PADERBORN UNIVERSITY
The University for the Information Society

Department of Communications Engineering

On the appropriateness of complex-valued neural networks for speech enhancement

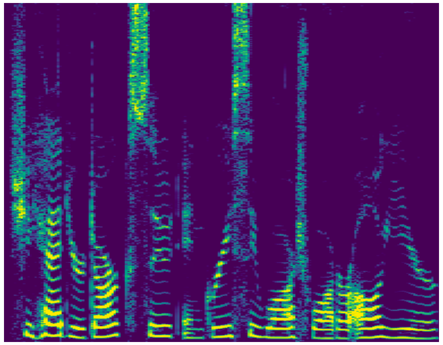
Lukas Drude
Bhiksha Raj
Reinhold Haeb-Umbach

Motivation

- Quite a few older papers on CVNNs, even on real-valued tasks
- Limited recent contributions with notable exceptions (i.e. complex valued weight matrix in RNNs)
- Real-valued NNs are universal approximators anyway
- Rather polarizing, so lets see if we can encourage vivid discussions.

Motivation, Speech Enhancement

- Real-valued data often more accessible in complex domain



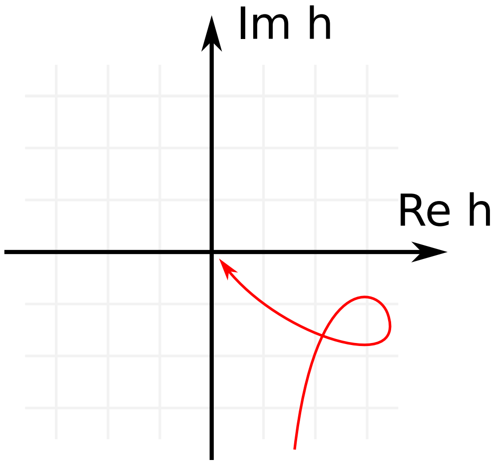
- SE algorithms often formulated in STFT/ spectral domain
  - Beamforming
  - Noise reduction/ Wiener filter
  - Speech recognition features
- A few networks trained on waveforms

BACKGROUND

Complex differentiability

- Complex differentiability:

$$\frac{\mathrm{d}f}{\mathrm{d}z} = \lim_{h \to 0} \frac{f(z+h) - f(z)}{h}$$



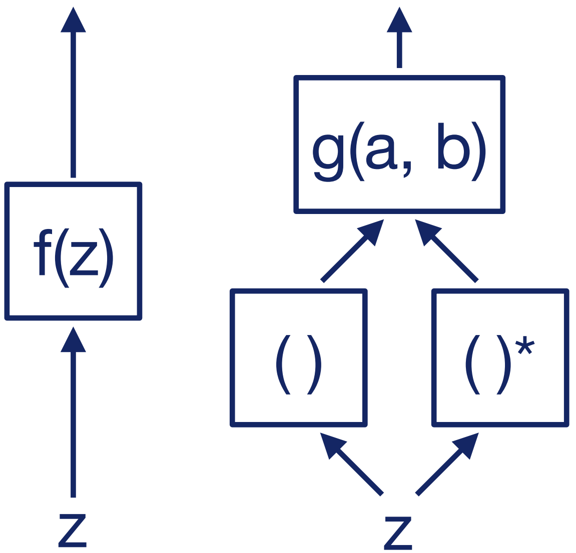
- Example: $f(z) = z^*$

$$\lim_{\eta \to 0} \frac{(z+\eta)^* - z^*}{\eta} = 1, \quad \lim_{\mathrm{j}\eta \to 0} \frac{(z+\mathrm{j}\eta)^* - z^*}{\mathrm{j}\eta} = -1.$$

Partial derivatives

- Non-holomorphic functions still partially differentiable
- Choose any rotation of the basis $x = \operatorname{Re} z, y = \operatorname{Im} z$
- One such choice for $f(z) = g(a, b), \;\; a(z) = z, \;\; b(z) = z^*$

$$\begin{aligned} \mathrm{d}f &= \frac{\partial g}{\partial a}\mathrm{d}a + \frac{\partial g}{\partial b}\mathrm{d}b \\ &= \frac{\partial g}{\partial a}\mathrm{d}z + \frac{\partial g}{\partial b}\mathrm{d}z^* \end{aligned}$$

