ILLINOIS
Coordinated Science Laboratory
GRAINGER COLLEGE OF ENGINEERING

Universal Approximation of Input-Output Maps by Temporal Convolutional Nets

Joshua Hanson and Maxim Raginsky
University of Illinois at Urbana-Champaign

CAEML
CENTER FOR ADVANCED ELECTRONICS THROUGH MACHINE LEARNING

Background:

- Convolutional nets are replacing recurrent nets in a growing number of sequence-to-sequence modeling applications (machine translation, audio generation, language modeling). Why?
- Both are inherently suited to modeling systems with limited long-term dependencies.
- Recurrent nets have theoretically infinite memory, but often fail to learn long sequences. Infinite memory is also usually unnecessary in practice.
- The lack of feedback elements in convolutional nets provides a computational advantage (copies of the input sequence can be processed in parallel).

Questions:

- When do convolutional architectures provide better approximation than recurrent architectures?
- How do we quantify "limited long-term dependencies"?

Definitions:

- *i/o map* — nonlinear operator $\mathsf{F} : \mathcal{S} \to \mathcal{S}$ where $\mathcal{S} := \{\boldsymbol{u} = (u_t)_{t\in\mathbb{Z}_+}\}$
- *Right shift* — $(\mathsf{R}\boldsymbol{u})_t := u_{t-1}\mathbf{1}_{\{t\ge 1\}}$ *Window* — $(\mathsf{W}_{t,m}\boldsymbol{u})_\tau := u_\tau \mathbf{1}_{\{\max\{t-m,0\}\le\tau\le t}$
- *Causal* — $\forall t \in \mathbb{Z}_+$, $(u_0,\dots,u_t) =: \boldsymbol{u}_{0:t} = \boldsymbol{v}_{0:t} \implies (\mathsf{F}\boldsymbol{u})_t = (\mathsf{F}\boldsymbol{v})_t$
- *Time-invariant* — $\forall k \in \mathbb{Z}_+$, $(\mathsf{F}\mathsf{R}^k\boldsymbol{u})_t = \begin{cases}(\mathsf{F}\boldsymbol{u})_{t-k}, & \text{for } t \ge k\\ 0, & \text{for } 0 \le t < k\end{cases}$
- *Approximately finite memory* on $\mathcal{M} \subset \mathcal{S}$ — $\forall \epsilon > 0\ \exists m \in \mathbb{Z}_+$,
$$\sup_{\boldsymbol{u}\in\mathcal{M}}\sup_{t\in\mathbb{Z}_+}\left|(\mathsf{F}\boldsymbol{u})_t - (\mathsf{F}\mathsf{W}_{t,m}\boldsymbol{u})_t\right| \le \epsilon$$
The smallest m such that this holds is denoted $m^*_{\mathsf{F}}(\epsilon)$.
- Set of uniformly bounded inputs — $\mathcal{M}(R) := \{\boldsymbol{u} \in \mathcal{S} : \sup_{t\in\mathbb{Z}_+}|u_t| \le R\}$

Main Result:

Assumption 1: The i/o map F has approximately finite memory on $\mathcal{M}(R)$

Assumption 2: For any $t \in \mathbb{Z}_+$, the functional $\tilde{\mathsf{F}}_t : \mathbb{R}^{t+1} \to \mathbb{R}$ is uniformly continuous on $[-R,R]^{t+1}$ with modulus and inverse modulus of continuity

$$\omega_{t,\mathsf{F}}(\delta) := \sup\left\{|\tilde{\mathsf{F}}_t(\boldsymbol{x}) - \tilde{\mathsf{F}}_t(\boldsymbol{x}')| : \boldsymbol{x},\boldsymbol{x}' \in [-R,R]^{t+1}, \|\boldsymbol{x}-\boldsymbol{x}'\|_\infty \le \delta\right\}$$
$$\omega^{-1}_{t,\mathsf{F}}(\epsilon) := \sup\{\delta > 0 : \omega_{t,\mathsf{F}}(\delta) \le \epsilon\}$$

Theorem: Let F be a causal, time-invariant i/o map satisfying Assumptions 1 and 2. Then for any $\epsilon > 0$, $\gamma \in (0,1)$ there exists a ReLU TCN $\widehat{\mathsf{F}}$ with context length $m = m^*_{\mathsf{F}}(\gamma\epsilon)$, width $m+2$, depth $\left(\frac{O(R)}{\omega^{-1}_{m,\mathsf{F}}((1-\gamma)\epsilon)}\right)^{m+2}$ such that

$$\sup_{\|\boldsymbol{u}\|_\infty \le R}\|\mathsf{F}\boldsymbol{u} - \widehat{\mathsf{F}}\boldsymbol{u}\|_\infty < \epsilon$$

Remark: The role of γ is to trade off context length and depth. Tuning γ can reduce the number of computation units needed to achieve ϵ-accuracy.

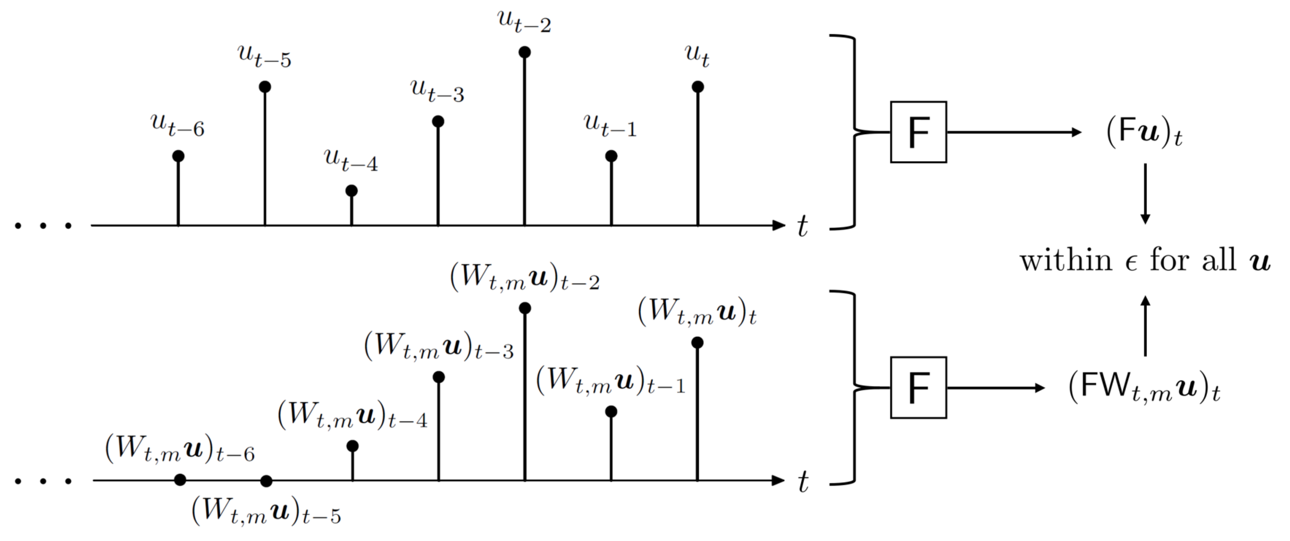
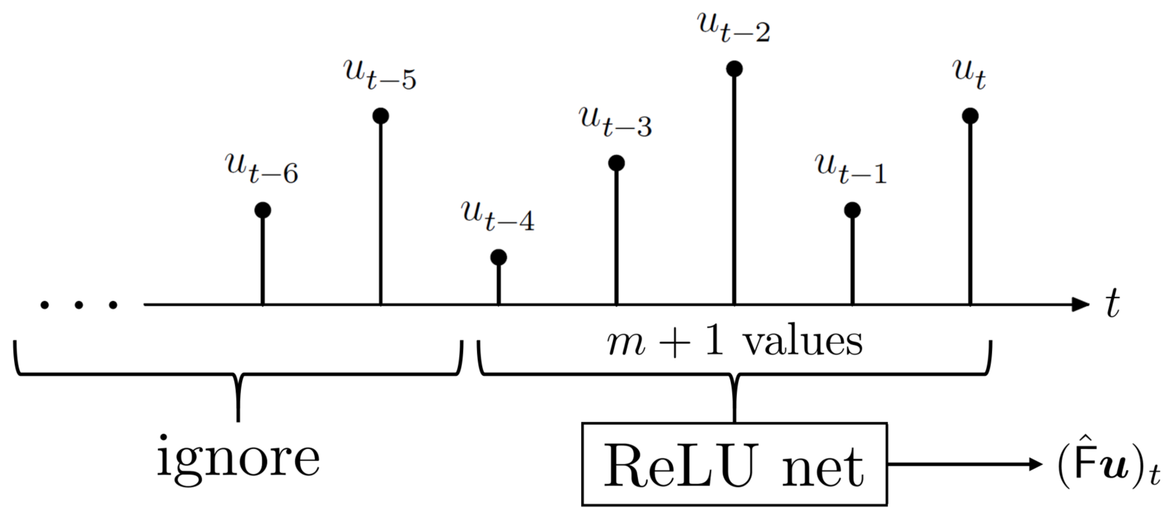


Figure 1 (left): Definition of approximately finite memory

Figure 2 (right): Sliding context window of ReLU TCN



Fading Memory:

