NAIS-Net: Stable Deep Networks from Non-Autonomous Differential Equations

Marco Ciccone * [1], Marco Gallieri * † [2], Jonathan Masci [2], Christian Osendorfer [2], Faustino Gomez [2]

marco.ciccone@polimi.it, { marco, jonathan, christian, tino } @ nnaisense.com

*: The authors equally contributed. †: The author derived the mathematical results.

[1]: Dipartimento di Elettronica, Informatica e Bioingegneria - Politecnico di Milano. [2]: NNAISENSE SA, Lugano, CH.

Introduction

Training Very Deep Networks has been made possible thanks to the use of additive non-linear transformations (Skip-connections), such as in Highway and Residual Networks [1] :

$$x(k+1) = x(k) + f\left(x(k), \theta(k)\right), 1 \leq k \leq K. \quad (1)$$

- Skip-connections solve vanishing gradient problem.
- Still require output normalization to train (e.g BatchNorm).
- The semantics of the forward path are still unclear (iterative estimation).

Note : Very Deep Networks sharing this structure can be considered as **Dynamical Systems**. Indeed, Eq. 1 can be seen as the Forward Euler Discretization of the ODE $\dot{x} = f(x)$.

- We want the system to **not have** any strange oscillations.
- We want the network **to have** very stable behavior (unless we are modeling inherently unstable behaviors).
- We want that the propagation of the state **does not** fluctuate.

Idea : Use **Control Theory** to analyze the behavior of these networks in terms of the stability of their underlined dynamical system.

Residual Networks are Autonomous Dynamical Systems.

- Input is connected only to the first layer.
- Stability means output $\to 0$ for each input : **useless for ML applications.**

Idea : Use **input connections** to define Non-Autonomous Systems.

Background : Stability Theory

Asymptotic Stability for Non-Linear Systems

A system is said to be asymptotically stable in $\mathcal{X}$ if there exists a $\bar{x}$ and $\mathcal{KL}$-function β such that $\forall x(0) \in \mathcal{X},\ k \geq 0$:

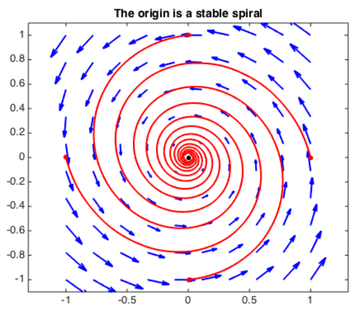
$$\|x(k) - \bar{x}\| \leq \beta(\|x(0) - \bar{x}\|, k). \quad (2)$$

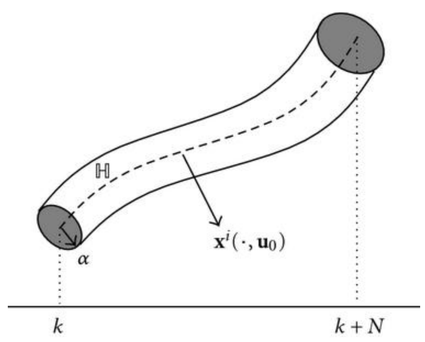
The vector $\bar{x}$ is called a steady state. β have to be strictly decreasing in k with $\lim_{k\to\infty} \beta(\cdot, k) \to 0$.

Input-Output Stability for Non-Linear System [2]

A system is said to be input-output stable (IOS) wrt bounded additive input perturbations, w, while $x \in \mathcal{X}$ if there exists a $\mathcal{KL}$-function β and a $\mathcal{K}_\infty$ function γ such that $\forall x(0) \in \mathcal{X}$:

$$\|x(k) - \bar{x}\| \leq \beta(\|x(0) - \bar{x}\|, k) + \gamma(\|w\|). \quad (3)$$





Non-Autonomous Residual Layer

Our fully connected layer is defined by the following non-autonomous system :

$$x(k+1) = x(k) + h\sigma\left(Ax(k) + Bu + b\right), \quad (4)$$

where $x \in \mathbb{R}^n$ is the latent state, $A \in \mathbb{R}^{n\times n}$ and $B \in \mathbb{R}^{n\times m}$ are the *hidden state* and input transfer matrices, $h \in (0, 1]$, $b \in \mathbb{R}^n$. Activation σ is tanh or ReLU.

