

液体時定数ネットワーク

Ramin Hasani,^{1,3} * Mathias Lechner,² * Alexander Amini,¹ Daniela Rus,¹ Radu Grosu³

¹ マサチューセッツ工科大学 (MIT)

² オーストリア科学技術研究所 (ISTオーストリア)

³ ウィーン工科大学 (TU Wien)

rhasani@mit.edu, mathias.lechner@ist.ac.at, amini@mit.edu, rus@csail.mit.edu, radu.grosu@tuwien.ac.at

要旨

時間連続リカレントニューラルネットワークモデルの新しいクラスを紹介する。暗黙的非線形性によって学習システムの力学を宣言する代わりに、我々は非線形相互リンクゲートによって変調された線形一次力学系のネットワークを構築する。その結果得られるモデルは、隠れ状態に結合された変化する（すなわち液体）時定数を持つ力学系を表し、出力は数値微分方程式ソルバーによって計算される。これらのニューラルネットは、安定で境界のある挙動を示し、ニューラル常微分方程式ファミリーの中で優れた表現力をもたらし、時系列予測タスクの性能を向上させる。これらの特性を実証するために、まず理論的アプローチにより、そのダイナミクスの境界を求め、潜在的な軌跡空間における軌跡の長さの測定により、その表現力を計算する。次に、一連の時系列予測実験を行い、液体時定数ネットワーク (LTC) の近似能力を明らかにする。

古典的なRNNや最新のRNNと比較し

ニューラルネットワーク f によって隠れ状態の導関数を直接定義するよりも、より安定な連続時間リカレントニューラルネットワーク (CT-RNN) を以下の式によって決定することができる (Funahashi and Nakamura 1993) :

$\frac{dx(t)}{dt} = -\frac{x(t)}{\tau} + f(x(t), I(t), t, \theta)$ である。 $x(t)$ は隠れた状態、 $I(t)$ は入力、 t は時間を表し、 f は θ でパラメータ化される。

我々は別の定式化を提案する。ネットワークの隠れた状態の流れを、 $dx(t)/dt = x(t)/\tau + S(t)$ という形の線形ODE系で宣言し、 $S(t) \in \mathbb{R}^M$ 、 $S(t) = f(x(t), I(t), t, \theta)A$ 、パラメータ θ と A で決まる次の非線形性を表すとする:

$$\frac{dx(t)}{dt} = -\frac{x(t)}{\tau} + f(x(t), I(t), t, \theta)A \quad (1)$$

1 はじめに

常微分方程式 (ODE) によって決定される連続時間の隠れ状態を持つリカレントニューラルネットワークは、医療、産業、およびビジネスセクタでユビキタスに使用されている時系列データをモデル化するための効果的なアルゴリズムである。ニューラルODEの状態 $x(t) \in \mathbb{R}^D$ は、この方程式の解によって定義される (Chen et al. 2018): $dx(t)/dt = f(x(t), I(t), t, \theta)$ 。次に、数

値ODEソルバーを使用して状態を計算し、逆モード自動微分 (Rumelhart, Hinton, and Williams 1986) を実

行することによって、ソルバーを介した勾配降下 (Lechner et al. 2019)、またはソルバーをブラックボックスとみなして (Chen et al. 2018; Dupont, Doucet, and Teh 2019; Gholami, Keutzer, and Biroş 2019)、アドジョイント法 (Pontryagin 2018) を適用することによって、ネットワークを訓練することができる。未解決の問題は、現在の形式論においてニューラルODEはどの程度表現力があるのか、そして、より豊かな表現学習と表現力を可能にするためにその構造を改善できるのか、ということである。

*同等の貢献をした著者

著作権 © 2021, 人工知能学会 (www.aai.org). 無断複写・転載

を禁じます。

¹コードとデータは次のサイトで入手可能:

<https://github.com/raminmh/液体の時定数ネットワーク>

式1は、いくつかの特徴と利点を持つ新しい時間連続RNNインスタンスを示している:

液体時定数。ニューラルネットワーク f は、隠れた状態 $\mathbf{x}(t)$ の導関数を解除するだけでなく、入力に依存して変化する時定数($\tau_{\text{sys}} =$

$\frac{\tau}{1 + \tau f}$) 時定数は

(この特性)により、隠れ状態の単一要素は、各時点に到着する入力特徴に特化した力学系を識別することができる。我々はこれらのモデルを液体時定数リカレント・ニューラル・ネットワーク (LTC) と呼んでいる。LTCは任意のODEソルバーによって実装することができる。セクション2では、陰的オイラー法の安定性と陽的オイラー法の計算効率を同時に享受できる実用的な固定ステップODEソルバーを紹介する。

LTCの逆モード自動微分。LTCは微分可能な計算グラフを実現する。神経ODEと同様に、LTCは勾配ベースの最適化アルゴリズムによって学習させることができる。我々は、アドジョイントベースの最適化手法の代わりに、バニラバックプロパゲーションスルータイムアルゴリズムを用いてLTCを最適化することで、後方パスの間、メモリをv-merical精度と交換することにした (Pontryagin 2018)。セクション3では、この選択の動機付けを徹底的に行う。

束縛されたダイナミクス - 安定性 セクション4では、LTCの状態と時定数が有限範囲に束縛されることを示す。この性質は出力ダイナミクスの安定性を保証し、システムへの入力が増加する場合に望ましい。

優れた表現力。 セクション5では、LTCの近似能力を理論的かつ定量的に分析する。LTCの一意性を示すために関数解析のアプローチをとる。次に、他の時間連続モデルと比較して、LTCの表現力を測定する。これは、潜在的軌跡表現におけるネットワークの活性化の軌跡長を測定することによって行う。軌跡長はフィードフォワード深層ニューラルネットワークの表現力の尺度として導入された (Raghu et al.) 我々はこれらの基準を連続時間再電流モデルのファミリーに拡張する。

時系列モデリング セクション6では、11の時系列予測実験を行い、最新のRNNと時系列連続モデルの性能を比較する。LTCによって達成されたケースの大半において、性能が向上していることが確認された。

なぜこのような特殊な表現なのか？ このような特殊な表現を選択する理由は、主に2つある：

I) LTCモデルは、シナプス伝達機構と組み合わされた、小型種の神経ダイナミクスの計算モデルと緩やかに関連している (Hasani et al.) $dv/dt = g_l v(t) + S(t)$ 。ここでSはシナプス前ソースから細胞へのすべてのシナプス入力の合計であり、 g_l は漏れコンダクタンスである。

セルへのすべてのシナプス電流は、定常状態では以下の非線形性で近似できる (Koch and Segev 1998; Wicks, R  ehrig, and Rankin 1996)： $S(t)=f(v(t), I(t)), (A v(t))$ 、ここで $f(\cdot)$ はシグモイド非直線性で、現在のセルにシナプス前接続している全てのニューロン、 $v(t)$ 、及びセルへの外部入力の状態に依存する、

$I(t)$ 。これら2つの方程式をプラグインすることで、式1と同様の方程式が得られる。LTCはこの基礎に着想を得ている。

II) 式1は、有名な動的因果モデル (Dynamic Causal Models: DCM) (Friston, Harrison, and Penny 2003) のバイリニア力学系近似 (Penny, Ghahra- mani, and

Friston 2005) と似ているかもしれない。DCMは、力学系 $dx/dt = F(x(t), I(t), \vartheta)$ の2次近似(Bilinear)によって定式化され、以下のような形式となる (Friston, Harrison, and Penny 2003)：

アルゴリズム1 融合ODEソルバーによるLTC更新

パラメータ: $\vartheta = \{\tau^{(N \times 1)} = \text{時定数}, \gamma^{(M \times N)} = \text{重み}, \nu^{(N \times N)} = \text{リカレント重み}, \mu^{(N \times 1)} = \text{バイアス}, A^{(N \times 1)} = \text{バイアスベクトル}, L = \text{展開ステップ数}, \Delta t = \text{ステップサイズ}, N = \text{ニューロン数}, \text{入力長さ } T \text{ の } M \text{ 次元入力 } \mathbf{I}(t), \mathbf{x}(0) \text{ 出力: 次のLTCニューラル状態 } \mathbf{x}_{t+\Delta t}$

関数 FusedStep($\mathbf{x}(t), \mathbf{I}(t), \Delta t, \vartheta$)

$$\mathbf{x}(t + \Delta t)^{(N \times T)} = \frac{\mathbf{x}(t) + \Delta t f(\mathbf{x}(t), \mathbf{I}(t), t, \vartheta) \odot A}{1 + \Delta t \frac{1}{\tau} + f(\mathbf{x}(t), \mathbf{I}(t), t, \vartheta)}$$

$df(\cdot)$ で、すべての分割は要素ごとに適用される。 d はハダマード積である。

end 関数

$\mathbf{x}_{t+\Delta t} = \mathbf{x}(t)$

for $i = 1 \dots L$ **を行う**

$\mathbf{x}_{t+\Delta t} = \text{FusedStep}(\mathbf{x}(t), \mathbf{I}(t), \Delta t, \vartheta)$

end for

return

$\mathbf{x}_{t+\Delta t}$

ODEソルバは、トラジェクトリ $\mathbf{x}(0)$ から $\mathbf{x}(T)$ までのシミュレーションを行います。ODEソルバーは、連続シミュレーション区間 $[0, T]$ を時間的離散化 $[t_0, t_1, \dots, t_n]$ に分解する。その結果、ソルバーのステップに $d\mathbf{x}/dt = (A + \mathbf{I}(t)B)\mathbf{x}(t) + \mathbf{C}\mathbf{I}(t)$ ただし、 $A = \frac{dF}{d\mathbf{x}(t)}$ 、 $B = \frac{dF}{d\mathbf{I}(t)}$ 。

$\frac{dF}{d\mathbf{x}(t)} \frac{d\mathbf{I}(t)}{dt} = \frac{dF}{d\mathbf{I}(t)}$ DCMとバイリニア動的システム
LTCは連続時間(CT)モデルの変種として導入された。
LTCは連続時間(CT)モデルの変種として導入され、二論理学にゆるやかにインスパイアされ、時系列のモデル化において優れた表現力、安定性、性能を示す。

2 融合ODEソルバーによるLTCフォワードパス

式1を解析的に解くことは、LTCセマンティクスの非線形性により、非自明である。しかし、任意の時点 T におけるODE系の状態は、数値計算で求めることができる。

は、 t_i から t_{i+1} までのニューロン状態の更新のみが含まれる。

LTCのODEは剛方程式系を実現する(Press et al. 2007)。このタイプのODEは、ルンゲクッタ (RK) ベースの積分器でシミュレートすると、指数関数的な離散化ステップ数を必要とします。その結果、Dormand-Prince (torchd-iffeqのデフォルト (Chen et al. 2018)) のようなRKに基づくODEソルバーは、LTCには適していません。そこで、陽解法と陰解法のオイラー法(Press et al. 2007)を融合した新しいODEソルバーを設計します。この離散化手法の選択により、陰解更新方程式の安定性を達成する。この目的のために、Fused Solverは与えられた $dx/dt = f(x)$ の形の力学系を数値的に展開します:

$$\mathbf{x}(t_{i+1}) = \mathbf{x}(t_i) + \Delta t f(\mathbf{x}(t_i), \mathbf{x}(t_{i+1})). \quad (2)$$

特に、 f に線形に現れる $\mathbf{x}(t_i)$ のみを $\mathbf{x}(t_{i+1})$ で置き換える。その結果、式2は $\mathbf{x}(t_{i+1})$ について記号的に解くことができる。LTC表現にFusedソルバーを適用し、 $\mathbf{x}(t + \Delta t)$ について解くと、次のようになる:

$$\mathbf{x}(t + \Delta t) = \frac{\mathbf{x}(t) + \Delta t f(\mathbf{x}(t), \mathbf{I}(t), t, \vartheta) A}{1 + \Delta t \frac{1}{\tau} + f(\mathbf{x}(t), \mathbf{I}(t), t, \vartheta)}. \quad (3)$$

式3は、LTCネットワークの1つの更新状態を計算する。Cor

f は、任意の活性化関数を持つと仮定する (例えば、 \tanh 非線形性 $f = \tanh(\gamma_r \mathbf{x} + \gamma_l + \mu)$)。長さ T の入力シーケンスに対するアルゴリズムの計算量は $O(LT)$ であり、 L は離散化ステップ数である。直感的には、 N 個のニューロンを持つLTCネットワークの密なバージョンと、 N 個のセルを持つ長短期記憶 (LSTM) (Hochreiter and Schmidhuber 1997) ネットワークの密なバージョンは、同じ複雑さになる。

アルゴリズム2 BPTTによるLTCのトレーニング

入力: 長さ T のトレース $[(t, y(t))]$ のデータセット,
 $RNN-cell = f(l, x)$
 パラメータ損失関数 $L(\vartheta)$, 初期パラメータ ϑ_0 , 学習
 レート α , 出力 $w = W_{out}$, バイアス b_{out}
for $i = 1 \dots$ 訓練ステップ数 **do**
 $(l_b, y_b) =$ サンプル訓練バッチ, $\sim x := x_{t_0} \quad p(x)_{t_0}$
 for $j = 1 \dots T$ **を行う**
 $\hat{x} = f(l(t, x)), \hat{y}(t) = W_{out} \cdot \hat{x} + b_{out}, L_{total} =$
 $\sum_{j=1}^T L(y_j(t), \hat{y}_j(t)), \nabla L(\vartheta) =$
 $\frac{\partial L_{tot}}{\partial \vartheta}$
 $\vartheta = \vartheta - \alpha \nabla L(\vartheta)$
end for
end for
 ϑ を返す

表1: バニラBPTTの複雑数とアドジョイント法の比較、単層ニューラルネットワークの場合 f

	バニラBPTT	アドジョイント
時間	$O(L \times T \times 2)$	$O((l_f + l_b) \times t)$
メモリ	$O(L \times T)$	$O(1)$
深さ	$O(L)$	$O(L)_b$
FF acc	高い	高い
BWD acc	高い	低い

注: L = 離散化ステップ数、 L_f = フォワードパス中の L 、 L_b = 後方パス中の L 。

$T = \frac{1}{\text{step_size}} \times \text{depth} \times \text{width} \times \text{height}$

Training LTC networks by BPTT

ニューラルODEは、逆モードの自動微分を行うためにアドジョイント感度法を適用することで、ニューラルネットワーク f の各層に対して一定のメモリコストで学習することが提案された(Chen et al.)しかし、アドジョイント法はリバースモードで実行すると数値誤差を伴う。この現象は、アドジョイント法がコミュニティによって繰り返し示された順方向時間の計算軌跡を忘れてしまうために起こる(Gholami, Keutzer, and Biros 2019; Zhuang et al.)

逆に、時間軸を介した直接バックプロパゲーションは

(BPTT)は、逆モード積分中のフォワードパスの正確なリカバリーとメモリを交換する(Zhuang et al. 2020)。そこで我々は、ソルバーを通して高精度の逆方向パス積分を維持する、バニラBPTTアルゴリズムを設計することにした。この目的のために、与えられたODEソルバーの出力(ニューラル状態のベクトル)を再帰的に折りたたんでRNNを構築し、アルゴリズム

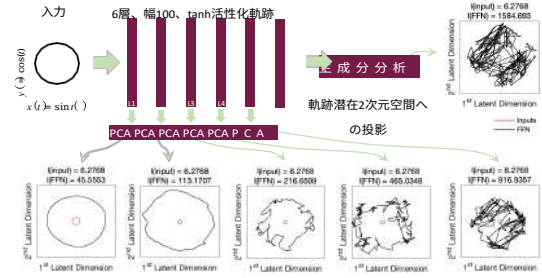


図1: 軌跡の潜在空間はより複雑になる

入力が隠れ層を通過するとき

LTCは無限の到着入力に対して安定を保つ(Hasani et al.) 本節では、定理1と定理2で述べたように、LTCニューロンの時定数と状態がそれぞれ有限の範囲に束縛されることを証明する。

定理1. x_i 、式1で識別されるLTCネットワーク内のニューロン i の状態を表し、ニューロンが M 個の接続を受信するとする。このとき、ニューロンの時定数 τ_{sys_i} は以下の範囲に束縛される:

$$\tau_i / (1 + \tau W_{ii}) \leq \tau_{sys_i} \leq \tau_i, \quad (4)$$

ム2で説明する学習アルゴリズムを適用してシステムを訓練することができる。アルゴリズム2は、バニラ確率的勾配降下(SGD)を使用する。これを、Adam(Kingma and Ba 2014)のような、SGDのより高性能な変種に置き換えることができる。

複雑さ。表1は、我々のバニラBPTTアルゴリズムの複雑さを、アドジョイント法と比較してまとめたものである。我々は、前方および後方積分軌道の両方において、大きなメモリコストで、同程度の計算複雑度で高い精度を達成しています。

4 τ の境界とLTCの神経状態

LTCは、入力に基づいて時定数を変化させるODEで表される。したがって

その証明は付録を参照されたい。これは、ニューラルネットワーク f の有界で単調増加するシグモイド非線形性と、LTCネットワークダイナミクスにおけるその置換に基づいて構成される。安定に変化する時間定数は、セクション5でより正式に発見するように、この形式の時間連続RNNの表現力を大幅に向上させる。

定理2. x_i 、式 (1) で識別されるLTC内のニューロンの状態を表し、ニューロン i が M 個の接続を受け取るとする。そして、任意のニューロンの隠れた状態は、有限区間 $t \in [0, T]$ 上で、次のように境界づけられる：

$$\min(0, A^{\min}_i) \leq x_i(t) \leq \max(0, A^{\max}_i) \quad (5)$$

その証明は付録の通りである。これは、LTCの方程式の区画の符号と、ODEモデルの陽解法的なオイラー離散化による近似に基づいて構成されている。定理2はLTCの望ましい性質、すなわち、入力が無限大に成長してもLTCの出力が爆発しないことを保証する状態安定性を示している。次に、CT-RNNやニューラル常微分方程式（Chen et al. 2018; Rubanova, Chen, and Duvenaud 2019）などの時間連続モデル群と比較したLTCの表現力について議論する。

5 LTCの表現力について

ニューラルネットの構造的特性が、どのような機能を計算できるかを決定する仕組みを理解することは、表現力問題として知られている。ニューラルネットの表現力を測定する初期の試みには、関数解析に基づく理論的研究がある。これらの研究は、3つの層を持つニューラルネットが、連続写像の任意の有限集合を任意の精度で近似できることを示している。これは普遍近似定理として知られている（Hornik, Stinchcombe, and White 1989; Funahashi 1989; Cybenko

表2: モデルの計算深度

活性	計算の深さ		
	化ニューラルODECT-RNN	LTC	
タン	0.56 ± 0.016	4.13 ± 2.19	9.19 ± 2.92
シグモイド	0.56 ± 0.00	5.33 ± 3.76	7.00 ± 5.36
ReLU	1.29 ± 0.10	4.31 ± 2.05	56.9 ± 9.03

注: 試行回数 10^4 、入力次元 $n=2$ 、 $\Delta t=0.01$ 、 100 秒の長さ。# 層数 = 1, 幅 = 100, $\sigma^2 = 2$, $\sigma^2 = 1$.

1989). 普遍性は標準RNN(Funahashi 1989)や連続時間RNN(Funahashi and Nakamura 1993)にも拡張された。注意深く考察することで、LTCも普遍的な近似器であることを示すことができる。

定理 3. $x \in \mathbb{R}^n$, $S \subset \mathbb{R}^n$, $x' = F(x)$ とする。

$n^1 D$ を S のコンパクトな部分集合とし、システムのシミュレーションが $t \in [0, T]$ 間において有界であると仮定する。このとき、正の ϵ について、 N 個の隠れユニット、 n 個の出力ユニット、 1 個の出力ユニットを持つLTCネットワークが存在する。

初期値 $x(0) \in S$ を持つシステムの任意のロールアウト $x(t) \in S$ と、適切なネットワークの初期化に対して、式1で記述される内部状態 $u(t)$ を置く、

$$\max_{t \in I} \|x(t) - u(t)\| < \epsilon \quad (6)$$

この証明の主な考え方は、 n 次元の力学系を定義し、それを高次元システムに配置することである。第二の系はLTCである。LTCの普遍性の証明とCT-RNN (Funahashi and Nakamura 1993) の証明の基本的な違いは、両システムのセマンティクスの違いにあり、LTCネットワークはその時間定数モジュールの中に非線形入力依存項を含み、これが証明の一部を非自明にしている。

普遍的近似定理は、ニューラルネットワークモデルの表現力を広く探るものである。しかし、この定理は、異なるニューラルネットアーキテクチャ間の隔たりがどこにあるのかについての基礎的な指標を与えてはくれない。したがって、モデル、特にLTCのような時空間データ処理に特化したネットワークを比較するためには、より厳密な表現力の尺度が必要である。静的ディープラーニングモデルの表現力に関する測定法の定義に関する進歩 (Pascanu, Montufar, and Bengio 2013; Montufar et al. 2014; Eldan and Shamir

そこで、2次元の潜在空間における出力軌跡の長さを測定し、その相対的な複雑さを明らかにする (図1参照)。軌跡の長さは、与えられた軌跡 $l(t)$ の弧の長さ (例えば2次元空間における円) として定義される (Raghu et al. 2017) : $l(l(t)) = \int_0^T \|dl(t)/dt\| dt$ を確立することにより軌跡の長さの成長に対する下界は、次のようになる。

は、ネットワークのウェイト構成に関する仮定に関係なく、浅いアーキテクチャーと深いアーキテクチャーのネットワークの間に障壁を設定する (Raghu et al. Pascanu, Montufar, and Bengio 2013; Montufar et al. 2014; Serra, Tjandraatmadja, and Ramalingam 2017; Gabriel et al. 2018; Hanin and Rolnick 2018, 2019; Lee, Alvarez-Melis, and Jaakkola 2019)。我々は、静的ネットワークの軌跡空間分析を以下のように拡張することに着手した。

時間連続(TC)モデル、およびモデルの表現力を比較するためのトラジェクトリ長の下限を設定する。この目的のために、Neural ODE、CT-RNN、およびLTCのインスタンスを共有 f で設計した。ネットワークは重み $\sim N(0, \sigma^2/k)$ 、バイアス $\sim N(0, \sigma^2)$ 。次に (2016; Poole et al. 2016; Raghu et al. 2017) は、理論的にも定量的にも、時間連続モデルの表現力を測定するのに役立つと考えられる。

5.1 軌跡の長さで表現力を測る

表現力の尺度は、ネットワークの容量 (深さ、幅、タイプ、重みの構成) が与えられたときに、学習システムがどのような複雑さを計算できるかを考慮しなければならない。静的ディープネットワークの統一的な表現力尺度は、(Raghu et al. 2017)で紹介された軌跡長である。この文脈では、ディープモデルが、与えられた入力軌跡 (例えば円形の2次元入力) を、より複雑なパターンに漸進的に変換する方法を評価する。

そして、得られたネットワークの活性化に対して、主成分分析 (PCA) を行うことができる。その後

異なるタイプのODEを用いたフォワードパスシミュレーションソルバーは、任意のウェイト・プロファイルに対して、 $t \in [0, 2\pi]$ に対して、 $I(t) = I_1(t) = \sin(t)$, $I_2(t) = \cos(t)$ の円入力軌道をネットワークに与えた。隠れ層の活性の最初の2つの主成分（平均分散説明率は80%以上）を見ることで、LTCについて一貫してより複雑な軌道が観察された。図2は、我々の経験的観察を垣間見ることができる。すべてのネットワークはDormand-Prince陽解法Runge-Kutta(4,5)ソルバー(Dormand and Prince 1980)で実装され、ステップサイズは可変である。我々は次のような**観察結果を得た**：**I)** Hard-tanhとReLU活性を持つNeural ODEとCT-RNNの軌跡長は指数関数的に成長し(図2A)、重みプロファイルに関わらず潜在空間の形状は変化しない。**II)** LTCは、Hard-tanhとReLUで設計した場合、軌跡長の成長速度が遅くなり、大きなレベルの複雑性を実現する妥協が見られる（図2A、2C、2E）。**III)** Hard-tanhとReLUによって構築された多層時間連続モデルは別として、全てのケースにおいて、LTCネットワークはより長く、より複雑な潜在空間挙動を観察した（図2B～図2E）。**IV)** 静的ディープネットワーク（図1）とは異なり、tanhとsigmoidで実現される多層連続時間ネットワークでは、深さによって軌跡長が成長しないことがわかった（図2D）。**V)** 結論として、我々は、TCモデルにおけるトラジェクトリ長は、モデルの活性度、重み分布、バイアス分布の分散、幅、深さによって変化することを観察した。図3では、これをより系統的に示した。**VI)** 軌跡の長さは、ネットワークの幅とともに直線的に成長する（図3B-対数スケールのY軸における曲線の対数的成長に注目）。**VII)** 分散が大きくなるにつれて、成長はかなり速くなる（図3C）。**VIII)** 軌跡の長さは、ODEソルバーの選択に依存しない（図3A）。**IX)** 活性化関数は、TCシステムによって探索される複雑なパターンを多様化し、ReLUとHard-tanhネットワークは、LTCに対してより高い複雑度を示す。その主な理由は、各層のセル間にリカレント・リンクが存在することである。**計算深度(L)の定義**。時間

連続ネットワークにおける f の隠れ層の場合、 L は各入力に対してソルバーがとる平均積分ステップ数である。 $\} \in$

サンプルとする。これらの観察から、連続時間ネットワークの軌跡の長さの成長に関する下界が形成された。

定理4.軌跡の長さ増加の境界。

ラルODEとCT-RNN。 $dx/dt = f_{n,k}(x(t), I(t), \vartheta)$ とする。

$\vartheta = \{W, b\}$ はニューラルODEを表し $\frac{dx(t)}{dt} = f_{n,k}(x(t), I(t), \vartheta)$ with $\vartheta = \{W, b, \tau\}$ a CT-RNN.

f はHard-tanh活性化でランダムに重み付けされる。 $I(t)$

を2次元入力軌跡とし、その漸進点(すなわち $I(t + \delta t)$)はすべての δt に対して $I(t)$ に垂直な成分を持ち、

L =ソルバーステップ数とする。次に

層 d の2次元潜在軌跡空間 $z^{(d)}(I(t)) = z^{(d)}(t)$ として、Neural ODEとCT-RNNのそれぞれについて、隠れ状態の最初の2つの主成分のスコアを互いに投影すると、次のようになる:

$$\|I(z^{(d)}(t))\| \geq \frac{\sigma_w^2 \sigma_b^2 k}{\sigma_w^2 + \sigma_b^2} \sqrt{\frac{d \times L}{\sigma_w^2 + \sigma_b^2}} \|I(t)\| \quad (7)$$

$$\|I(z^{(d)}(t))\| \geq \frac{(\sigma_w - \sigma_b)^2}{\sigma_w^2 + \sigma_b^2} \sqrt{\frac{d \times L}{\sigma_w^2 + \sigma_b^2}} \|I(t)\| \quad (8)$$

その証明は付録に記載されている。この証明は、連続時間セットアップのために注意深く考慮しながら、区分的線形活性化を持つディープネットワークについて確立された軌跡長境界に関する(Raghu et al. 2017)と同様のステップに従う。その証明は、原理成分領域における $d+1$ 層の隠れ状態勾配のノルム $\|dz/dt^{(d+1)}\|$ と、ニューラルODEとCT-RNNの微分方程式の右辺のノルムの期待値との間の再帰を定式化するように構築される。次に、再帰をロールバックして入力に到達する。

問題の複雑さを軽減するために、隠れた状態の直交成分のみを制限したことに注意。

image $\frac{dz}{dt}^{(d+1)}(t)$ 、従って、定理 (Raghu et al.

2017).次に、LTCネットワークの下限を見つける。

定理5.LTCの軌道長の成長率。式(1)で $\vartheta = W, b, \tau, A$ のLTCを決めるとする。 f と $I(t)$ に定理4と同じ条件をつけると、次のようになる:

$$\|I(z^{(d)}(t))\| \geq \frac{\sigma_w^2 \sigma_b^2 k}{\sigma_w^2 + \sigma_b^2} \sqrt{\frac{d \times L}{\sigma_w^2 + \sigma_b^2}} \|I(t)\| \quad (9)$$

その証明は付録を参照されたい。大まかな概要: 隠れた状態の勾配のノルムとLTCの右辺の成分との間に漸進的に境界を構築する再帰を別々に構築する。

5.2 理論的境界の議論

I) 予想通り、ニューラルODEの境界は、ソルバーステップ数 L に対する指数依存性を除いて、 n 層の静的ディープネットワークの境界と非常によく似ている。

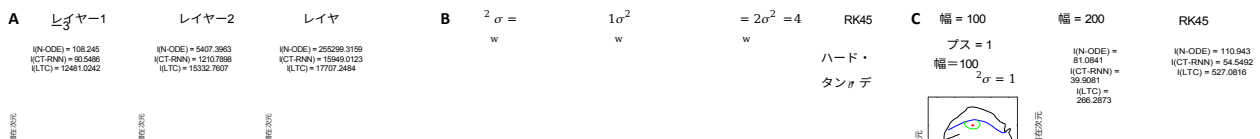
この結果は、図2と図3に示した実験結果と一致している。この結果は、図2と図3に示した我々の実験と一貫して一致する。III) Fig.

図2Bと図3Cは、LTCの軌跡の長さが、重量分布のばらつきの関数として、直線よりも速く成長していることを示している。これは、式(1)で示されるLTCの下限值によって確認される。

9.IV)LTCの下界はまた、幅 k と共に軌跡長が直線的に成長することを示し、これは3Bで示された結果を検証する。V)表2のHard-tanh活性化モデル L の計算深度を考慮すると、ニューラルODE、CT-RNN、LTCの下界は、セクション5の実験において、LTCネットワークのより長い軌跡長を正当化する。次に、現実の時系列予測タスクにおいてLTCの表現力を評価する。

6 実験的評価

6.1 時系列予測。提案されたFused ODEソルバーによって実現されたLTCの性能を、時系列予測に対して評価した。



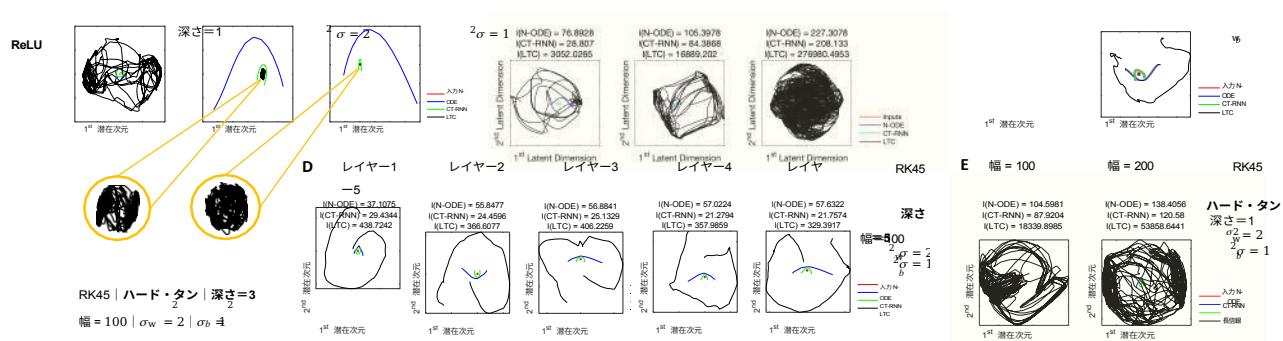


図2: Hard-tanh活性を持つネットワーク層における軌跡長の変形A)、B)重み分散スケーリング係数の関数として、C)ネットワーク幅(ReLU)の関数として、D)ロジスティック・シグモイド活性を持つネットワーク層における、E)幅(Hard-tanh)の関数として。

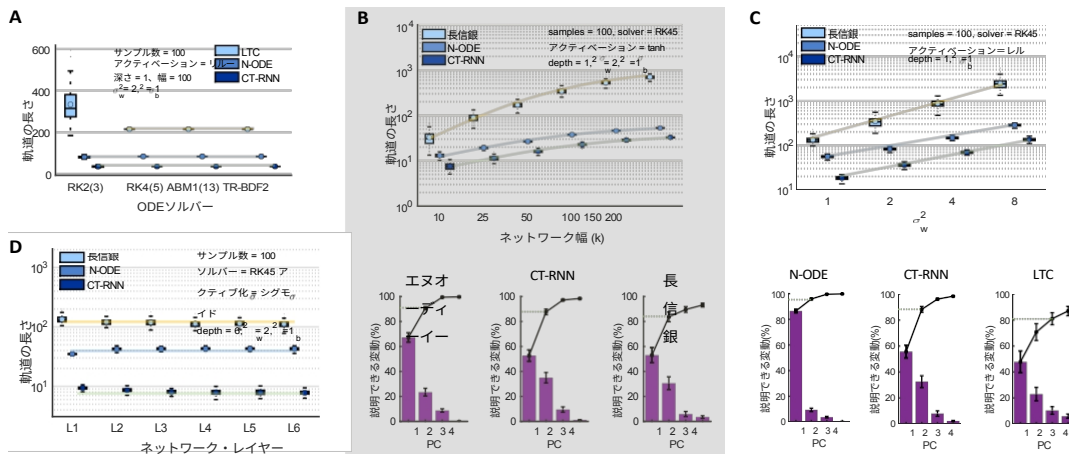


図3: 軌跡長依存性。A) 軌跡長と異なるソルバー(可変ステップソルバー)との比較。RK2(3): Bogacki-Shampine Runge-Kutta (2,3) (Bogacki and Shampine 1989).RK4(5): Dormand-Prince explicit RK (4,5) (Dormand and Prince 1980).ABM1(13): Adams-Bashforth-Moulton (Shampine 1975).TR-BDF2: 第1段台形則と第2段後方微分による陰解法 RKソルバー (Hosea and Shampine 1996).B) 上: 軌跡の長さ。下: 主成分の分散 (紫の棒) とその累積値 (黒の実線)。

C) 軌跡の長さ重量分布の分散。D) 軌跡の長さ層数。(表3: 時系列予測 平均と標準偏差, $n=5$)

データ集合	メトリック	エルエスティーエム	CT-RNN	ニューラルODE	CT-GRU	LTC
ジェスチャー	精度	64.57% ± 0.59	59.01% ± 1.22	46.97% ± 3.03	68.31% ± 1.78	69.55% ± 1.13
占有率	精度	93.18% ± 1.66	94.54% ± 0.54	90.15% ± 1.71	91.44% ± 1.67	94.63% ± 0.17
アクティビティ認識	精度	95.85% ± 0.29	95.73% ± 0.47	97.26% ± 0.10	96.16% ± 0.39	95.67% ± 0.575
シーケンシャル MNIST	精度	98.41% ± 0.12	96.73% ± 0.19	97.61% ± 0.14	98.27% ± 0.14	97.57% ± 0.18
交通	(自乗誤差)	0.169 ± 0.004	0.224 ± 0.008	1.512 ± 0.179	0.389 ± 0.076	0.099 ± 0.0095
パワー	(二乗エラー)	0.628 ± 0.003	0.742 ± 0.005	1.254 ± 0.149	0.586 ± 0.003	0.642 ± 0.021
オゾン	(F1スコア)	0.284 ± 0.025	0.236 ± 0.011	0.168 ± 0.006	0.260 ± 0.024	0.302 ± 0.0155

表4: 人物の活動、1番目の設定 - $n=5$

アルゴリズム	精度
LSTM83	.59% ± 0.40
CT-RNN81	.54% ± 0.33
潜在ODE76	.48% ± 0.56
CT-GRU85	.27% ± 0.39
LTC (ours)	85.48% ± 0.40

で提案されているように、最先端の離散化RNN、LSTM (Hochreiter and Schmidhuber 1997)、CT-RNN (ODE-RNN) (Funahashi and Nakamura 1993; Rubanova, Chen, and Duvenaud 2019)、連続時間ゲートリカレントユニット (CT-GRU) (Mozier, Kazakov, and Lindsey 2017)、および4th次のルンゲクッタソルバによって構築されたニューラルODE (Chen et al.2018)で提案されているような4次ルンゲクッタソルバーによって構築されたニューラル

ODEを、一連の多様な実生活の超視覚学習タスクで使用した。結果は表

3.実験セットアップは付録に記載されている。その結果、7つの実験のうち4つでLTCが他のRNNモデルと比較して5%から70%の性能向上を達成し、残りの3つでは同等の性能を達成したことが確認された (表3参照)。

6.2 人活動データセット。我々は、(Rubanova, Chen, and Duvenaud)に記述されている "Human Activity" データセットを使用する。

2019) を2つの異なるフレームワークで評価した。データセットは6554個のヒトの活動シーケンス（例：横たわる、歩く、座る）で構成され、周期は211msである。1つ目の設定では、ベースラインは前述のモデルであり、入力表現は変更しない（詳細は付録参照）。表4に示すように、LTCはすべてのモデル、特にCT-RNNとニューラルODEを大きなマージンで上回る。CT-RNNアーキテクチャは、(Rubanova, Chen, and Duvenaud 2019)に記載されているODE-RNNと同等であり、状態減衰係数 τ を持つという違いがあることに注意。

2番目の設定では、LTCと(Rubanova, Chen, and Duvenaud 2019)で議論されたより多様なRNNバリエーションのセットとの公正な比較を得るために、(Rubanova, Chen, and Duvenaud 2019)による修正に合わせて実験を慎重に設定した（補足参照）。LTCは他のモデルと比較して高いマージンで優れた性能を示す。結果は表5) にまとめられている。

6.3 ハーフチーターの運動モデリング。我々は、連続時間モデルが物理的ダイナミクスをどの程度捉えることができるかを評価することを意図した。これを実行するために、我々は、HalfCheetah-v2ジム環境 (Brockman et al.

表5: 人物のアクティビティ、第2セッティング

アルゴリズム	精度
rnn Δ_t^*	0.797 \pm
RNN-Decay [*]	0.800 \pm
0.003	
0.010	
rnn-gru-d0	.806 \pm 0.007
RNN-VAE [*]	0.343 \pm
0.040	
潜在ODE (Dエンコード)	0.835
	\pm 0.010
ODE-RNN [*]	0.829 \pm
潜在ODE(Cエンコード) [*]	0.016
	0.846 \pm
	0.013
LTC (当社比)	0.882 \pm 0.005

注: *で示したアルゴリズムの精度は、(Rubanova, Chen, and Duvenaud 2019) から直接引用した。RNN Δt = classic RNN + 入力遅延(Rubanova, Chen, and Duvenaud 2019)。RNN-Decay = 隠れ状態に指数関数的減衰を持つRNN (Mozier, Kazakov, and Lindsey 2017)。GRU-D = gated recurrent unit + exponential decay + input imputation (Che et al. 2018)。D-enc. = RNNエンコーダ(Rubanova, Chen, and Duvenaud 2019)。C-enc = ODEエンコーダ (Rubanova, Chen, およびDuvenaud 2019)。

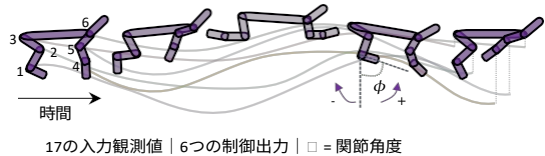


図4: ハーフチーターの物理シミュレーション

JoCo物理エンジン (Todorov, Erez, and Tassa 2012)。タスクは、観測空間の時系列をau-to-regressiveにフィットさせることである (図4)。難易度を上げるために、行動の5%をランダムな行動で上書きする。テストの結果は表6に示されており、LTCの性能が他のモデルに比べて優れていることを裏付けている。

7 関連作品

時間連続モデル。TCネットワークは前例がないほど普及している。これは、適応計算、より優れた連続時系列モデリング、メモリ、パラメータ効率など、

表6: シーケンスモデリング。ハーフチーターダイナミクス $n=5$

アルゴリズム	MSE
LSTm2	.500 \pm 0.140
CT-RNN2	.838 \pm 0.112
ニューラルODE3	.805 \pm 0.313
ct-gru3	.014 \pm 0.134
LTC (ours)	2.308\pm 0.015

案した (Montufar et al. 2013) は、リニア・リレーの数を数えることを提

いくつかの利点の発現によるものである (Chen et al.) 多くの代替的なアプローチが、アドジョイント法の改善と安定化 (Gholami, Keutzer, and Birois 2019)、特定の文脈でのニューラルODEの使用 (Rubanova, Chen, and Duvenaud 2019; Lechner et al. 本研究では、ニューラルODEの表現力を調査し、その表現力と性能を向上させる新しいODEモデルを提案した。

表現力の尺度。より深いネットワークや特定のアーキテクチャが優れた性能を発揮するのはなぜか、浅いネットワークと深いネットワークの近似能力の境界はどこにあるのか、といった疑問に対する答えを見出そうと、現代の多くの研究が試みられている。このコンテキストでは、(Montufar et al.

の表現力の尺度として、ニューラルネットワークの指数関数が存在することを示し、(El-dan and Shamir 2016)は、より小さなネットワークが生成できない放射状関数のクラスが存在することを示し、(Poole et al. 2016)は、過渡的カオスによるニューラルネットワークの指数関数的表現力を研究した。

これらの方法は説得力があるが、(Serra, Tjandraatmadja, and Ramalingam 2017; Gabrie' et al. 2018; Hanin and Rolnick 2018, 2019; Lee, Alvarez-Melis, and Jaakkola 2019)に類似した表現力を下限するために、与えられたネットワークの特定の重み構成に縛られている。(Raghu et al. 2017) は、軌跡の長さによって与えられた静的ネットワークの表現力を定量化する、相互に関連する概念を導入した。我々は彼らの表現力分析を時間連続ネットワークに拡張し、軌跡長の成長に対する下界を提供し、LTCの優れた近似能力を宣言した。

8 結論、範囲、限界

我々は、線形ODEニューロンと特殊な非線形ウェイト構成の組合せによって得られる、新しいクラスの時間連続ニューラルネットワークモデルの使用について研究した。その結果、任意の可変ステップおよび固定ステップのODEソルバーによって効果的に実装できること、および時間を通してバックプロパゲーションによって学習できることを示した。また、教師あり学習の時系列予測タスクにおいて、標準的なディープラーニングモデルや最新のディープラーニングモデルと比較し、その拘束された安定したダイナミクス、優れた表現力、優れた性能を実証した。

長期依存性。時間連続モデルの多くのバリエーションと同様に、LTCは勾配降下法で学習すると、消失勾配現象 (Pascanu, Mikolov, and Bengio 2013; Lechner and Hasani 2020) を表現する。このモデルは様々な時系列予測タスクで有望であるが、現在の形式では長期的な依存関係を学習するための明らかな選択肢ではない。

ODEソルバーの選択。時間連続モデルの性能は、その数値的実装方法に大きく左右される(Hasani 2020)。LTCは高度な可変ステップソルバーや今回紹介するFused固定ステップソルバーで良好な性能を発揮しますが、その性能は市販の陽解法オイラー法を使用した場合に大きく影響されます。

時間と記憶。ニューラルODEは、LTCのようなより洗練されたモデルに比べて、驚くほど高速である。しかし、表現力には欠ける。我々の提案するモデルは、現在の形式では、TCモデルの表現力を大幅に向上させるが、その代償として時間とメモリの複雑さが増大する。

因果関係。時間連続差分方程式セマンティクスによって記述されるモデルは、本質的に因果構造を持っている (Schoenkopf 2019)。LTCのようなリカレントモデルの意味論は、*動的因果モデル* (Friston, Harrison, and Penny 2003) のバイリニア力学系近似 (Penny, Ghahramani, and Friston 2005) に似ているため、その因果関係を研究することは、将来的な再探索の方向性として興味深い。従って、LTCのような因果構造が推論を向上させるのに役立つ、連続時間観測・行動空間におけるロボットの制御が自然な応用領域となる (Lechner et al.)

謝辞

R.H.とD.R.はボーイング社から一部援助を受けている。R.H.とR.G.は、Horizon-2020 ECSEL プロジェクト助成金 No.783163 (iDev40) の一部支援を受けた。M.L.は、オーストリア科学基金 (FWF) の助成金 Z211-N23 (ウィトゲンシュタイン賞) の一部を受けた。A.A.は National Science Foundation (NSF) Graduate Research Fellowship Program の支援を受けている。本研究は、R.H.の博士論文から一部抜粋したものである。

参考文献

Anguita, D.; Ghio, A.; Oneto, L.; Parra, X.; and Reyes-Ortiz, J.L. 2013. スマートフォンを使った人間の行動認識のためのパブリックドメインデータセット。In *Esann*.

Bogacki, P.; and Shampine, L. F. 1989. 3(2)組のルンゲクッタ公式. *応用数学レターズ* 2(4): 321-325.

Brockman, G.; Cheung, V.; Pettersson, L.; Schneider, J.; Schulman, J.; Tang, J.; and Zaremba, W. 2016. Openai gym. *arXiv preprint arXiv:1606.01540*.

Candanedo, L. M.; and Feldheim, V. 2016. 統計的学習モデルを用いた、光、温度、湿度、CO₂測定値からのオフィスルームの正確なオキュパンシー検知。 *Energy and Buildings* 112: 28-39.

Che, Z.; Purushotham, S.; Cho, K.; Sontag, D.; and Liu, Y. 2018. Recurrent neural networks for multivariate time series with missing values. *Scientific reports* 8(1): 1-12.

Chen, T. Q.; Rubanova, Y.; Bettencourt, J.; and Duvenaud, D.K. 2018. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, 6571-6583.

Cybenko, G. 1989. シグモイド関数の重ね合わせによる近似. *Mathematics of control, signals and systems* 2(4): 303-314.

Dormand, J. R.; and Prince, P. J. 1980. 埋め込みルンゲクッタ公式の一群. *Journal of computational and applied mathematics* 6(1): 19-26.

Dua, D.; and Graff, C. 2017. UCI Machine Learning Repository. URL <http://archive.ics.uci.edu/ml>.

Dupont, E.; Doucet, A.; and Teh, Y. W. 2019. Augmented neural odes. In *Advances in Neural Information Processing Systems*, 3134-3144.

Durkan, C.; Bekasov, A.; Murray, I.; and Papamakarios, G. 2019. Neural spline flows. In *Advances in Neural Information Processing Systems*, 7509-7520.

Eldan, R.; and Shamir, O. 2016. The power of depth for feedforward neural networks. In *Conference on learning theory*, 907-940.

Friston, K. J.; Harrison, L.; and Penny, W. 2003. 動的因果モデリング。 *Neuroimage* 19(4): 1273-1302.

船橋啓一 1989. ニューラルネットワークによる連続写像の近似的実現について. *ニューラルネットワーク* 2(3): 183-192.

船橋慶一郎; 中村由行, 1993. 連続時間リカレントニューラルネットワークによる力学系の近似. *ニューラル・ネットワーク* 6(6): 801-806.

Gabrie, M.; Manoel, A.; Luneau, C.; Macris, N.; Krzakala, F.; Zdeborova, L.; et al. Deep Neural Network のモデルにおけるエントロピーと相互情報量。 In *Advances in Neural Information Processing Systems*, 1821-1831.

Gholami, A.; Keutzer, K.; and Biro, G. 2019. Anode : *arXiv preprint arXiv:1902.10298*.

Hanin, B.; and Rolnick, D. 2018. トレーニングの始め方 : 初期化とアーキテクチャの効果。 In *Advances in Neural Information Processing Systems*, 571-581.

Hanin, B.; and Rolnick, D. 2019. *ArXiv preprint arXiv:1901.09021*.

韓秀, Y.; Jiawei, D.; Vincent, T.; and Jiashi, F. 2020. ニューラル常微分方程式の頑健性について. In *International Conference on Learning Representations*.

Hasani, R. 2020. 連続時間制御環境における解釈可能なリカレントニューラルネットワーク. 博士論文、ウィーン工科大学。

Hasani, R.; Amini, A.; Lechner, M.; Naser, F.; Grosu, R.; and Rus, D. 2019. Response characterization for auditing cell dynamics in long short-term memory networks. In *2019 International Joint Conference on Neural Networks (IJCNN)*, 1-8. IEEE.

Hasani, R.; Lechner, M.; Amini, A.; Rus, D.; and Grosu, R. 2020. 自然な宝くじ当選者: 普通の神経回路による強化学習。 *2020 年機械学習国際会議講演論文集 JMLR*.

Hirsch, M. W.; and Smale, S. 1973. 微分方程式、力学系と線形代数. Academic Press college division.

Hochreiter, S.; and Schmidhuber, J. 1997. 長期短期記憶。 *神経計算* 9(8): 1735-1780.

Holl, P.; Koltun, V.; and Thuerey, N. 2020. 微分可能な物理でPDEを制御する学習。

Hornik, K.; Stinchcombe, M.; and White, H. 1989. 多層フィードフォワードネットワークは普遍的な近似器である。 *Neural networks* 2(5): 359-366.

- Hosea, M.; and Shampine, L. 1996. TR-BDF2の解析と実装. *応用数値数値* 20(1-2): 21-37.
- Jia, J.; and Benson, A. R. 2019. Neural jump stochastic differential equations. In *Advances in Neural Information Processing Systems*, 9843-9854.
- Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- Koch, C.; and Segev, K. 1998. *Methods in Neuronal Modeling - From Ions to Networks*. MIT press, second edition.
- Lapicque, L. 1907. 分極を利用した神経興奮の定量的研究。 *Journal de Physiologie et de Pathologie Generale* 9: 620-635.
- Lechner, M.; and Hasani, R. 2020. 不規則にサンプリングされた時系列における長期依存性の学習. *arXiv preprint arXiv:2006.04418*.
- レヒナー、M.; ハサニ、R.; アミニ、A.; ヘンツィンガー、T. A.; ルス、D.; およびグロス、R. 2020a. 自律可能な神経回路政策。 *Nature Machine Intelligence* 2(10): 642-652.
- レヒナー、M.; ハサニ、R.; ルス、D.; そしてグロス、R. 2020b. Gershgorin Loss Stabilizes the Recurrent Neural Network Compartment of an End-to-end Robot Learning Scheme. *2020 International Conference on Robotics and Automation (ICRA)*. IEEE.
- