How a Nonlinear Version of the KKL Observer Can Provide Estimation Guarantees for some RNN-Based Algorithms

V. Andrieu

Works in collaboration with P. Bernard, L. Brivadis, V. Pachy, L. Praly

June 4, 2024

Approaching the observation problem

Observation Problem:

Estimate state variables (x) from measured variables (y)

- The "object" solving this problem is called an observer
- Measurements make what is called the a posteriori information It evolves with time as data accumulate
- we have also a priori information: a model that links x and y!

$$\dot{x} = f(x) \ , \ \ y = h(x) \ , \ \ x \in \mathbb{R}^n$$

Observer:

$$\begin{matrix} \{t \mapsto y(t)\} \\ (f, h) \end{matrix} \quad \rightarrow \boxed{\text{Observer}} \rightarrow \text{Estimated state variables } \hat{x}(t)$$

Dynamic observer approach

General structure of the observer:

- History of the measurement is stored in a finite dimensional "state"
- The estimate is given as a function of this state

Observer:

$$\{t \mapsto y(t)\} \rightarrow \boxed{\begin{array}{c}\text{Dynamical System}\\ \dot{z}(t) = \varphi(z(t), y(t))\\ z \in \mathbb{R}^m\end{array}} \rightarrow \boxed{\begin{array}{c}\text{Observer output}\\ \hat{x}(t) = \gamma(z(t))\end{array}} \rightarrow \hat{x}(t)$$

Observer question : How to design φ and γ such that $\hat{x}(t)$ is a good estimate of $x(t)$

Asking a computer science guy to solve the problem

The case of linear activation functions

The case of monotonic activation functions

Why employing nonlinear activation functions?

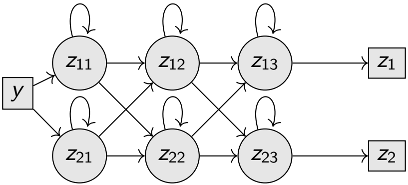
Conclusion

A popular approach in computer science

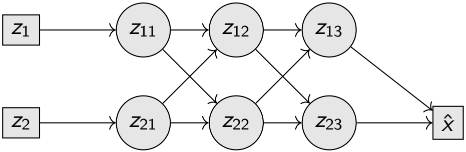
A computer science observer structure:

$\{t \mapsto y(t)\} \rightarrow$ Dynamical System *RNN* $\rightarrow$ Observer output *MLP* $\rightarrow \hat{x}(t)$

RNN is a recurrent neural network



MLP is a Multilayer Perceptron



A universal approach

- *RNN* and *MLP* depend on activation functions and parameters denoted Ω
- In the simplest case, a continuous time model of a *RNN* with one layer:

$$\dot{z}_i = W_0 \sigma (W_1 z_i + W_2 y + W_3)$$

where σ is an activation function and the $\Omega = (W_i)$ are parameters (matrices)

Computer science approach for state observer

1. Define a cost which quantifies what is a good estimate
2. Optimize the parameters of *RNN* and *MLP* based on data (or model) to get an observer

Question: Can we give guaranty that it may work ? Can it be tunnable ?

Tunnable observers

$$\{t \mapsto y(t)\} \rightarrow \boxed{\begin{array}{c}\text{Dynamical System}\\ RNN_{\Omega}\end{array}} \rightarrow \boxed{\begin{array}{c}\text{Observer output}\\ MLP_{\Omega}\end{array}} \rightarrow \hat{x}(t)$$

Tunnable observer structure

Given

- a compact (invariant) set $\mathcal{X} \subset \mathbb{R}^n$ of initial conditions
- an observation time t_o
- an estimation treshold ϵ

There exist parameters Ω such that

$$|\hat{x}(t) - x(t)| \leq \epsilon \ , \ \forall t > t_o \ , \forall \ (x_0, z_0) \in \mathcal{X} \times \mathcal{Z}_0$$

Question: For which activation functions σ in the *RNN* do we get this property ?

Asking a computer science guy to solve the problem

The case of linear activation functions

The case of monotonic activation functions

Why employing nonlinear activation functions?

Conclusion

A particular case

The case of a linear activation function in the *RNN*

$$\dot{z}_i = k\lambda_i z_i + y \ , \ i = 1, \dots, m$$

$\Rightarrow$ We recognize KKL observer dynamics.

KKL Paradigme: *If the system is observable, picking m sufficiently large, there exists* $\mathbf{T}^{\mathrm{inv}} : \mathbb{R}^m \mapsto \mathbb{R}^n$*, such that* $\hat{x}(t) = \mathbf{T}^{\mathrm{inv}}(z(t))$ *gives a KKL asymptotic observer !*

- Local version: Shoshitaishvili-90, Kazantzis-Kravaris-98
- Global version: Kreisselmeier-Engel-2003, VA-Praly-2006, Brivadis-VA-Bernard-Serres-2023
- Time varying version: Bernard-VA 2019
- Discrete time version: Tran-Bernard 2024

KKL observers

Given m linear filters

$$\dot{z}_i = k\lambda_i z_i + y \ , \ i = 1, \ldots, m$$

KKL is a two steps procedure

Step 1: The state of the filter provides new information

Theorem (VA-Praly-2006)

Let $\mathcal{X}$ be a compact invariant subset of $\mathbb{R}^n$. For all $k > 0$ and all $\lambda_1, \ldots, \lambda_m$ negative, there exists a (continuous) function $\mathbf{T}_k : \mathbb{R}^m \mapsto \mathbb{R}$ such that

$$|z(t) - \mathbf{T}_k(x(t))| \leq e^{-k \max_i \{\lambda_i\} t} |z_0 - \mathbf{T}_k(x_0)| \ , \ \forall (z_0, x_0) \in \mathbb{R}^m \times \mathcal{X}$$

$\Rightarrow$ If $\mathbf{T}_k$ is invertible, we get a state observer

KKL observers

Assumption: Differential observability in $\mathcal{X}$

There exists an integer $m \geq 1$ such that the map $\mathbf{H}_m : \mathbb{R}^n \to \mathbb{R}^m$ defined by:

$$\mathbf{H}_m : x \mapsto \begin{pmatrix} h(x) & L_f h(x) & \dots & L_f^{m-1} h(x) \end{pmatrix}$$

is Lipschitz injective on $\mathcal{X}$.

Theorem (VA-Praly-2006, VA-2014)

Let $\mathcal{X} \subset \mathbb{R}^n$ be compact invariant. There exists k^* such for all $k \geq k^*$, $\mathbf{T}_k$ is C^1 and Lipschitz injective

If $\mathbf{T}_k$ is injective, there exists $\mathbf{T}^{\mathrm{inv}}$ such that $\mathbf{T}^{\mathrm{inv}}(\mathbf{T}_k(x)) = x$!

KKL observers

An (asymptotic) observer is:

$$\hat{x} = \mathbf{T}^{\mathrm{inv}}(z) \ , \ \dot{z}_i = k\lambda_i z_i + y \ , \ i = 1, \dots, m$$

Theorem (VA-2014)

Let $\mathcal{X} \subset \mathbb{R}^n$ be compact invariant. There exists k^* such for all $k \geq k^*$, there exists a C^1 mapping $\mathbf{T}^{\mathrm{inv}} : \mathbb{R}^m \mapsto \mathbb{R}^n$ and a constant c such that

$$|\mathbf{T}^{\mathrm{inv}}(z(t)) - x(t)| \leq ck^m e^{-k \max_i\{\lambda_i\} t}(|z_0| + 1) \ , \forall (z_0, x_0) \in \mathbb{R}^m \times \mathcal{X}$$

$\Rightarrow$ For each (ϵ, t_o) there exists k^* such that for all $k \geq k^*$

$$|\mathbf{T}^{\mathrm{inv}}(z(t)) - x(t)| \leq \epsilon \ , \ \forall t > t_o \ , \forall \ (x_0, z_0) \in \mathcal{X} \times \mathcal{Z}_0$$

$\Rightarrow$ We have a tunnable asymptotic observer

Question: How to compute $\mathbf{T}^{\mathrm{inv}}$?

