Professorship of Data Mining and Analytics
Department of Informatics
Technical University of Munich

Certifiable Robustness and Robust Training for Graph Convolutional Networks

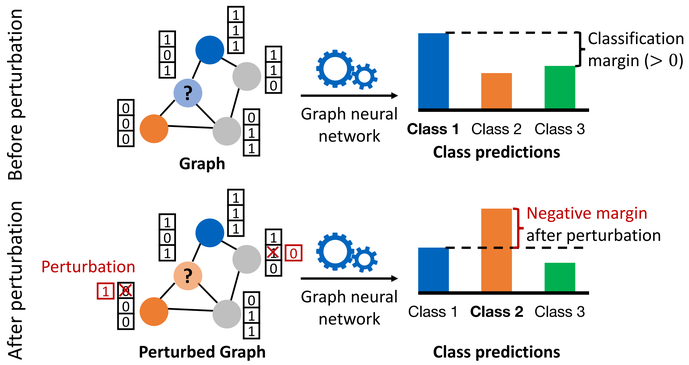
Daniel Zügner, Stephan Günnemann

tl;dr

- ❑ **Robustness certification** of GNNs via **convex relaxation**.
- ❑ GCN with standard training is **highly non-robust**.
- ❑ Our **robust training increases** robust nodes by up to **4x** without sacrificing accuracy.

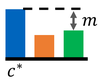
Semi-supervised node classification

- ❑ Given an (attributed) graph and a small number of labeled nodes, predict the labels of the remaining unlabeled nodes.
- ❑ Graph neural networks (GNNs) excel at this task. But: they are not robust.



Research questions

- Robustness certification: How can we verify whether a GNN is robust?
- Robust training: How can we improve GNNs' robustness?



Preliminaries

Worst-case classification margin $m^* = \underset{\text{perturbations}}{\text{minimize}} \; \underbrace{\min_{\text{class } c \neq c^*} \log p(c^*) - \log p(c)}_{\text{Classification margin } m}$

m^*: worst possible outcome among all admissible perturbations.

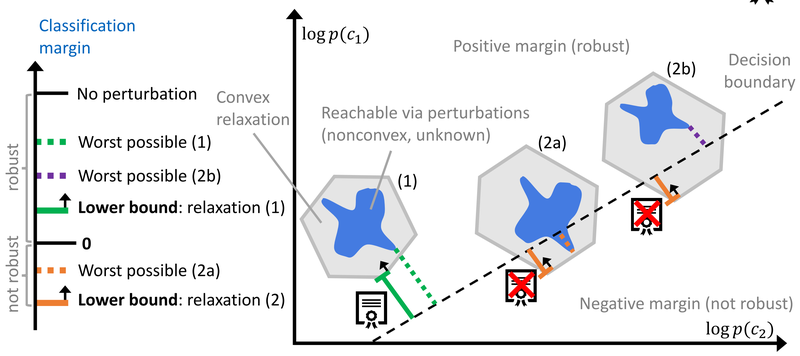
$m^* > 0$ ➔ model is robust; $m^* < 0$ ➔ model is not robust.

Certificate: prediction doesn't change under any perturbation ($m^* > 0 \; \forall \; \Delta$).

Attack scenario:

- ❑ Perturbations can be performed only to the node attributes.
- ❑ Binary node attributes, e.g. multi-hot vectors indicating words in abstract.
- ❑ Perturbations are L_0-bounded: at most q perturbations per individual node; global perturbation budget Q.

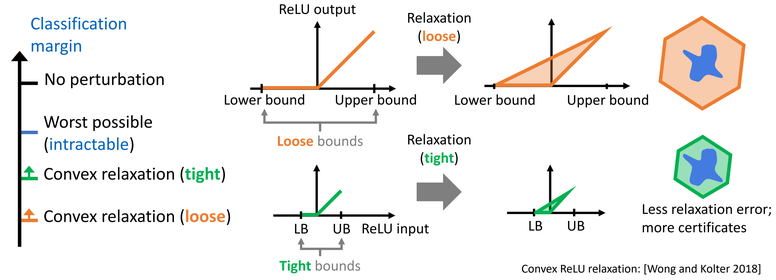
Robustness Certification: Overview



➔ False negatives are possible. But no false positives!

Relaxation 1: Convex ReLU Relaxation

Goal: Transform the GNN into a linear program that is tractable to optimize.

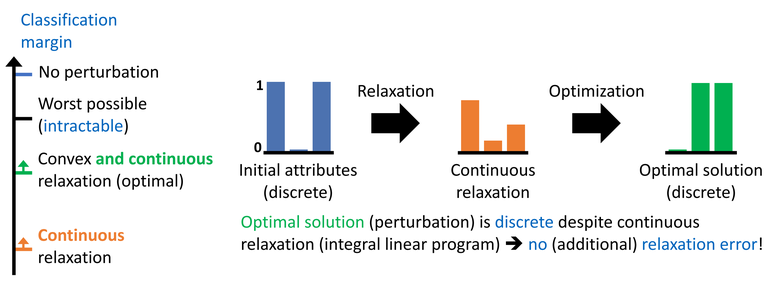


Convex ReLU relaxation: [Wong and Kolter 2018]

We derive **tight bounds** on the hidden neurons' activations exploiting the data discreteness and the graph structure.

Relaxation 2: Continuous Attributes

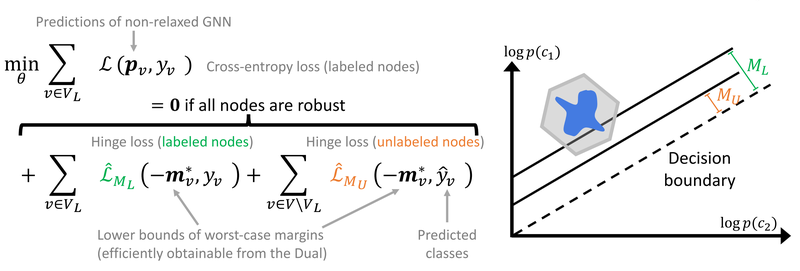
Goal: Relax the discreteness of the (binary) node attributes.



Optimal solution (perturbation) is discrete despite continuous relaxation (integral linear program) ➔ no (additional) relaxation error!

Feed the optimal (binary) perturbed features into the original GNN. Does its prediction change? If yes ➔ we have found an adversarial example.

Robust Training: Overview



- Reduces to standard cross entropy loss on the non-relaxed GNN when robust.
- Semi-supervised setting: mitigates overconfidence in possibly incorrect predictions.

Results

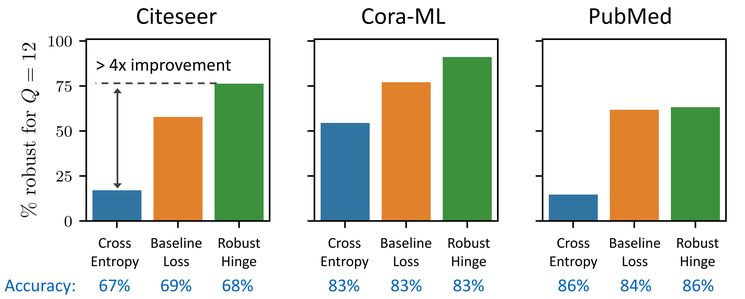


Figure: Training with our robust hinge loss increases the number of certifiably robust nodes by up to a 4x, not sacrificing accuracy.

Figure: Robust training for $Q = 12$ vs. standard training.

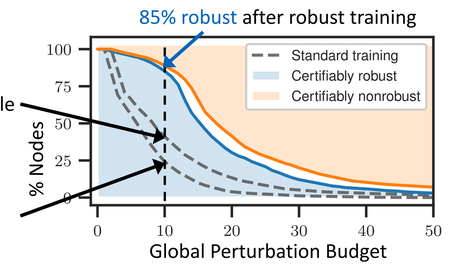
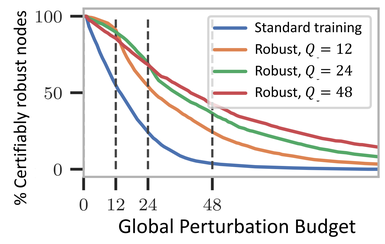


Figure: Varying the perturbation budget Q used for training:

Small Q: many nodes robust for small perturbations.

Large Q: fewer robust nodes, but handle larger perturbations.



github.com/danielzuegner/robust-gcn