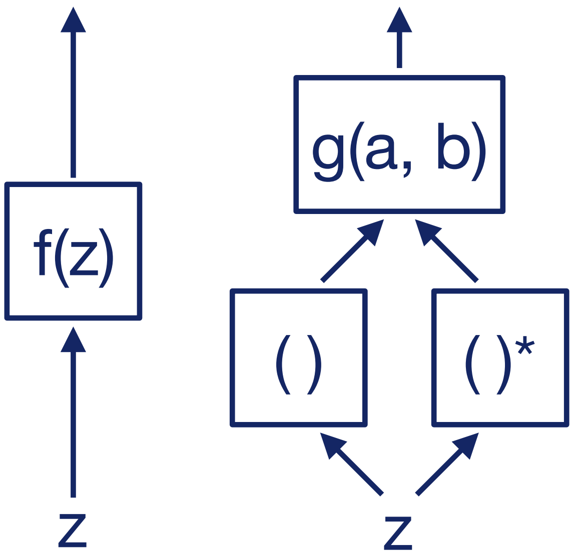


[Wirtinger1927]

Partial derivatives

- Non-holomorphic functions still partially differentiable
- Choose any rotation of the basis $x = \operatorname{Re} z, y = \operatorname{Im} z$
- One such choice for $f(z) = g(a, b),\ \ a(z) = z,\ \ b(z) = z^*$

$$\begin{aligned} \mathrm{d}f &= \frac{\partial g}{\partial a}\mathrm{d}a + \frac{\partial g}{\partial b}\mathrm{d}b \\ &= \frac{\partial g}{\partial a}\mathrm{d}z + \frac{\partial g}{\partial b}\mathrm{d}z^* \end{aligned}$$



- Finally, each network element needs to provide:

$$\frac{\partial \mathbf{g}}{\partial \mathbf{z}}, \quad \frac{\partial \mathbf{g}}{\partial \mathbf{z}^*}, \quad \nabla_{\mathbf{z}^*} = \left((\nabla_{\mathbf{g}^*})^* \frac{\partial \mathbf{g}}{\partial \mathbf{z}^*} + \nabla_{\mathbf{g}^*} \left(\frac{\partial \mathbf{g}}{\partial \mathbf{z}} \right)^* \right)$$

[Wirtinger1927]

Building blocks

- Do we need an extended linear-layer?

$$f(z) = \mathbf{A}z + \mathbf{B}z^* + \mathbf{b}$$

Building blocks

- Do we need an extended linear-layer?

  $f(z) = \mathbf{A}z + \mathbf{B}z^* + \mathbf{b}$

- How to we get a valid non-linearity?
  - Bounded and holomorphic = constant.
  - Variants: $f_{\mathrm{mt}}(z) = \tanh|z|\mathrm{e}^{\mathrm{j}\arg z}$

    $f_{\mathrm{st}}(z) = \tanh \operatorname{Re} z + \mathrm{j} \tanh \operatorname{Im} z$

    $f_{\mathrm{sr}}(z) = \max(0, \operatorname{Re} z) + \mathrm{j} \max(0, \operatorname{Im} z)$

Building blocks

- Do we need an extended linear-layer?

  $f(z) = \mathbf{A}z + \mathbf{B}z^* + \mathbf{b}$

- How to we get a valid non-linearity?
  - Bounded and holomorphic = constant.
  - Variants: $f_{\mathrm{mt}}(z) = \tanh|z|\mathrm{e}^{\mathrm{j}\arg z}$

    $f_{\mathrm{st}}(z) = \tanh \operatorname{Re} z + \mathrm{j}\tanh \operatorname{Im} z$

    $f_{\mathrm{sr}}(z) = \max(0, \operatorname{Re} z) + \mathrm{j}\max(0, \operatorname{Im} z)$

- Optimizer?
  - SGD, Momentum SGD, AdaGrad and NesterovAG generalize nicely.

EXPERIMENTS

Signal model and beamforming

- Clean speech from TIMIT, random ATF $\mathbf{H}_f$

$$\mathbf{Y}_{tf} = \mathbf{H}_f S_{tf} + \mathbf{N}_{tf}$$

- Objective: Maximize signal to noise ratio
- Known analytic solution for spatially white noise:

$$\mathbf{\Phi}_f = \sum_t \mathbf{Y}_{tf} \mathbf{Y}_{tf}^{\mathsf{H}}, \quad \mathbf{W}_f = \mathrm{PCA}\left\{\mathbf{\Phi}_f\right\}, \quad Z_{tf} = \mathbf{W}_f^{\mathsf{H}} \mathbf{Y}_{tf}$$

Signal model and beamforming

- Clean speech from TIMIT, random ATF $\mathbf{H}_f$

  $$\mathbf{Y}_{tf} = \mathbf{H}_f S_{tf} + \mathbf{N}_{tf}$$

- Objective: Maximize signal to noise ratio
- Known analytic solution for spatially white noise:

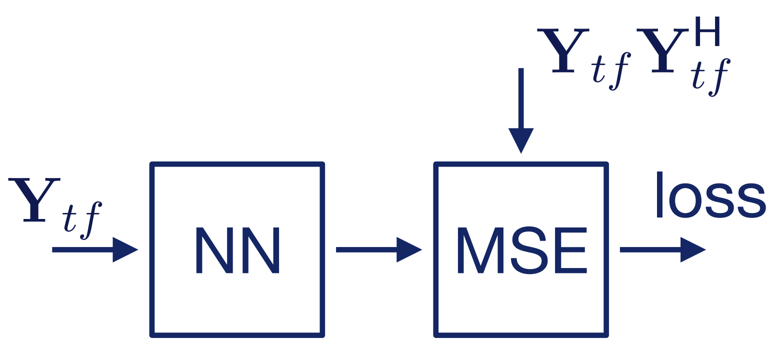
  $$\mathbf{\Phi}_f = \sum_t \mathbf{Y}_{tf}\mathbf{Y}_{tf}^{\mathsf{H}}, \quad \mathbf{W}_f = \mathrm{PCA}\left\{\mathbf{\Phi}_f\right\}, \quad Z_{tf} = \mathbf{W}_f^{\mathsf{H}}\mathbf{Y}_{tf}$$

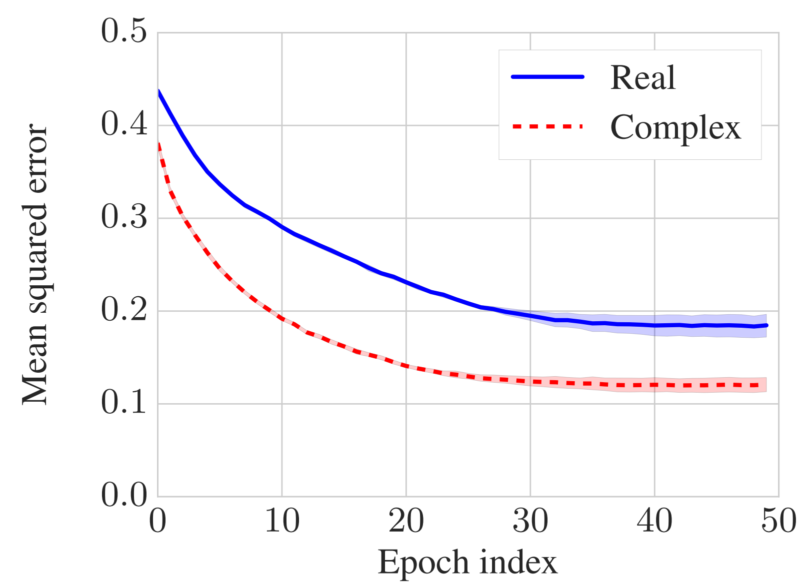
- Analytic solution can now be split into different tasks:
  a) Map observations to outer product
  b) Map covariance matrix to principal component
  c) Map observations to principal component directly

