Time Series Analysis Tutorial

Exercise 10

Leonard Bereska

Central Institute of Mental Health

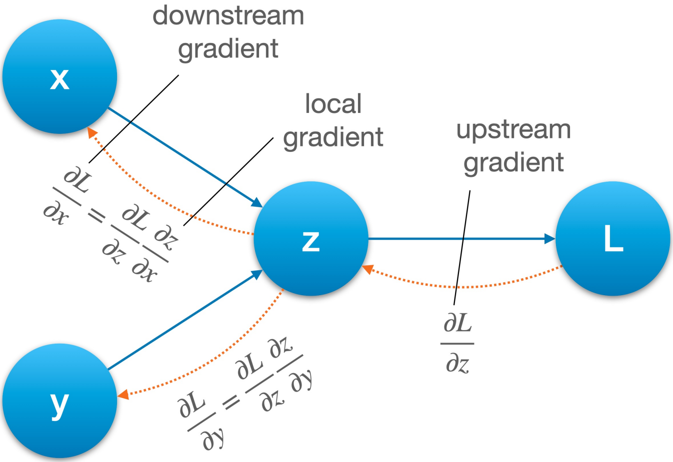
June 30, 2021

Zentralinstitut für Seelische Gesundheit
Landesstiftung des öffentlichen Rechts

Universität Heidelberg
Zukunft seit 1386

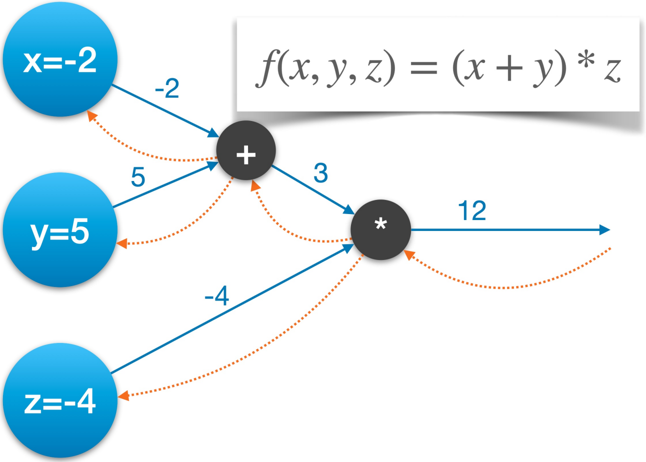
Vanishing & Exploding gradients

Backpropagation



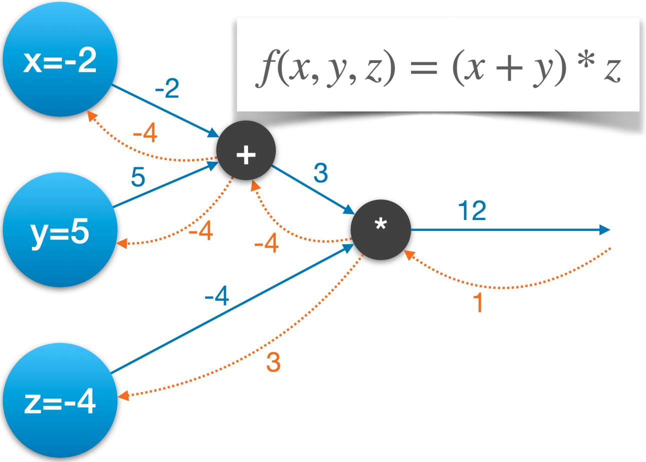
Downstream = Upstream * Local gradient

BACKPROPAGATION EXAMPLE



- Multiplication swaps input activations
- Addition passes gradient on

Backpropagation example



- Multiplication swaps input activations
- Addition passes gradient on

Vanishing & Exploding gradients

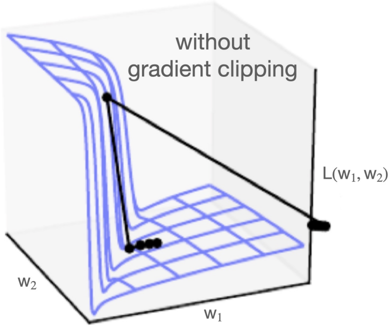
- Recurrent connections, e.g. $h_t = \tanh(Wh_{t-1} + Ux_t + b)$, can be approximated linearly with $h_t = Wh_{t-1}$.
- Repeated application of W leads to exponential growth along the eigenvector direction of maximum absolute eigenvalue:[1]

$$h_t = W^t h_0$$
$$h_t = Q\Lambda^t Q^T h_0$$

Hence maximum absolute eigenvalue indicates both stability as the occurrence explosion or vanishing of gradients.

