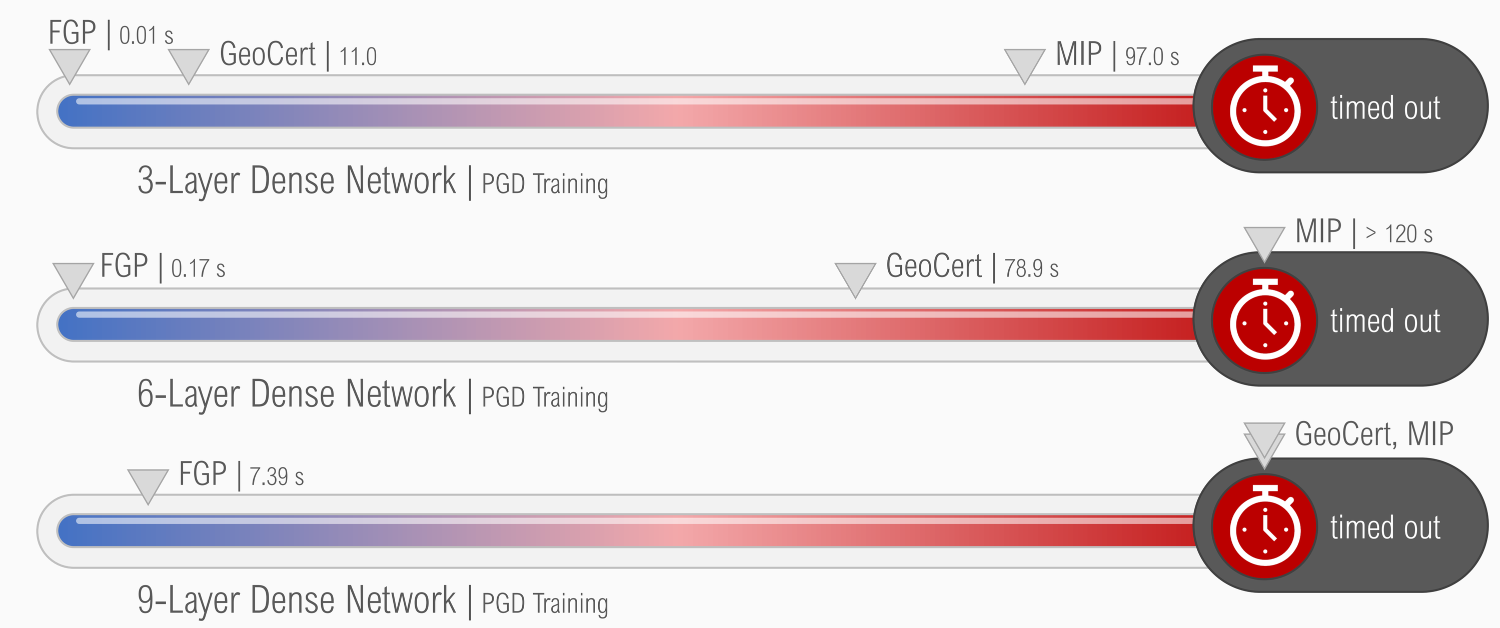
E.g., **M**aximum **M**argin **R**egularization (MMR) and ReLU Stability encourage pushing the boundaries of the polyhedral complex away from the training points, resulting in fewer regions to explore.