Learn outer product

RVNN	CVNN
2D → 50	D → 25
ReLU	SplitReLU
50 → $2D^2$	25 → D^2
MSE	MSE

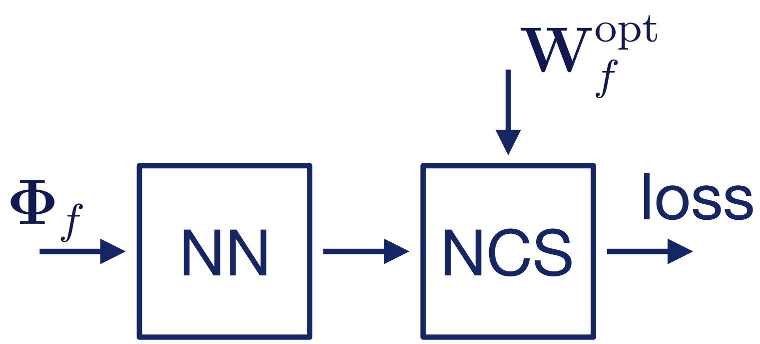
- Momentum SGD, learning rate 0.001
- D = 3 Channels
- White noise with 10 dB SNR

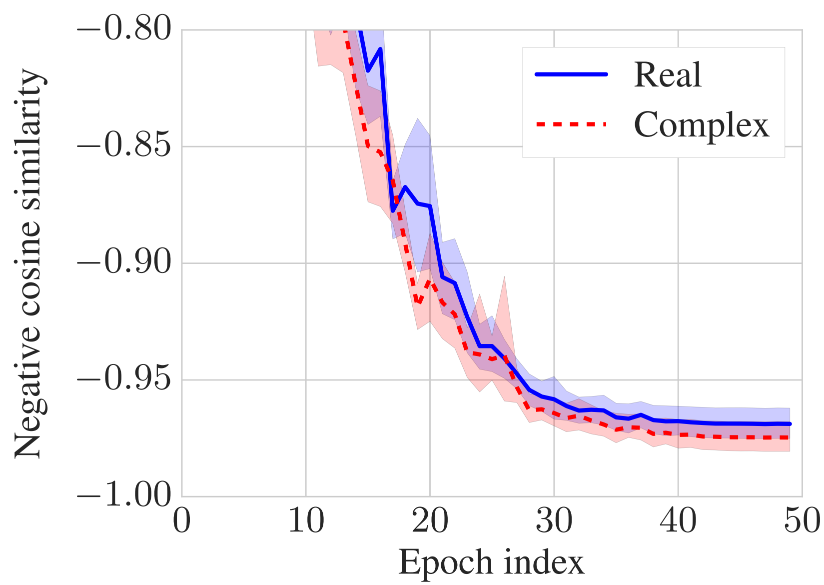




Learn PCA

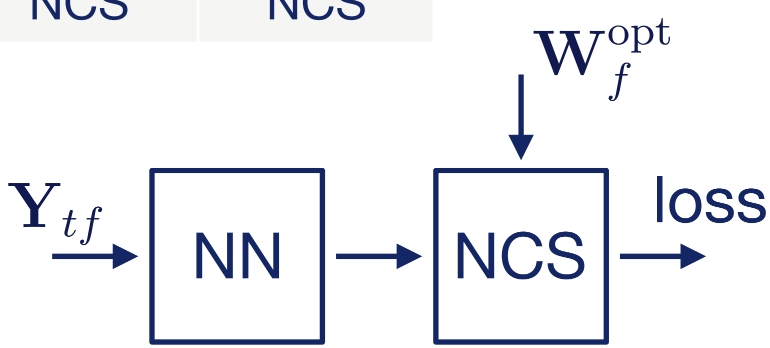
RVNN	CVNN
$2D^2 \rightarrow 50$	$D^2 \rightarrow 25$
ReLU	SplitReLU
$50 \rightarrow 2D$	$25 \rightarrow D$
NCS	NCS