[1]If W admits an eigendecomposition: $W^t = (Q\Lambda Q^T)^t = Q\Lambda^t Q^T$.

Gradient Clipping [2]



- High curvature walls lead to bad gradient updates without clipping

Image from *Deep Learning* [1]

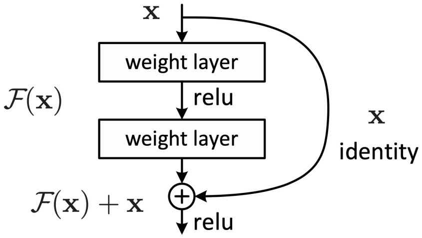
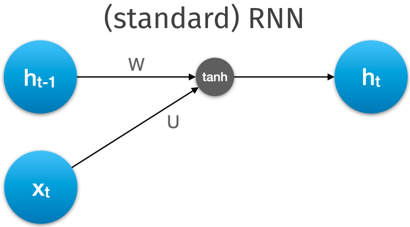


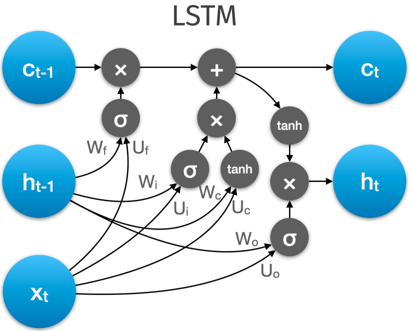
Image from He et al., 2015 [3]

- Residual connections $\rightarrow$ *ResNet*
- Addition of identity maintains information
- Deep network training much easier

Long short term memory (LSTM) [4]

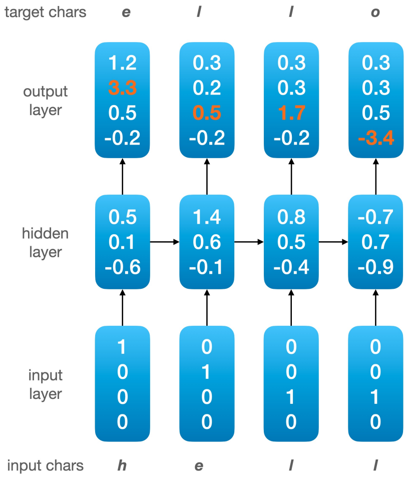


$$h_t = \tanh(Wh_{t-1} + Ux_t + b)$$



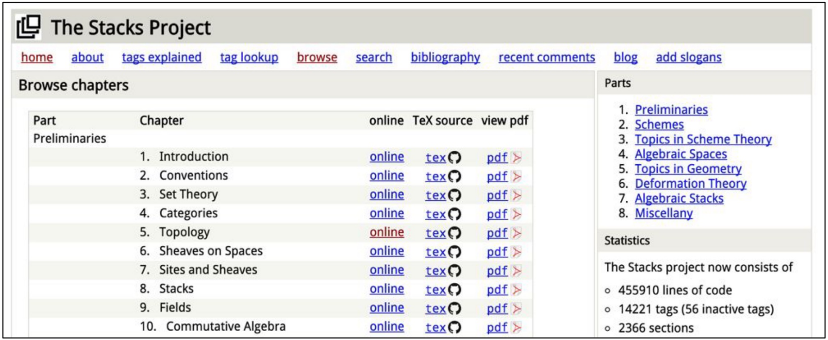
$$f_t = \sigma(W_f h_{t-1} + U_f x_t + b_f)$$
$$i_t = \sigma(W_i h_{t-1} + U_i x_t + b_i)$$
$$o_t = \sigma(W_o h_{t-1} + U_o x_t + b_o)$$
$$\tilde{c}_t = \tanh(W_c h_{t-1} + U_c x_t + b_c)$$
$$c_t = f_t \circ c_{t-1} + i_t \circ \tilde{c}_t$$
$$h_t = o_t \circ \tanh(c_t)$$

Example: Char-level RNN



From `http://vision.stanford.edu/teaching/cs231n/slides/2016/winter1516_lecture10.pdf`

Dataset: open source textbook



From `http://vision.stanford.edu/teaching/cs231n/slides/2016/winter1516_lecture10.pdf`

Generated Latex

For $\bigoplus_{n=1,\ldots,m}$ where $\mathcal{L}_{m_\bullet} = 0$, hence we can find a closed subset $\mathcal{H}$ in $\mathcal{H}$ and any sets $\mathcal{F}$ on X, U is a closed immersion of S, then $U \to T$ is a separated algebraic space.

Proof. Proof of (1). It also start we get

$$S = \mathrm{Spec}(R) = U \times_X U \times_X U$$

and the comparicoly in the fibre product covering we have to prove the lemma generated by $\coprod Z \times_U U \to V$. Consider the maps M along the set of points Sch_{fppf} and $U \to U$ is the fibre category of S in U in Section, **??** and the fact that any U affine, see Morphisms, Lemma **??**. Hence we obtain a scheme S and any open subset $W \subset U$ in $Sh(G)$ such that $\mathrm{Spec}(R') \to S$ is smooth or an

$$U = \bigcup U_i \times_{S_i} U_i$$

which has a nonzero morphism we may assume that f_i is of finite presentation over S. We claim that $\mathcal{O}_{X,x}$ is a scheme where $x, x', s'' \in S'$ such that $\mathcal{O}_{X,x'} \to \mathcal{O}'_{X',x'}$ is separated. By Algebra, Lemma **??** we can define a map of complexes $\mathrm{GL}_{S'}(x'/S'')$ and we win. □

To prove study we see that $\mathcal{F}|_U$ is a covering of $\mathcal{X}'$, and $\mathcal{T}_i$ is an object of $\mathcal{F}_{X/S}$ for $i > 0$ and $\mathcal{F}_p$ exists and let $\mathcal{F}_i$ be a presheaf of $\mathcal{O}_X$-modules on $\mathcal{C}$ as a $\mathcal{F}$-module. In particular $\mathcal{F} = U/\mathcal{F}$ we have to show that

$$\widetilde{M}^\bullet = \mathcal{I}^\bullet \otimes_{\mathrm{Spec}(k)} \mathcal{O}_{S,s} - i_X^{-1}\mathcal{F})$$

is a unique morphism of algebraic stacks. Note that

$$\mathrm{Arrows} = (Sch/S)^{opp}_{fppf}, (Sch/S)_{fppf}$$

and

$$V = \Gamma(S, \mathcal{O}) \longmapsto (U, \mathrm{Spec}(A))$$

is an open subset of X. Thus U is affine. This is a continuous map of X is the inverse, the groupoid scheme S.

Proof. See discussion of sheaves of sets. □

The result for prove any open covering follows from the less of Example **??**. It may replace S by $X_{spaces,étale}$ which gives an open subspace of X and T equal to S_{Zar}, see Descent, Lemma **??**. Namely, by Lemma **??** we see that R is geometrically regular over S.

Lemma 0.1. *Assume (3) and (3) by the construction in the description.*

Suppose $X = \lim |X|$ *(by the formal open covering* X *and a single map* $\underline{Proj}_X(\mathcal{A}) = \mathrm{Spec}(B)$ *over* U *compatible with the complex*

$$Set(\mathcal{A}) = \Gamma(X, \mathcal{O}_{X,\mathcal{O}_X}).$$

When in this case of to show that $\mathcal{Q} \to \mathcal{C}_{Z/X}$ *is stable under the following result in the second conditions of (1), and (3). This finishes the proof. By Definition* **??** *(without element is when the closed subschemes are catenary. If* T *is surjective we may assume that* T *is connected with residue fields of* S. *Moreover there exists a closed subspace* $Z \subset X$ *of* X *where* U *in* X' *is proper (some defining as a closed subset of the uniqueness it suffices to check the fact that the following theorem*

(1) f *is locally of finite type. Since* $S = \mathrm{Spec}(R)$ *and* $Y = \mathrm{Spec}(R)$.

Proof. This is form all sheaves of sheaves on X. But given a scheme U and a surjective étale morphism $U \to X$. Let $U \cap U = \coprod_{i=1,\ldots,n} U_i$ be the scheme X over S at the schemes $X_i \to X$ and $U = \lim_i X_i$. □

The following lemma surjective restrocomposes of this implies that $\mathcal{F}_{x_0} = \mathcal{F}_{x_0} = \mathcal{F}_{\mathcal{X},\ldots,0}$.

Lemma 0.2. *Let* X *be a locally Noetherian scheme over* S, $E = \mathcal{F}_{X/S}$. *Set* $\mathcal{I} = \mathcal{J}_1 \subset \mathcal{I}'_n$. *Since* $\mathcal{I}^n \subset \mathcal{I}^n$ *are nonzero over* $i_0 \leq \mathfrak{p}$ *is a subset of* $\mathcal{J}_{n,0} \circ \overline{A}_2$ *works.*

Lemma 0.3. *In Situation* **??**. *Hence we may assume* $\mathfrak{q}' = 0$.

Proof. We will use the property we see that $\mathfrak{p}$ is the mext functor (**??**). On the other hand, by Lemma **??** we see that

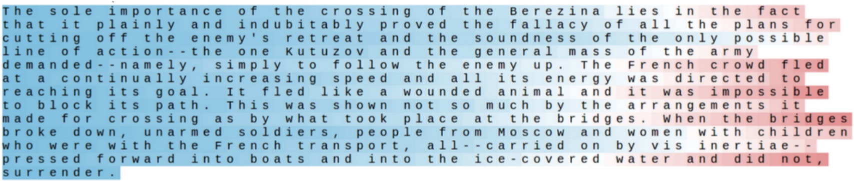
$$D(\mathcal{O}_{X'}) = \mathcal{O}_X(D)$$

where K is an F-algebra where δ_{n+1} is a scheme over S. □

From http://vision.stanford.edu/teaching/cs231n/slides/2016/winter1516_lecture10.pdf

Visualizing activation [5]

Cell indicating position in line



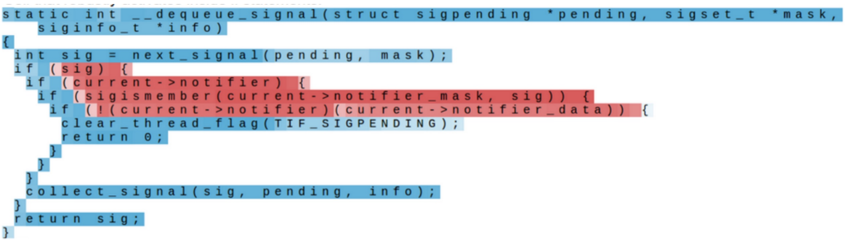
Visualizing activation [5]

Cell indicating depth of expression

```
#ifdef CONFIG_AUDITSYSCALL
static inline int audit_match_class_bits(int class, u32 *mask)
{
  int i;
  if (classes[class]) {
    for (i = 0; i < AUDIT_BITMASK_SIZE; i++)
      if (mask[i] & classes[class][i])
        return 0;
  }
  return 1;
}
```