MLP as approximator of $\mathbf{T}^{\mathrm{inv}}$

$\mathbf{T}^{\mathrm{inv}}$ is C^1 $\Rightarrow$ MLP can approximate it !

Universal Approximation Theorem my MLP (Cybenko 80')

There exist activation functions such that for each ϵ, for each compact $\mathcal{Z} \subset \mathbb{R}^m$ there exists parameters Ω_{MLP} such that with $\gamma(z) = MLP_{\Omega_{MLP}}(z)$

$$\sup_{z \in \mathcal{Z}} |\gamma(z) - \mathbf{T}^{\mathrm{inv}}(z)| \leq \epsilon$$

Hence, with $\hat{x}(t) = \gamma(z(t))$

$$|\hat{x}(t) - x(t)| \leq \underbrace{|\gamma(z(t)) - \mathbf{T}^{\mathrm{inv}}(z(t))|}_{\leq \epsilon} + \underbrace{|\mathbf{T}^{\mathrm{inv}}(z(t)) - x(t)|}_{\leq \epsilon}$$

Theorem

$$y \to \boxed{\begin{array}{c}\text{Linear Filter}\\ \dot{z}_i = k\lambda_i z_i + y\end{array}} \to \boxed{\begin{array}{c}\text{Observer output}\\ MLP_{\Omega_{MLP}}\end{array}} \to \hat{x}(t)$$

is a tunable observer structure

Question: What can we say for motonic activation function ?

Asking a computer science guy to solve the problem

The case of linear activation functions

The case of monotonic activation functions

Why employing nonlinear activation functions?

Conclusion

A contracting nonlinear filter dynamics

Consider a continuous time model of RNN

$$\dot{z}_i = k\lambda_i \sigma(z_i, y)$$

Where the function σ satisfies:

$$0 < \gamma \leq \left| \frac{\partial \sigma}{\partial y}(z, y) \right| \ , \ -\beta \leq \frac{\partial \sigma}{\partial z}(z, y) \leq -\alpha < 0$$

- In the following, $k >> 1$ we are following a high-gain approach
- λ_i are taken different for each i

Question: Can we follow the same procedure as the linear case ?

A contraction

It can be noticed that the map $\boldsymbol{\sigma}$ verifies

$$\frac{\partial \boldsymbol{\sigma}}{\partial \mathbf{z}}(\mathbf{z}, y) + \frac{\partial \boldsymbol{\sigma}}{\partial \mathbf{z}}(\mathbf{z}, y)^\top < -\mu \, \mathrm{I}_m \ , \forall (\mathbf{z}, y) \in \mathbb{R}^m \times \mathbb{R},$$

$\Rightarrow$ This system defines a contraction

Theorem (Pachy-VA-Bernard-Brivadis-Praly-2024)

Let $\mathcal{X} \subset \mathbb{R}^n$ be a compact invariant set. For all $(\lambda_0, \dots, \lambda_{m-1})$ and for all $k > 0$ there exists a continuous function $\mathbf{T}_k : \mathbb{R}^n \to \mathbb{R}^m$ such that,

$$|z(t) - \mathbf{T}_k(\hat{x}(t))| \leq e^{-\alpha k \max_i \{\lambda_i\} t} |z_0 - \mathbf{T}_k(\hat{x}_0)|$$

Sketch of the proof:

- $\mathcal{X}$ is invariant $\Rightarrow t \mapsto y(t)$ is a bounded signal in $\mathbb{R}$
- Pavlov 2004 $\Rightarrow \exists$ a unique bounded solution $t \in \mathbb{R} \mapsto \bar{z}(t)$ exp. attractive
- $\mathbf{T}_k(x) = \bar{z}(0)$

Question: What can we say about its regularity and its injectivity ?

A contraction

Assumption: Differential observability in $\mathcal{X}$

There exists an integer $m \geq 1$ such that the map $\mathbf{H}_m : \mathbb{R}^n \to \mathbb{R}^m$ defined by:

$$\mathbf{H}_m : x \mapsto \begin{pmatrix} h(x) & L_f h(x) & \dots & L_f^{m-1} h(x) \end{pmatrix}$$

is Lipschitz injective on $\mathcal{X}$

Theorem (Pachy-VA-Bernard-Brivadis-Praly-2024)

Let $\mathcal{X} \subset \mathbb{R}^n$ be compact invariant. There exists k^* such for all $k \geq k^*$,

- $\mathbf{T}_k$ is C^1 and Lipschitz injective.
- There exists a C^1 mapping $\mathbf{T}^{\text{inv}} : \mathbb{R}^m \mapsto \mathbb{R}^n$ and a constant c such that

$$|\mathbf{T}^{\text{inv}}(z(t)) - x(t)| \leq c k^m e^{-\alpha k \max_i \{\lambda_i\} t} (|z_0| + 1) \ , \ \forall x \in \mathcal{X}$$

$\Rightarrow$ We have a tunnable asymptotic observer

Sketch of the proof

Question: How to check regularity and injectivity ?

Formally, if $\mathbf{T}_k = (\mathbf{T}_{k1}, \ldots, \mathbf{T}_m)$ is C^1, it is solution to the PDE:

$$\frac{\partial \mathbf{T}_{ki}}{\partial x} f(x) = (k\lambda_i)\sigma(\mathbf{T}_{ki}(x), h(x))$$

Key idea: Make an approximation of $\mathbf{T}_{ki}$ in $\frac{1}{(k\lambda_i)^m}$ and work on it !

Approximation of $\mathbf{T}_k$

There exists $\phi_1, \ldots, \phi_m$ such that

$$\mathbf{T}_{ki}(x) = \sum_{\ell=0}^{m-1} \frac{\phi_\ell(x)}{(k\lambda_i)^\ell} + R_i(x)$$

and, if $k >> 1$, there exist positive real numbers (independant of k)

$$|R_i(x)| \leq \frac{c}{k^m} \ , \ |R_i(x_a) - R_i(x_b)| \leq \frac{c}{k^m} |x_a - x_b|$$

Sketch of the proof

In conclusion $\phi(x) = (\phi_1(x), \dots \phi_m(x))$, $\mathbf{R}(x) = (R_1(x), \dots, R_m(x))$

$$\mathbf{T}(x) = \mathcal{V} K^{-1} \phi(x) + \mathbf{R}(x),$$

with $K = \operatorname{diag}\left(1, \dots, k^{m-1}\right)$ and $\mathcal{V}$ is the Vandermonde matrix

$$\mathcal{V} = \begin{pmatrix} 1 & \lambda_0^{-1} & \dots & \lambda_0^{-(m-1)} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \lambda_{m-1}^{-1} & \dots & \lambda_{m-1}^{-(m-1)} \end{pmatrix}. \tag{1}$$

The function ϕ_i depends on $h(x), L_f h(x), \dots, L_f^{i-1} h(x)$ $\Rightarrow$ with observability assumption, ϕ is Lipschitz injective

$\Rightarrow$ There exists c, such that for $k >> 1$

$$|\mathbf{T}(x_a) - \mathbf{T}(x_b)| \geq \frac{c}{k^m} |x_a - x_b|$$

There exists a Lipschitz function $\mathbf{T}^{\mathrm{inv}}$ such that the result holds

$$|\mathbf{T}^{\mathrm{inv}}(z(t)) - x(t)| \leq c k^m e^{-\alpha k \max_i \{\lambda_i\} t} |\mathbf{T}^{\mathrm{inv}}(z_0) - x_0|$$

MLP as approximator of $\mathbf{T}^{\mathrm{inv}}$

$$y \to \boxed{\begin{array}{c}\text{Contractive Filter}\\ \dot{z}_i = k\lambda_i\sigma(z_i, y)\end{array}} \to \boxed{\begin{array}{c}\text{Observer output}\\ \hat{x} = \mathbf{T}^{\mathrm{inv}}(z)\end{array}} \to \hat{x}(t)$$

is a tunnable observer !

