Gated Complex Recurrent Neural Networks

Moritz Wolter, Angela Yao

wolter@cs.uni-bonn.de

July 13, 2018

UNIVERSITÄT BONN

Motivation

- RNN and neural networks in general suffer from unstable gradients.
- Distribution over a sum using gating is one fix for vanishing gradients (GRU, LSTM, ...)
- Norm preserving matrices are another way to fix this. $\|\mathbf{W}h\|_2 = \|h\|_2$
- Orthogonal (real) and unitary (complex) matrices are norm preserving.

Motivation

- Unitary matrices are more expressive than orthogonal ones.
- Complex networks must be interoperable with real components.
- Mappings from $\mathbb{C}$ to $\mathbb{R}$ are not complex differentiable.

Wirtinger-Calculus [Wir27][MG09][KD09]

For a complex function $f(z) = u(x, y) - iv(x, y)$ we have:

$$\mathbb{R}\text{-derivative} \triangleq \frac{\partial f}{\partial z}\Big|_{\bar{z}=\text{const}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\frac{\partial f}{\partial y}\right), \tag{1}$$

$$\overline{\mathbb{R}}\text{-derivative} \triangleq \frac{\partial f}{\partial \bar{z}}\Big|_{z=\text{const}} = \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\frac{\partial f}{\partial y}\right). \tag{2}$$

Based on these derivatives, one can define the chain rule for a function $g(f(z))$ as follows:

$$\frac{\partial g(f(z))}{\partial z} = \frac{\partial g}{\partial f}\frac{\partial f}{\partial z} + \frac{\partial g}{\partial \bar{f}}\frac{\partial \bar{f}}{\partial z} \text{ where } \bar{f} = u(x, y) - iv(x, y). \tag{3}$$

Unitary Evolution matrix RNN-Motivation [ASB16][Pas13]

$$\mathbf{x}_t = \mathbf{W}_{\text{rec}} f(\mathbf{x}_{t-1}) + \mathbf{W}_{\text{in}} \mathbf{u}_t + \mathbf{b}. \tag{4}$$

$$\frac{\partial \mathcal{E}}{\partial \theta} = \sum_{1 \leq t \leq T} \frac{\mathcal{E}_t}{\partial \theta}, \tag{5}$$

$$\frac{\partial \mathcal{E}_t}{\partial \theta} = \sum_{1 \leq k \leq t} \left(\frac{\mathcal{E}_t}{\partial \mathbf{x}_t} \frac{\partial \mathbf{x}_t}{\mathbf{x}_k} \frac{\partial^+ \mathbf{x}_k}{\partial \theta} \right), \tag{6}$$

$$\frac{\partial \mathbf{x}_t}{\partial \mathbf{x}_k} = \prod_{t \geq i > k} \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_{i-1}} = \prod_{t \geq i > k} W_{\text{rec}}^T \text{diag}(f'(\mathbf{x}_{i-1})). \tag{7}$$

Stiefel Manifold Weight Updates [WPH+16]

$$\mathbf{W}_{k+1} = (\mathbf{I} + \frac{\lambda}{2}\mathbf{A}_k)^{-1}(\mathbf{I} - \frac{\lambda}{2}\mathbf{A}_k)\mathbf{W}_k, \tag{8}$$

where $$\mathbf{A} = \mathbf{W}\overline{\nabla_{\mathbf{w}} F}^T - \overline{\mathbf{W}}^T \nabla_{\mathbf{w}} F. \tag{9}$$

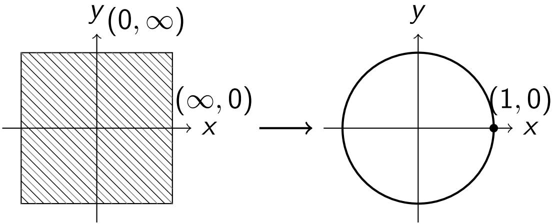


Figure: Fix the optimized matrix eigenvalues onto the unit circle. The key idea behind stiefel-manifold optimization.

Unitary evolution network performance

$$\mathbf{x}_t = \mathbf{U}_{\text{rec}} f(\mathbf{x}_{t-1}) + \mathbf{W}_{\text{in}} \mathbf{u}_t + \mathbf{b}. \tag{10}$$

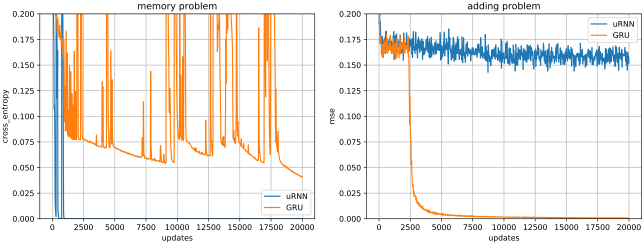
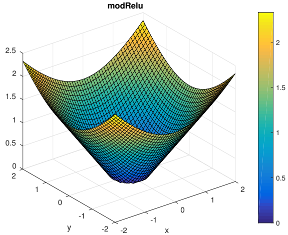
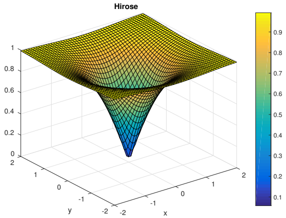


Figure: Current state of the art performance on memory and adding problem for T=250. Models have approximately 40k weights.

Complex equivalents of tanh and Relu





$$f_{\text{Hirose}}(z) = \tanh\left(\frac{|z|}{m^2}\right) e^{-i \cdot \theta_z} = \tanh\left(\frac{|z|}{m^2}\right) \frac{z}{|z|}, \tag{11}$$

$$f_{\text{modReLU}}(z) = \text{ReLU}(|z| + b) e^{-i \cdot \theta_z} = \text{ReLU}(|z| + b) \frac{z}{|z|}. \tag{12}$$

We will compare their performance as state-to-state non-linearities.

Complex gated Recurrent Recurrent Nets

Gate equation:

$$\mathbf{g}_r = f_g(\mathbf{z}_r), \qquad \text{where} \qquad \mathbf{z}_r = \mathbf{W}_r\mathbf{h} + \mathbf{V}_r\mathbf{x}_t + \mathbf{b}_r, \tag{13}$$
$$\mathbf{g}_z = f_g(\mathbf{z}_z), \qquad \text{where} \qquad \mathbf{z}_z = \mathbf{W}_z\mathbf{h} + \mathbf{V}_z\mathbf{x}_t + \mathbf{b}_z, \tag{14}$$

Update equations:

$$\widetilde{\mathbf{z}}_t = \mathbf{W}(\mathbf{g}_r \odot \mathbf{h}_{t-1}) + \mathbf{V}\mathbf{x}_t + \mathbf{b}, \tag{15}$$
$$\mathbf{h}_t = \mathbf{g}_z \odot f_a(\widetilde{\mathbf{z}}_t) + (1 - \mathbf{g}_z) \odot \mathbf{h}_{t-1}, \tag{16}$$

$\mathbb{C} \rightarrow \mathbb{R}$, mapping:

$$\mathbf{o}_r = \mathbf{W}_o[\Re(\mathbf{h})\ \Im(\mathbf{h})] + \mathbf{b}_o. \tag{17}$$

Complex gate activations

$$f_{\text{prod}}(\mathbf{z}) = \sigma(\Re(\mathbf{z})) \cdot \sigma(\Im(\mathbf{z})), \tag{18}$$

$$f_{\text{gate hirose}} = \tanh(\frac{|z|}{m^2})\sigma(a\frac{z}{|z|} + b), \tag{19}$$

$$f_{\text{mod sigmoid}}(\mathbf{z}) = \sigma(\alpha\Re(\mathbf{z}) + \beta\Im(\mathbf{z})). \tag{20}$$

With $\alpha \in [0, 1]$ and $\beta = (1 - \alpha)$.

Comparison to state of the art

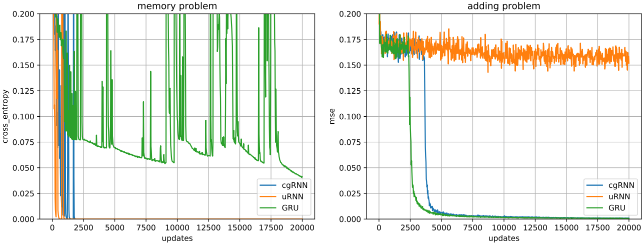


Figure: Comparison of our complex gated RNN (cgRNN, blue, $n_h = 80$) with the unitary RNN [ASB16](uRNN, orange, $n_h = 140$) and standard GRU [CvMG+14](orange, $n_h = 112$) on the memory (left) and adding (right) problem for $T = 250$.

Stiefel optimization and activations

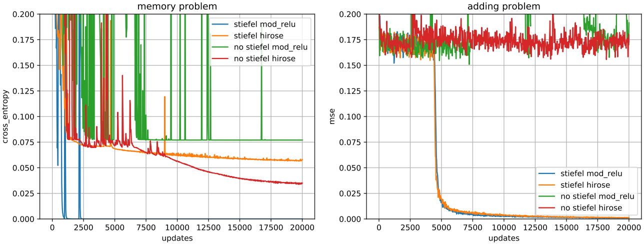
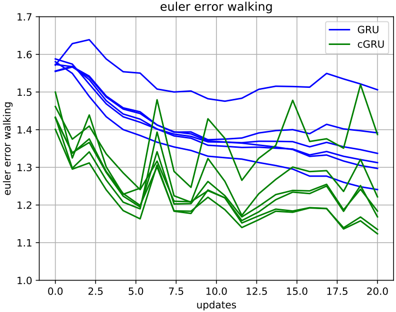


Figure: Comparison of non-linearities and norm preserving state transition matrices on the complex gated RNNs for the memory (a) and adding (b) problems for T=250. We use $n_h = 80$ for all experiments.

Motion prediction

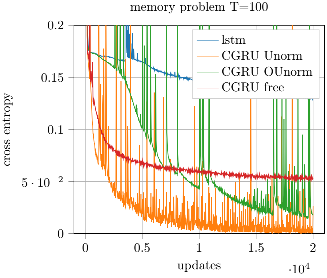


seed	cgRNN-error	GRU-error
0080	**1.13**	1.24
0160	**1.14**	1.30
0320	**1.19**	1.31
0400	**1.17**	1.34
0560	**1.23**	1.39
1000	**1.39**	1.51
average	**1.21**	1.35

Figure: Motion prediction Euler angle errors for the complex gated RNN (green) versus GRU (blue), where each line indicates a separate test sequence. The final error after 20,000 iterations is shown in the adjacent table.

Gates must be able to saturate to work!

In order to further stabilize the gradients we explored normalizing the recurrent matrices in the gate equations



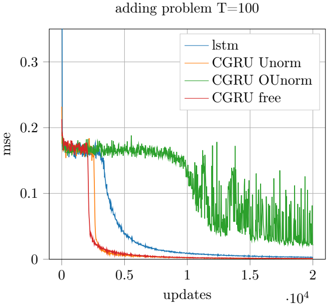


Figure: Orthogonal recurrent gate matrices prevent the gates from functioning.

Future Work

- Complex gate coupling. Just one complex gate equation, $\mathbf{r} = \sigma(\Re(\mathbf{g}))$, $\mathbf{z} = \sigma(\Im(\mathbf{g}))$. Reduces complex overhead.
- Explore frequency domain networks using Hilbert or Fourier transformed input data.
- Explore dynamic mode decomposition as an alternative complex input representation.

References I

- Martin Arjovsky, Amar Shah, and Yoshua Bengio, *Unitary evolution recurrent neural networks*, ICML, 2016.
- Kyunghyun Cho, Bart van Merriënboer, Çalar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio, *Learning phrase representations using RNN encoder–decoder for statistical machine translation*, EMNLP, October 2014.
- Ken Kreutz-Delgado, *The complex gradient operator and the cr-calculus*, arXiv preprint arXiv:0906.4835 (2009).
- Danilo P Mandic and Vanessa Su Lee Goh, *Complex valued nonlinear adaptive filters: noncircularity, widely linear and neural models*, vol. 59, John Wiley & Sons, 2009.

References II

- Pascanu, *On the difficulty of training recurrent neural networks*, Journal of Machine Learning Research (2013).
- W. Wirtinger, *Zur formalen theorie der funktionen von mehr komplexen veränderlichen*, 1927.
- Scott Wisdom, Thomas Powers, John R. Hershey, Jonathan Le Roux, , and Les Atlas, *Full-capacity unitary recurrent neural networks*, Advances in Neural Information Processing Systems, 2016.

Feedback

Thanks for your attention and feedback.
Later: wolter@cs.uni-bonn.de