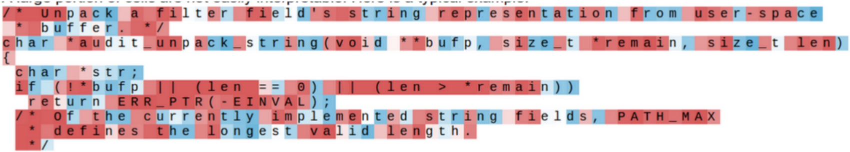
Cell active inside comments

Cell active inside if statements



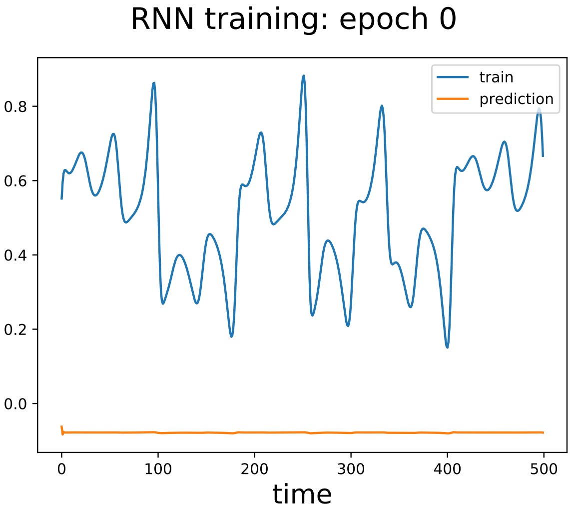
VISUALIZING ACTIVATION [5]

Cell activation that is hard to interpret

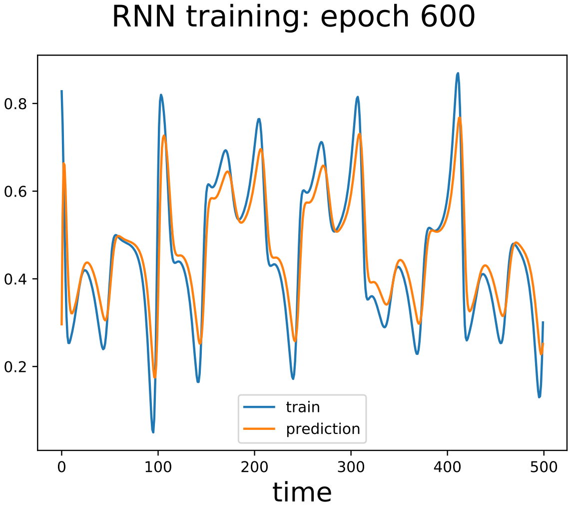


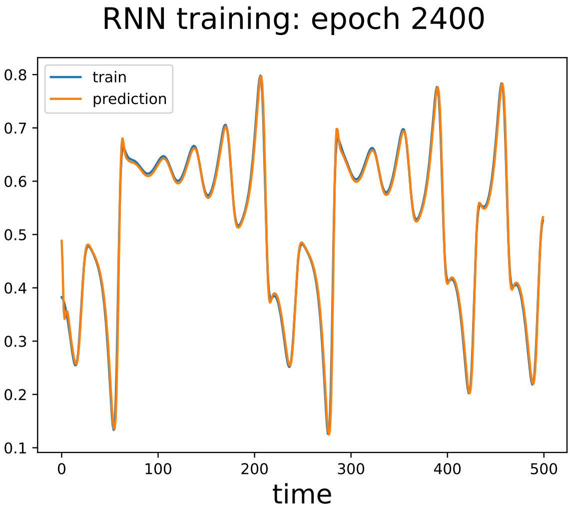
Most cells are activating in this manner!

Solution exercise 10

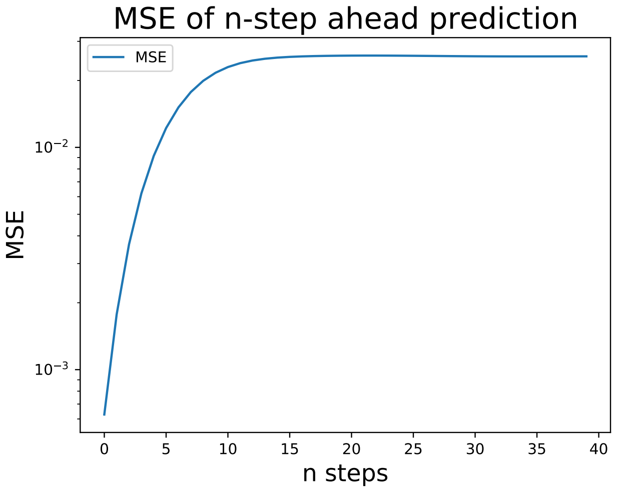

RNN training: epoch 0
train
prediction
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
time