$\mathbf{T}^{\mathrm{inv}}$ is globally Lipschitz $\Rightarrow$ MLP can approximate it !

Theorem : With RNN modeled as a continuous time dynamics

$$y \to \boxed{\begin{array}{c}\text{RNN}\\ \dot{z}_i = k\lambda_i\sigma(z_i, y)\end{array}} \to \boxed{\begin{array}{c}\text{Observer output}\\ MLP_{\Omega_{MLP}}\end{array}} \to \hat{x}(t)$$

is a tunable observer structure

Asking a computer science guy to solve the problem

The case of linear activation functions

The case of monotonic activation functions

Why employing nonlinear activation functions?

Conclusion

Nonlinear or linear activation function for the RNN

Is it better to use linear or nonlinear activation functions ?

Given a linear KKL observer:

$$\hat{x} = \mathbf{T}^{\mathrm{inv}}(z) \ , \ \dot{z}_i = k\lambda_i z_i + y \ , \ i = 1, \ldots, m$$

Observer paradigm : Two cases may be distinguished

1. If k is large:
   - Convergence rate is high
   - Less robustness to measurement noise
2. if k is small:
   - Convergence rate is slow
   - Better robustness to measurement noise

Question: How to combine both good points ?

Nonlinear or linear activation function for the RNN

We want

- Fast observer in the transient
- Slow/robust observer at "steady state"

Note that at steady state, $z \approx y$

A possible nonlinear structure for the filter could be

$$\dot{z} = \lambda(a_{\text{fast}}(z - y) + (a_{\text{slow}} - a_{\text{fast}})\tanh(z - y))$$

$\Rightarrow$ monotonic function $\Rightarrow$ We can learn the mapping $\mathbf{T}^{\mathrm{inv}}$

Nonlinear or linear activation function for the RNN

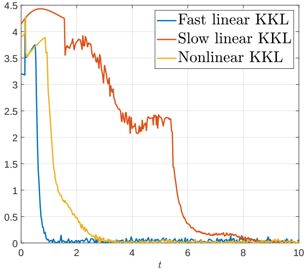
Consider a nonlinear Duffing oscillator

$$\begin{cases} \dot{x}_1 = x_2 \\ \dot{x}_2 = -0.2x_1 - x_1^3 \end{cases}, \quad y = x_1,$$

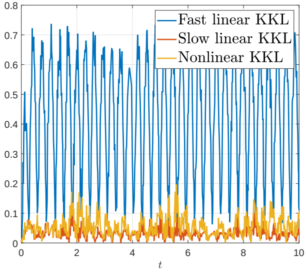
3 activation functions

$$\dot{z} = \lambda(a_{\text{fast}}(z-y) + (a_{\text{slow}} - a_{\text{fast}})\tanh(z-y))$$

$$\dot{z} = \lambda a_{\text{fast}}(z-y), \quad \dot{z} = \lambda a_{\text{slow}}(z-y),$$



(a) Scenario 1: $|\hat{x}(t) - x(t)|$



(b) Scenario 2: $|\hat{x}(t) - x(t)|$

In Conclusion

- It is possible to show that a continuous time model of an observer based on RNN and MLP gives a tunnable observer
- The proof is based on the use of a nonlinear version of KKL observer
- The use of nonlinear KKL observer may be interesting to combine different behavior
- What about discrete time version ?