NB : Same holds for convolutional layers, where A is Toeplitz

If $B = 0$, $x(0) = u$, then we have a classic ResNet (autonomous).

NAIS-Net block Stability

The state-transfer Jacobian for layer k is :

$$J(x(k), u) = \frac{\partial x(k+1)}{\partial x(k)} = I + \underbrace{h\frac{\partial \sigma(\Delta x(k))}{\partial \Delta x(k)}A}_{\text{residual Jacobian}}, \quad (5)$$

where $\Delta x(k)$ is the argument of the activation function σ.

NB : Same analysis applies to convolutional layers.

Stability Condition (from Lyapunov indirect method)

For any $\underline{\sigma} > 0$, the *state* Jacobian, $J(x, u)$, satisfies :

$$\bar{\rho} := \sup_{(x,u)\in\mathcal{P}} \rho(J(x, u)) < 1, \quad (6)$$

where $\rho(\cdot)$ is the spectral radius.

Theorem 1 (Asymptotic stability for shared weights)

If $\bar{\rho} < 1$, then the NAIS-Net block is Asymptotically Stable :

- For *tanh*, $\bar{x} = -A^{-1}(Bu + b)$.
- For ReLU, $\bar{x}$ is *continuous, piecewise affine* in $x(0)$ and u. The network is Locally Asymptotically Stable with respect to each $\bar{x}$.

If $\bar{\rho} < 1$, then the NAIS-Net block is Input-Output Stable :

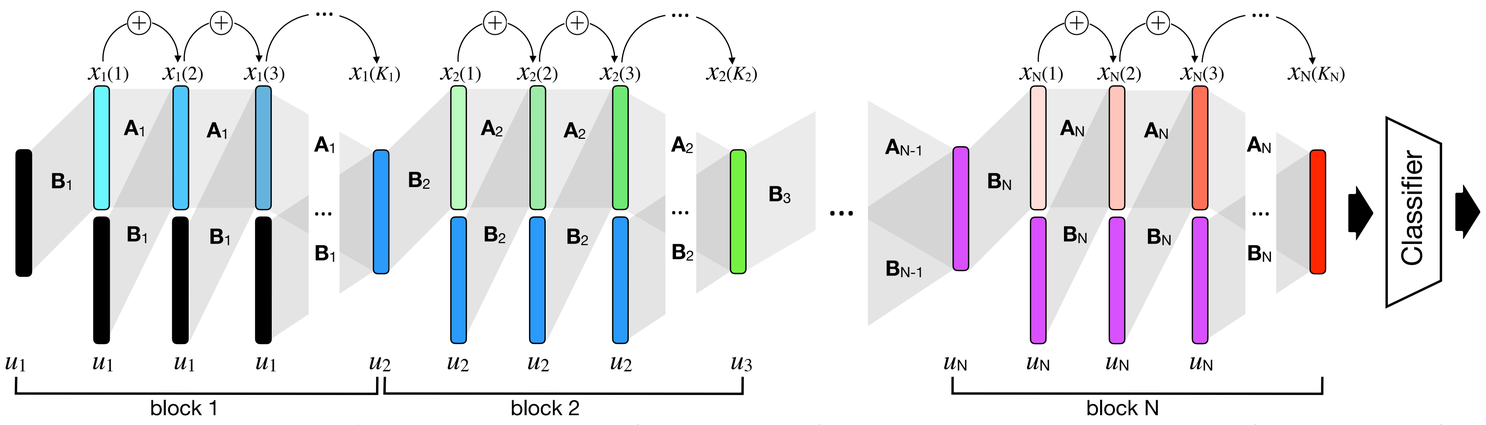
$$\lim_{k\to\infty} \|x(k) - \bar{x}\| \leq \gamma(\|w\|) \quad (7)$$

The Input-output gain is :

$$\gamma(\|w\|) = \underbrace{h\frac{\|B\|}{(1-\bar{\rho})}}_{L_w < \infty}\|w\|. \quad (8)$$

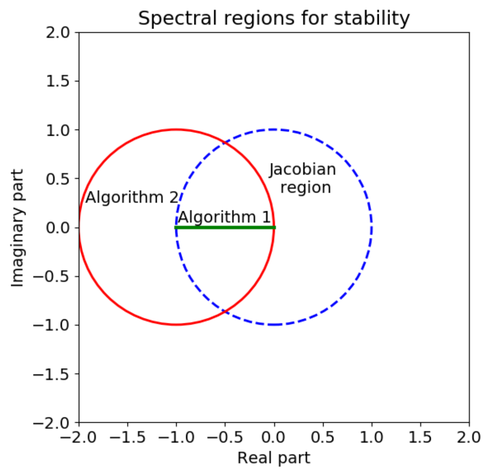
L_w is a Lipschitz constant for infinite layers.