1.4 RNN TRAINING



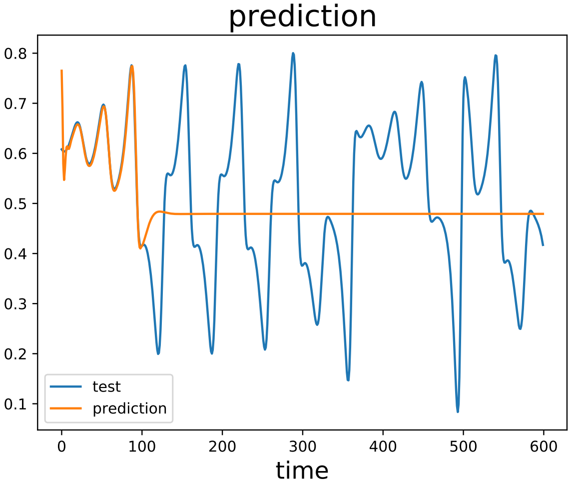

RNN training: epoch 2400
train
prediction
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
time

1.4 RNN MSE

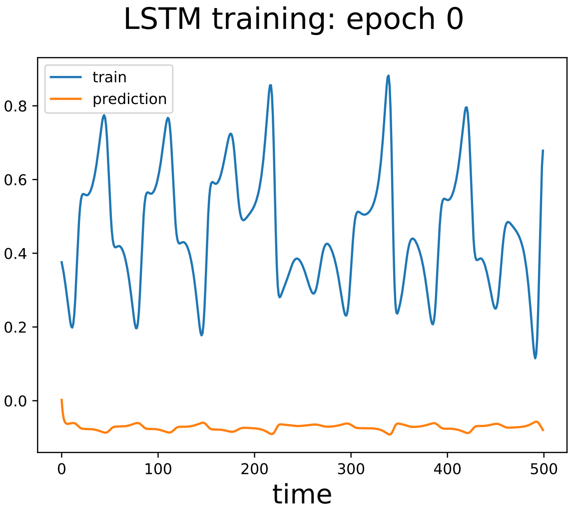


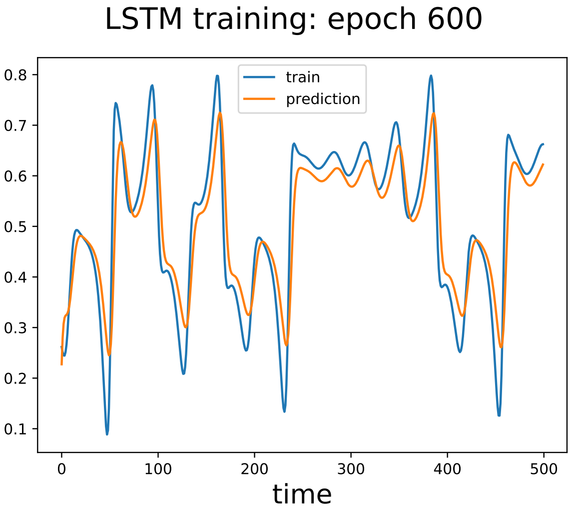
MSE of n-step ahead prediction
MSE
10^{-2}
10^{-3}
MSE
0
5
10
15
20
25
30
35
40
n steps