MMR: Croce et al. 2019; ReLU Stability: Xiao et al. 2019

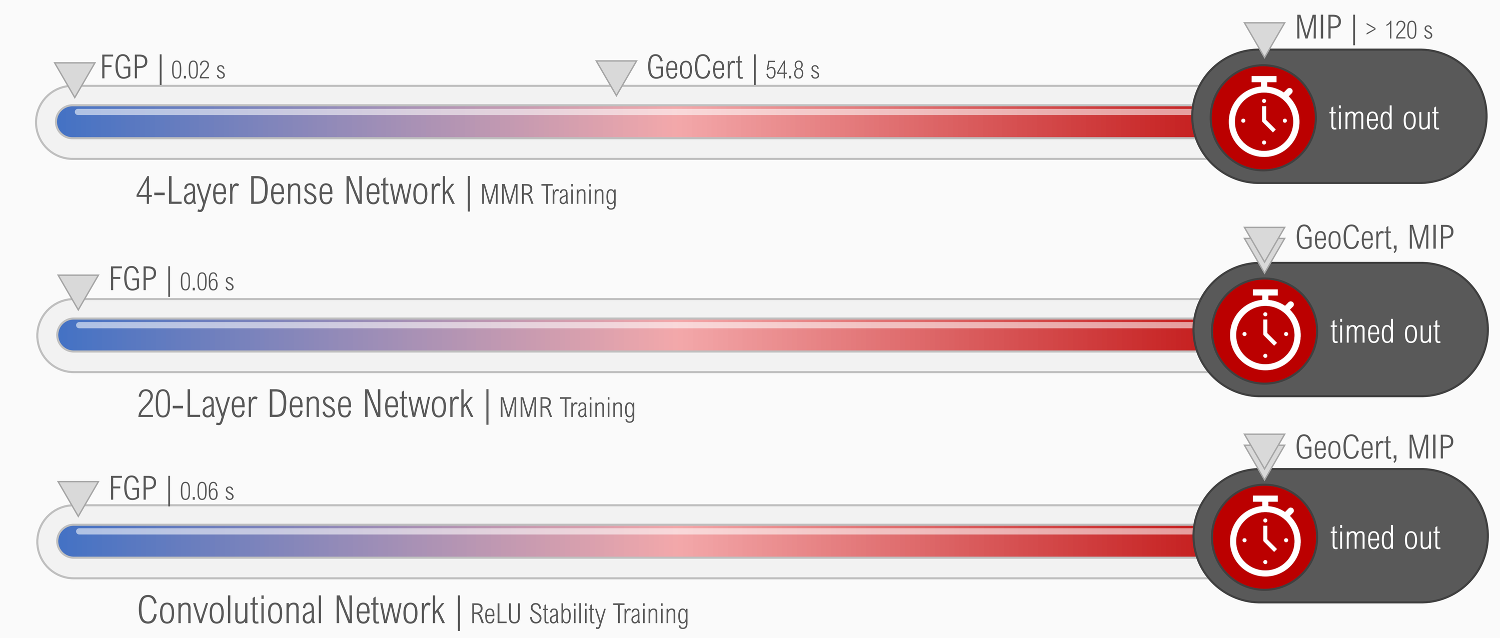


Certifying Faster with FGP

Median ℓ_2 certification times on adversarially-trained networks using FGP, GeoCert, and MIP.



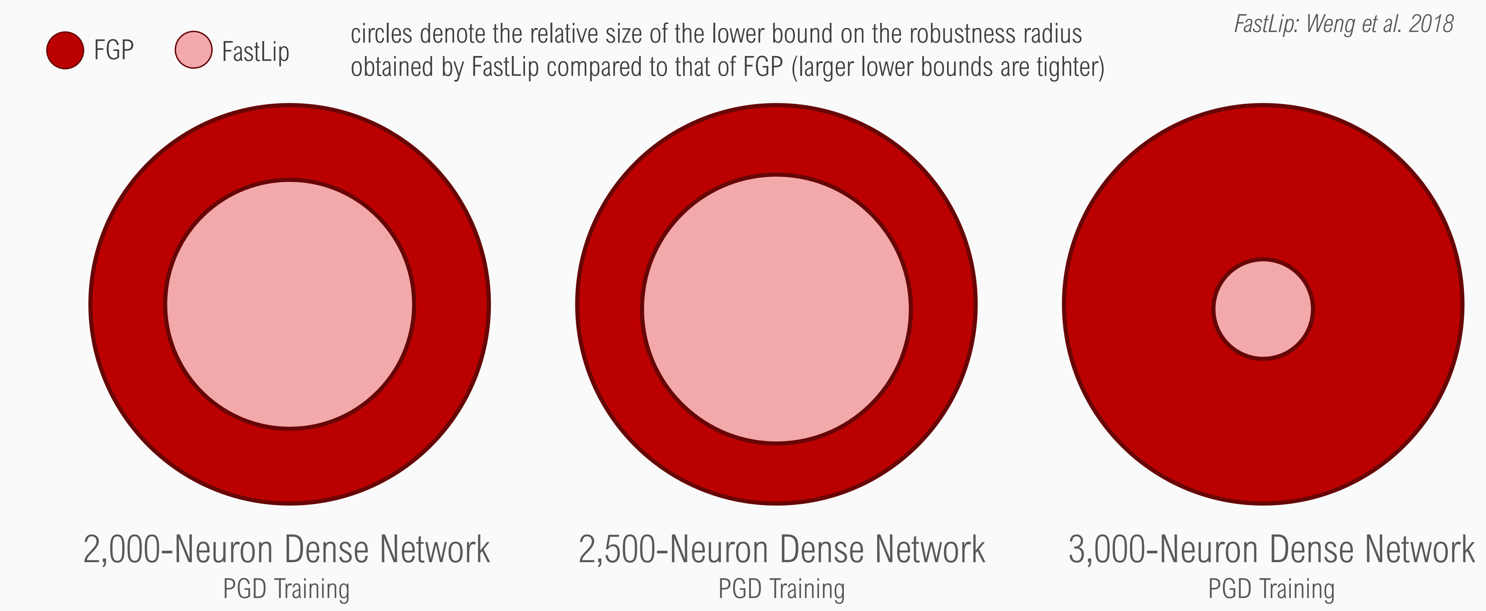
Median ℓ_2 certification times on larger networks trained for efficient certification.



Tighter Bounds with FGP

Comparison of mean lower bounds on the robustness radius obtained via FGP and FastLip.

FastLip: Weng et al. 2018



learn more — check out our spotlight talk and the full paper for more!

full paper

https://tinyurl.com/fgp-iclr2021