- Set of weighting sequences — $\mathcal{W} := \{\boldsymbol{w} \in \mathcal{S} : w_t \in (0,1],\ w_t \downarrow 0 \text{ as } t \to \infty\}$
- $\boldsymbol{w}$-fading memory on $\mathcal{M} \subset \mathcal{S}$ — $\forall \epsilon > 0\ \exists \delta > 0\ \forall \boldsymbol{u},\boldsymbol{v} \in \mathcal{M}\ \forall t \in \mathbb{Z}_+$
$$\max_{s\in\{0,\dots,t\}} w_{t-s}|u_s - v_s| =: \|\boldsymbol{u}-\boldsymbol{v}\|_{\boldsymbol{w}} < \delta \implies |(\mathsf{F}\boldsymbol{u})_t - (\mathsf{F}\boldsymbol{v})_t| < \epsilon$$

Proposition: A continuous i/o map F satisfies Assumptions 1 and 2 if and only if it has $\boldsymbol{w}$-fading memory on $\mathcal{M}(R)$ for arbitrary $\boldsymbol{w} \in \mathcal{W}$.

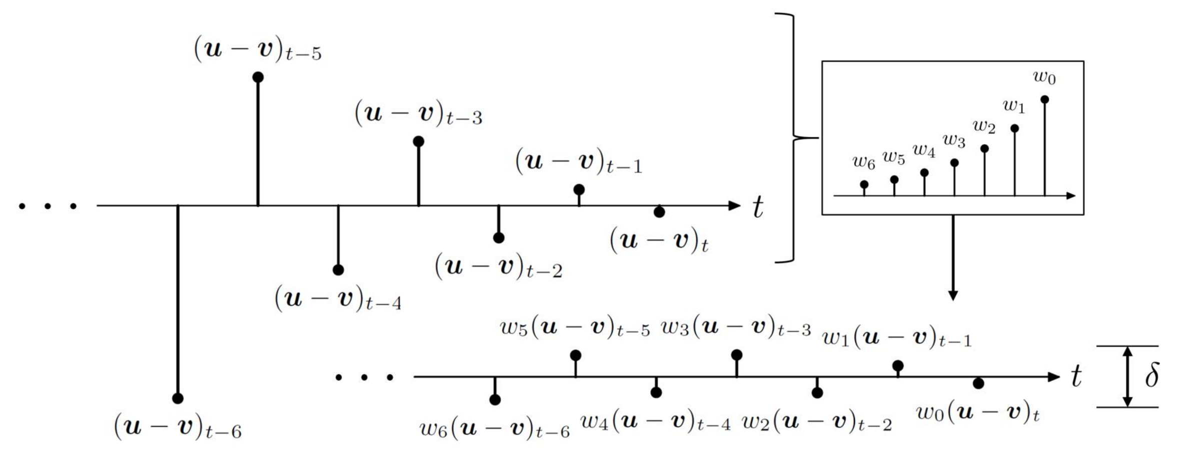


Figure 3 (left): Definition of fading memory

Recurrent Systems:

Many i/o maps $\mathsf{F} : \boldsymbol{u} \mapsto \boldsymbol{y}$ admit state space realizations

$$x_{t+1} = f(x_t, u_t)$$
$$y_t = g(x_t)$$

where $f : \mathbb{R}^n \times \mathbb{R} \to \mathbb{R}^n$, $g : \mathbb{R}^n \to \mathbb{R}$, $x_0 = \xi$. We consider systems that are

- *Uniformly asymptotically incrementally stable* on $\mathcal{M} \subset \mathcal{S}$ — there exists a function $\beta : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}_+$ of class $\mathcal{KL}$ such that $\forall \boldsymbol{u} \in \mathcal{M}$
$$\|\varphi^{\boldsymbol{u}}_{s,t}(\xi) - \varphi^{\boldsymbol{u}}_{s,t}(\xi')\| \le \beta(\|\xi - \xi'\|, t-s)$$
where $\varphi^{\boldsymbol{u}}_{s,t}(\xi)$ is the solution to the system above for $x_s = \xi$ and input $\boldsymbol{u}$.

Theorem: Assume f and g are Lipschitz, $\varphi^{\boldsymbol{u}}_{0,t}(\xi)$ remains in a compact set for all $\boldsymbol{u} \in \mathcal{M}$ and all $t \in \mathbb{Z}_+$, and β is summable over its second argument. Then the i/o map for the above system satisfies Assumptions 1 and 2.

Demidovich criterion: $\exists P \succ 0\ \exists \mu \in (0,1)\ \frac{\partial f}{\partial x}^{\mathsf{T}} P \frac{\partial f}{\partial x} - \mu P \preceq 0 \implies \beta(C,t) \propto C\mu^{t/2}$