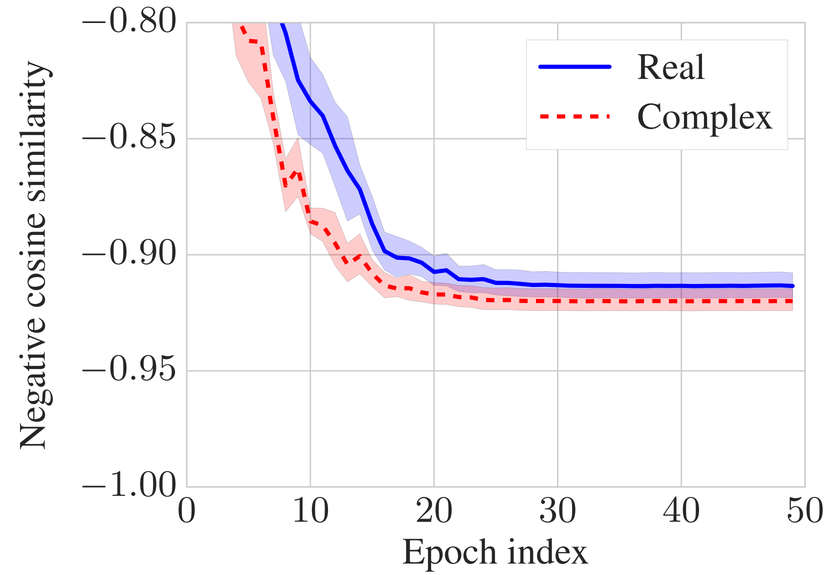


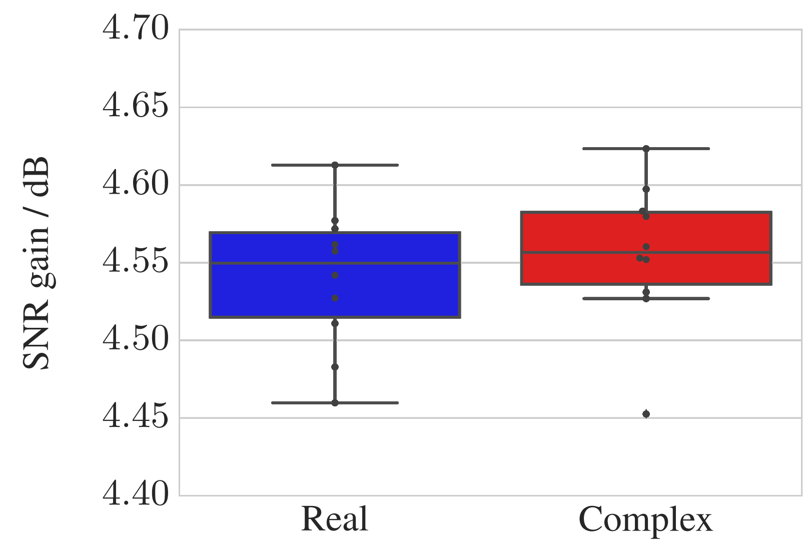


Learn both

RVNN	CVNN
2D → 50	D → 25
ReLU	SplitReLU
50 → 2D^2	25 → D^2
Pooling	Pooling
2D^2 → 50	D^2 → 25
ReLU	SplitReLU
50 → 2D	25 → D
NCS	NCS







Summary/ Discussion

- Limited benefits in (this) classical feed forward setting, use real-valued NNs
- Use complex-valued networks to propagate through…
  - …complex valued PCA or other beamformer,
  - …feature extraction pipeline.
- Allows to use physically motivated models within deep learning framework.

Summary/ Discussion

- Limited benefits in (this) classical feed forward setting, use real-valued NNs
- Use complex-valued networks to propagate through…
  - …complex valued PCA or other beamformer,
  - …feature extraction pipeline.
- Allows to use physically motivated models within deep learning framework.

Thank you for listening!