NAIS-Net : Non-Autonomous Input-Output Stable Architecture



NAIS-Net architecture is a cascade of a time-invariant dynamical systems. Each block is an **iterative process** as the first layer in the i-th block, $x_i(1)$, is unrolled into a pattern-dependent number, K_i, of processing stages, using weight matrices $\mathbf{A}_i$ and $\mathbf{B}_i$. The skip connections from the input, u_i, to all layers in block i make the process non-autonomous. **Latent space dynamics :** each block is modeling the trajectories of the input in different latent space. IO-stability and asymptotic stability make the trajectories to be bounded with respect to noise perturbations. Moreover, each block converges to input-dependent attractors (latent representations).

Stability Implementation



Fully Connected Stability Reprojection

Input : $R \in \mathbb{R}^{\tilde{n}\times n}$, $\tilde{n} \leq n$, $\delta = 1 - 2\epsilon \in (0, 1)$.
if $\|R^T R\|_F > \delta$ **then**
  $\tilde{R} \leftarrow \sqrt{\delta}\frac{R}{\sqrt{\|R^T R\|_F}}$
else
  $\tilde{R} \leftarrow R$
end if
Output : $\tilde{R}$

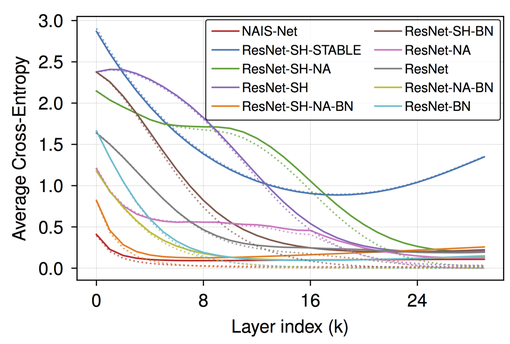
CNN Stability Reprojection

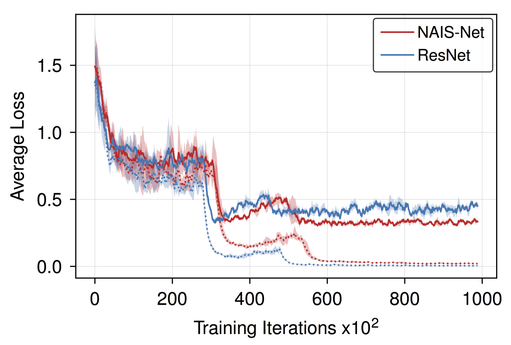
Input : $\delta \in \mathbb{R}^{N_c}$, $C \in \mathbb{R}^{n_X\times n_X\times N_C\times N_C}$, and $0 < \epsilon < \eta < 1$
for each feature map c **do**
  $\tilde{\delta}_c \leftarrow \max\left(\min\left(\delta_c, 1-\eta\right), -1+\eta\right)$
  $\tilde{C}^c_{i_{\text{centre}}} \leftarrow -1 - \tilde{\delta}_c$
  if $\sum_{j\neq i_{\text{centre}}} \left|C^c_j\right| > 1 - \epsilon - |\tilde{\delta}_c|$ **then**
    $\tilde{C}^c_j \leftarrow \left(1 - \epsilon - |\tilde{\delta}_c|\right)\frac{C^c_j}{\sum_{j\neq i_{\text{centre}}}\left|C^c_j\right|}$
  end if
end for
Output : $\tilde{\delta}$, $\tilde{C}$

Fully-Connected Layer. In order to enforce the stability of the system A needs to satisfy $\rho\left(I + h\frac{\partial\sigma(\Delta x(k))}{\partial \Delta x(k)}A\right) < 1$. Because of the Identity sum the stability region is translated (similarly as in Forward Euler). **Solution** : Eig(A) inside complex unit circle around $(-1, j0)$. We restrict the matrix A to be symmetric and negative definite by choosing the parametrization $A = -R^T R - \epsilon I$, where $R \in \mathbb{R}^{n\times n}$ is trained, and $0 < \epsilon \ll 1$ is a hyper-parameter. Them we apply the reprojection in the algorithm if the spectral norm of A is out of the stability region.

Convolutional Layer. We can write the convolution operator as matrix multiplication by vectorizing the input in a tall matrix, where each row is a channel (feature map). Then the form of the A matrix has a specific structure, where each row contains the element of the filter $C^{(c)}_{(i)}$ plus some zero elements, with central element of the filter on the diagonal. **Geršgorin's Circle Theorem** (GCT) states the eigenvalues of a squared matrix A are located within the union of the complex circles centered around the diagonal values a_{ii} of A with radius $\sum_{j=1, j\neq i}^{n} |a_{ij}|$ equal to the sum of the absolute values of the non-diagonal entries in each row of A. Thanks to GCT we can build the stability reprojection for the convolution layer : (i) for each state channel c, we set the central element of $C^{(c)}_{(c)}$ (on the diagonal) to $-1 + \delta_c$, $\|\delta_c| < 1$. Then we stack the remaining of $C^{(c)}_{(c)}$, and all of $C^{(c)}_{(j)}, \forall j \neq c$, into $C^{(c)}$. Finally we set $\sum_j \left|C^{(c)}_j\right| \leq 1 - \epsilon - |\delta_c|$. The algorithm Scales with number of channels.

Results for Image Classification - MNIST, CIFAR10-100





Model	CIFAR-10 train/test	CIFAR-100 train/test
ResNet	99.86±0.03	97.42 ± 0.06
	91.72±0.38	66.34 ± 0.82
NAIS-Net1	99.37±0.08	86.90 ± 1.47
	91.24±0.10	65.00 ± 0.52
NAIS-Net10	99.50±0.02	86.91 ± 0.42
	91.25±0.46	66.07 ± 0.24

CIFAR10-100 : accuracy averaged over 5 runs.

Preliminary Results on MNIST (left). We trained a single NAIS-Net block unrolled for a fixed length of 30 steps and we compared it against different ResNet configurations. **SH** : shared weights (time-invariant). **NA** : non-autonomous (input skip connections). **BN** : with Batch Normalization. **Stable** : stability enforced by Algorithm 1. At test time we analyzed the behavior of each network by passing the activation, $x(i)$, though the softmax classifier and measuring the cross-entropy loss. NAIS-Net is able to learn even with the stability constraint, showing that non-autonomy is key to obtaining representations that are stable and, at the same time, good for learning the task. **Generalization gap on CIFAR10-100. (center and right)** We compared NAIS-Net with ResNet three sets of 18 residual blocks with 16, 32, and 64 filters, respectively, for a total of 54 stacked blocks (single convolution). Two experiments where NAIS-Net block is unrolled 1 and 10 times. NAIS-Net is less prone to overfitting than a classic ResNet, reducing the generalization gap. This is a consequence of the stability constraint which imparts a degree of robust invariance to input perturbations. **Normalization.** NAIS-Net can unroll up to 540 layers, and still train **without requiring any output normalization at each step** (batchNorm is applied only to each bottleneck layer to speed up the training).

Pattern-Dependent Processing Depth

Pattern-Dependent Processing Depth. NAIS-Net can be unrolled for a variable number of processing steps until it reaches convergence. Thanks to the stability property, we can set a stopping criteria on the norm of the difference between consecutive states such as $\|x(k+1) - x(k)\| \leq \eta$, where $\eta = 10^{-4}$.

Image samples with corresponding NAIS-Net depth. The figure shows samples from CIFAR-10 grouped by final network depth, for four different classes. The qualitative differences evident in images inducing different final depths indicate that NAIS-Net adapts processing systematically according to the characteristics of the data. The variable depth of the network can be considered as an additional degree of freedom of the model.

- R. K. Srivastava, K. Greff, and J. Schmidhuber, "Highway networks," *arXiv preprint arXiv :1505.00387*, May 2015.
- H. Izadi, B. W. Gordon, and Y. Zhang, "Decentralized model predictive control for cooperative multiple vehicles subject to communication loss," May 2011.