1-Path-Norm Regularization of Deep Neural Networks

Fabian Latorre[1], Antoine Bonnet[1], Paul Rolland[2], Nadav Hallak[3] and Volkan Cevher[1]

[1]EPFL [2]Swiss Data Science Center [3]Technion, Israel. Correspondence to: fabian.latorre@epfl.ch

EPFL

Motivation & Overview

The *1-path-norm* provides *width-independent* generalization bounds for ReLU networks (Neyshabur et al. 2015). The path-norm expression is nonconvex and nonsmooth, making it hard to handle in an optimization framework.

Our contributions.

- Connection between 1-path-norm and the Lipschitz constant of networks with arbitrary depth/width.
- Approximate Proximal Gradient for 1-path-norm regularization that requires only forward/backward passes through a modified network (Algorithm 4).
- Experiments show 1-path-norm regularization improves classification error and robustness of Fully connected architectures, vs L2 (weight decay) or no regularization. Proximal Gradient methods perform better than automatic differentiation (AD) in the robustness task.

Definition (1-path-norm)

For an L-layer neural network $f_W(x) := W^L\sigma(W^{L-1}\sigma(\cdots\sigma(W^1x)\cdots))$ with activation function $\sigma:\mathbb{R}\to\mathbb{R}$, its 1-path-norm can be defined as:

$$P_1(W) := \mathbb{1}^T|W^L||W^{L-1}|\cdots|W^1|\mathbb{1} \tag{1}$$

- $|W^\ell|$ is the matrix obtained by application of the absolute value.
- $\mathbb{1}$ denotes an all-ones column vector.

Theorem

Let $f_W:\mathbb{R}^{d_0}\to\mathbb{R}$, $f_W(x) := W^L\sigma(W^{L-1}\sigma(\cdots\sigma(W^1x)\cdots))$ be a network such that

$$0 \le \sigma'(x) \le 1 \qquad \text{or} \qquad \sigma(x) = ReLU(x)$$

Choose the ℓ_∞-norm for the input space and $|\cdot|$ for the output space. The Lipschitz constant of the network, denoted by L_W is bounded as follows:

$$L_W \le P_1(W) \le \prod_{\ell=1}^{L}\|W^\ell\|_\infty. \tag{2}$$

The right-hand-side of Equation (2) is usually referred to as the trivial bound based on the product of the norms of each weight matrix.

Regularized Objective and Proximal Mapping

Let $(x_i, y_i)\in\mathbb{R}^{d_0}\times\mathbb{R}^{d_L}$, be n labeled training samples:

$$\min_W \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f_W(x_i), y_i) + \lambda P_1(W) \tag{3}$$

This is a composite non-convex and non-smooth objective.

The Proximal Mapping is defined as:

$$\mathrm{prox}_{\lambda P_1(W)} \in \arg\min_Z \frac{1}{2}\|Z-W\|_F^2 + \lambda P_1(Z). \tag{4}$$

Algorithm 1 1-PN regularization using AD (Path-AD)

for $t = 1,\dots,T$ **do**
  Sample $i_1,\dots,i_b \sim \mathrm{Unif}[n]$
  $W_{t+1} \leftarrow W_t - \gamma\nabla_W\left[\frac{1}{b}\sum_{j=1}^{b}\mathcal{L}\left(f_{W_t}(x_{i_j}), y_{i_j}\right) + \lambda P_1(W_t)\right]$
return W_T

Algorithm 2 (Stochastic) Proximal Gradient Descent

for $t = 1,\dots,T$ **do**
  Sample $i_1,\dots,i_b \sim \mathrm{Unif}[n]$
  $W_{t+1/2} \leftarrow W_t - \gamma\nabla_W\frac{1}{b}\sum_{j=1}^{b}\mathcal{L}\left(f_{W_t}(x_{i_j}), y_{i_j}\right)$
  $W_t \leftarrow \mathrm{prox}_{\gamma\lambda P_1}(W_{t+1/2})$
return W_T

Lemma

Let P be a function satisfying $P(W) = P(|W|)$. Its proximal mapping satisfies

$$\mathrm{prox}_P(W) = \mathrm{sign}(W)\odot\mathrm{prox}_P^+(|W|)$$
$$\mathrm{prox}_P^+(X) := \arg\min_{Z\in\mathbb{R}_+^d}\frac{1}{2}\|X-Z\|_F^2 + P(Z). \tag{5}$$

Algorithm 4 Differentiable Proximal training of 1-path-norm regularized NNs (Prox-DIF)

for $t = 0,\dots,T-1$ **do**
  Sample $i_1,\dots,i_b \sim \mathrm{Unif}[n]$
  $W_{t+1/2} \leftarrow W_t - \gamma\nabla_W\frac{1}{b}\sum_{j=1}^{b}\mathcal{L}\left(f_{W_t}(x_{i_j}), y_{i_j}\right)$
  if $t = 0 \,(\mathrm{mod}\, B)$ **then**
    $Z_0 = |W_{t+1/2}|$
    for $t' = 0,\dots,T'-1$ **do**
      $Z_{t'+1/2} = Z_t - \gamma'\nabla_Z\left[\frac{1}{2}\||W_{t+1/2}| - Z_{t'}\|_2^2 + \lambda\gamma P_1(Z_{t'})\right]$
      $Z_{t'+1} = \max(0, Z_{t'+1/2})$

    $W_{t+1} = \mathrm{sign}(W_{t+1/2})\odot Z_{T'}$
  else
    $W_{t+1} = W_{t+1/2}$
return W_T

Experiments

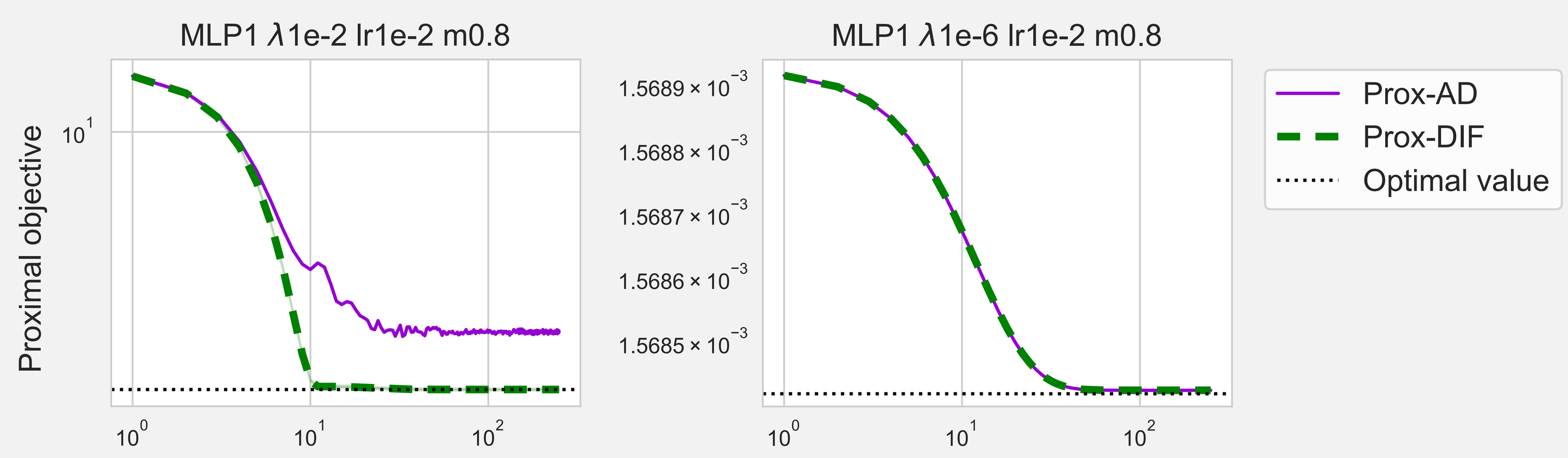

Figure: Proximal operator objective vs iteration. Randomly initialized networks.

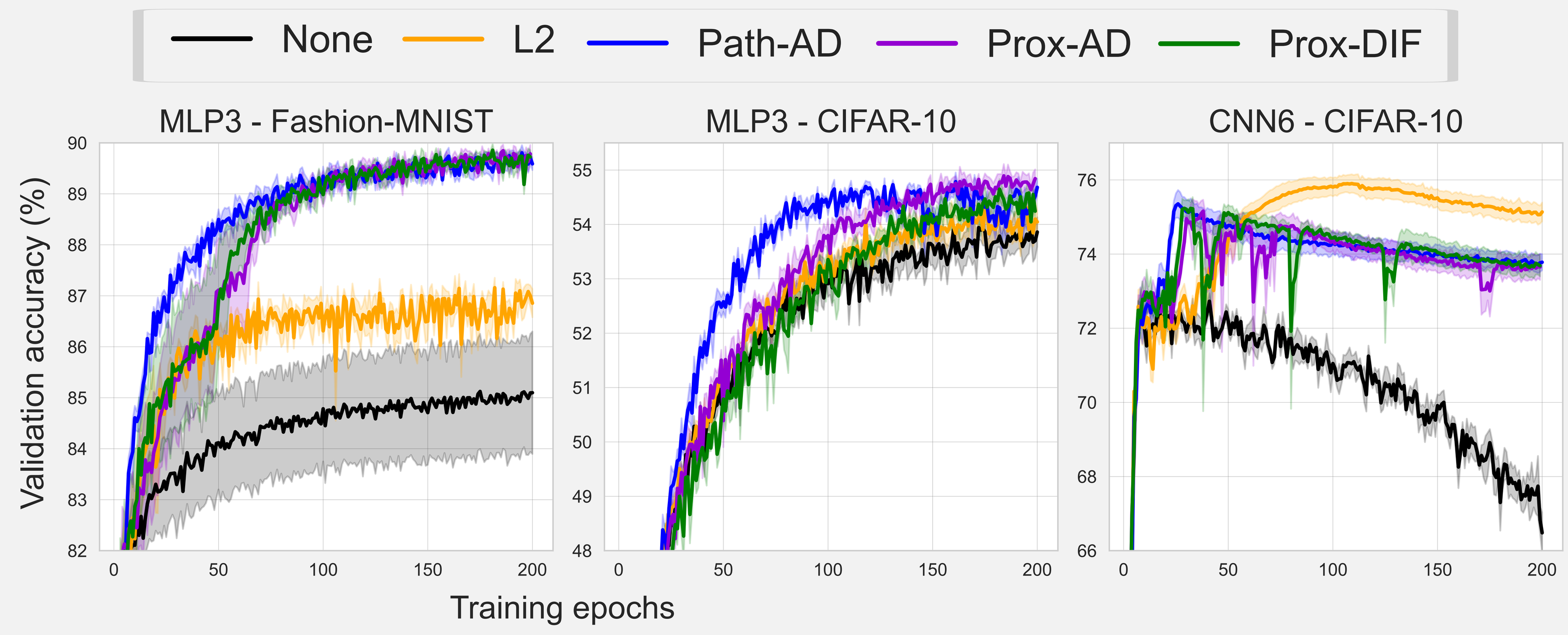

Figure: Validation accuracy vs. epoch, for different training algorithms. Averaged over 5 independent runs.

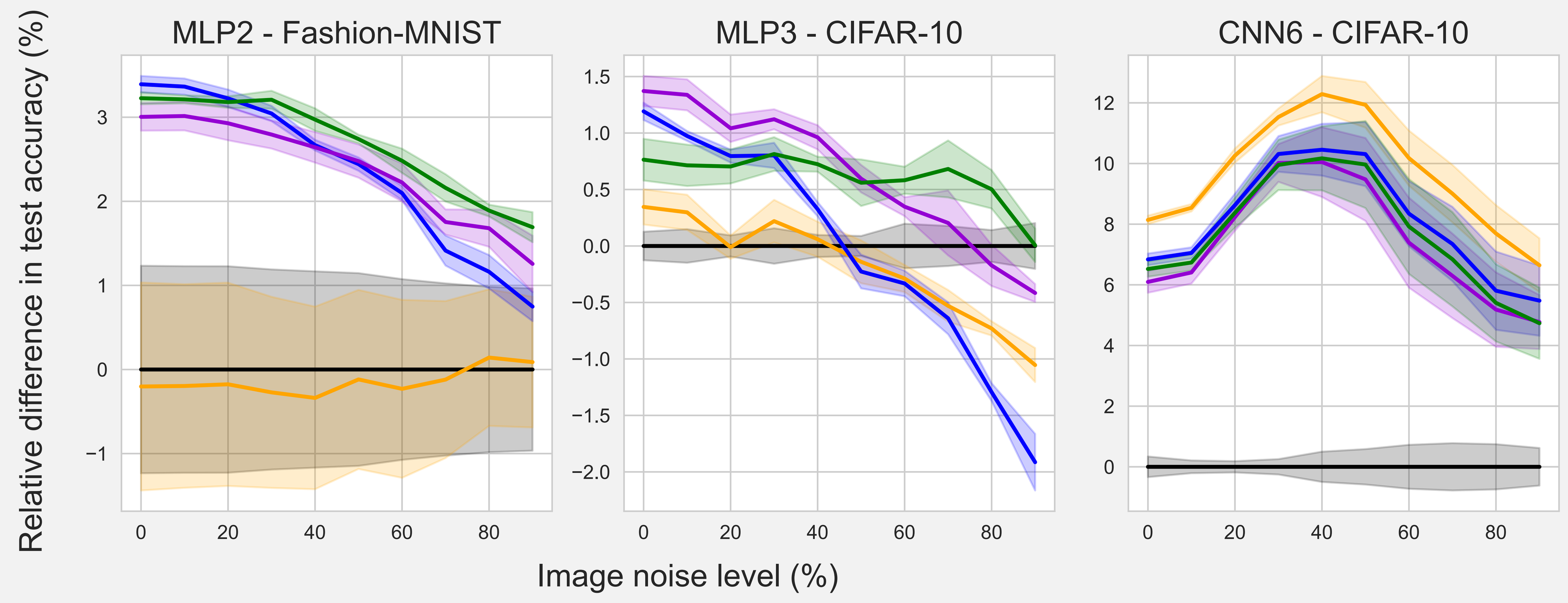

Figure: Absolute difference in test accuracy with regards to the unregularized model vs image noise level, for different training algorithms. Averaged over 5 independent runs.

This work is funded (in part) through a PhD fellowship of the Swiss Data Science Center, a joint venture between EPFL and ETH Zurich. This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n. 725594 - time-data). This work was supported by the Swiss National Science Foundation (SNSF) under grant number 200021_178865 / 1.