Presented as late breaking results contribution
21st IFAC World Congress (Virtual)
Berlin, Germany, July 12-17, 2020

Robust neural networks via Lipschitz regularization and enforced Lipschitz bounds ⋆

Patricia Pauli * **Anne Koch** * **Julian Berberich** * **Frank Allgöwer** *

* *Institute of Systems Theory and Automatic Control, University of Stuttgart, 70569 Stuttgart, Germany, e-mail: {patricia.pauli, anne.koch, julian.berberich, frank.allgower}@ist.uni-stuttgart.de*

Abstract: Neural networks are successful in many fields, yet they can be easily fooled by imperceptible adversarial perturbations. One measure of robustness of an NN to such perturbations in the input is the Lipschitz constant of the input-ouput map defined by the NN. In this work, we therefore propose a framework for training of single-hidden layer NNs that not only minimizes the underlying loss but also encourages the NN's robustness by keeping its Lipschitz constant small. The resulting optimization problem is solved using the alternating direction method of multipliers that splits the problem into two subproblems, the minimization of the training loss and a semidefinite programming-based regularizer that penalizes the Lipschitz constant. A variation of the framework allows not only for minimization of the Lipschitz constant but for enforcing a desired Lipschitz bound during training.

⋆ This work was funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy - EXC 2075 - 390740016. The authors thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS) for supporting Patricia Pauli, Anne Koch, and Julian Berberich.