Enhancing Gradient-based Attacks With Symbolic Intervals
Shiqi Wang, Yizheng Chen, Ahmed Abdou, Suman Jana

GRADIENT-BASED ATTACKS
- Adversarial examples are defined in small Lp-norm bounded range
- First-order adversary: Gradient ascent the loss

LIMITATIONS
- NN is highly non-convex and non-linear
  => Easily get stuck at local optima
- Attack performance can be improved with multiple starting points.

VERIFICATION METHODS:
SOUND BOUND PROPAGATION
- Relax nonlinearity with convex function
- Over-approximate the output range
- Provide a broader view within surrounding area

LIMITATIONS
- NN is highly non-convex and non-linear
  => Easily get stuck at local optima
- Attack performance can be improved with multiple starting points.

OVERESTIMATION ERROR
(1) Cannot converge to optima if only rely on interval gradient
  => Using interval gradient ascent to locate interesting area and then use regular gradient ascent to converge
(2) The error is proportional to the input range
  => Dynamically balance the range used for each step

EXPERIMENTAL RESULTS
- 100,000 might still not enough to locate all adversarial examples
  => Stronger attacks are still needed before model is verified
  The strongest attack so far on MadryLab MNIST challenge examples
  On average 47% relatively more than PGD attack