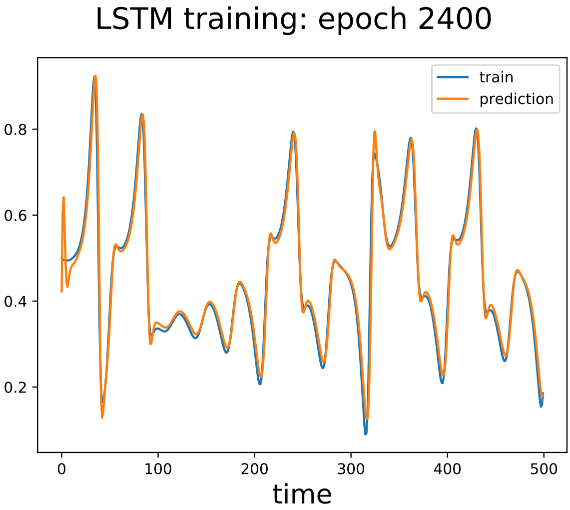
1.4 RNN PREDICTION



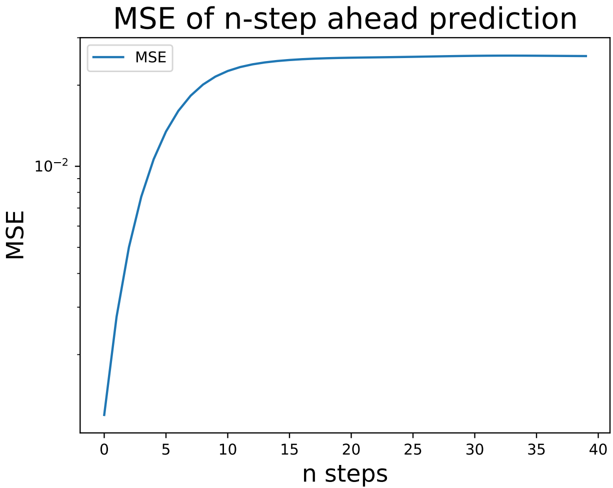
1.4 LSTM TRAINING

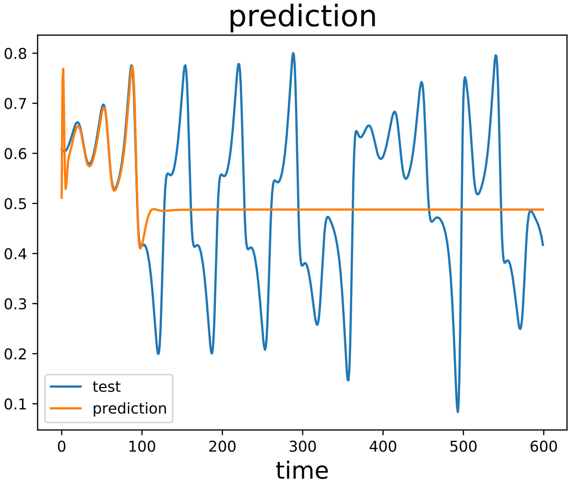



LSTM training: epoch 600
train
prediction
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
time

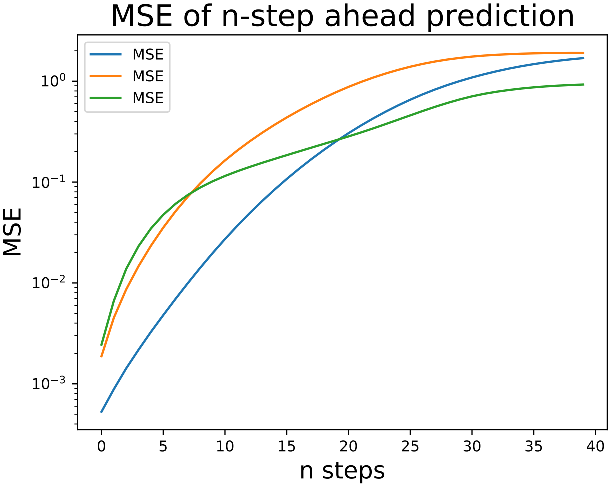

LSTM training: epoch 2400
train
prediction
0.8
0.6
0.4
0.2
0
100
200
300
400
500
time

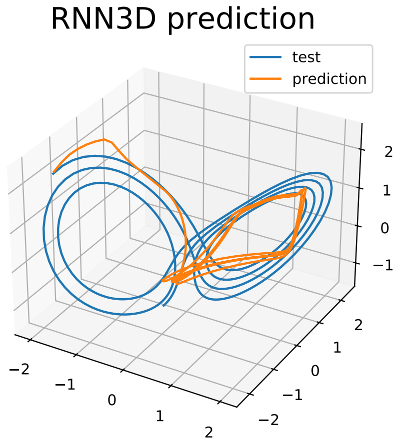
1.4 LSTM MSE


MSE of n-step ahead prediction
MSE
10^{-2}
MSE
0
5
10
15
20
25
30
35
40
n steps

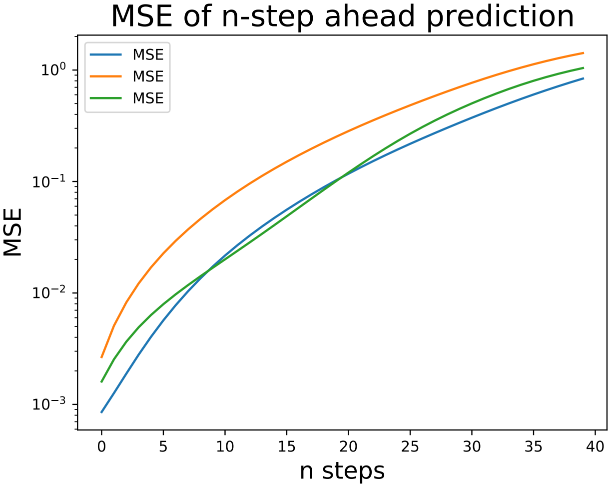

prediction
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
100
200
300
400
500
600
time
test
prediction

1.4 RNN3D MSE

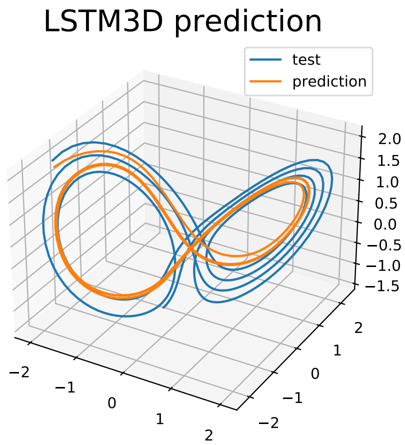

MSE of n-step ahead prediction
MSE
MSE
MSE
10^0
10^{-1}
10^{-2}
10^{-3}
MSE
0
5
10
15
20
25
30
35
40
n steps


RNN3D prediction
test
prediction
2
1
0
−1
−2
−1
0
1
2
2
1
0
−1
−2

1.4 LSTM3D MSE



1.4 LSTM3D PREDICTION



LECTURE REVIEW

Extended Kalman Filter

Assume:

$$
\begin{aligned}
z_1 &\sim \mathcal{N}(\mu_0, \Sigma_0) \\
z_t &= F_\theta(z_{t-1}) + \epsilon_t,\ \epsilon_t \sim \mathcal{N}(0, \Sigma) \\
x &= Bz_t + \eta_t,\ \eta_t \sim \mathcal{N}(0, \Gamma)
\end{aligned}
$$

Idea: linearize non-linear functions

$$F_\theta(z_{t-1}) \approx F_\theta(\mu_{t-1}) + F'_\theta(\mu_{t-1})(z_{t-1} - \mu_{t-1})$$

EKF update

$$
\begin{aligned}
\mu_t &= F(\mu_{t-1} + K_t[x_t - G(F(\mu_{t-1}))]) \\
V_t &= L_{t-1} - K_t \nabla_{t-1} L_{t-1} \\
L_{t-1} &= J_{t-1} V_{t-1} J_{t-1}^T + \Sigma \\
K_t &= L_{t-1} \nabla_{t-1}^T (\nabla_{t-1} L_{t-1} \nabla_{t-1}^T + \Gamma)^{-1}
\end{aligned}
$$

$$
\nabla_{ij,t-1} := \frac{\partial G_i}{\partial F(\mu_{j,t-1})}, \; J_{ij,t-1} := \frac{\partial F_i}{\partial \mu_{j,t-1}}
$$

References I

Ian Goodfellow, Yoshua Bengio, and Aaron Courville.
Deep Learning.
MIT Press, 2016.
http://www.deeplearningbook.org.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio.
On the difficulty of training Recurrent Neural Networks.
arXiv:1211.5063 [cs], February 2013.
arXiv: 1211.5063.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.
Deep Residual Learning for Image Recognition.
arXiv:1512.03385 [cs], December 2015.
arXiv: 1512.03385.

S. Hochreiter and J. Schmidhuber.
Long short-term memory.
Neural Computation, 9:1735–1780, 1997.

References II

Andrej Karpathy, Justin Johnson, and Li Fei-Fei.
Visualizing and Understanding Recurrent Networks.
arXiv:1506.02078 [cs], November 2015.
arXiv: 1506.02078.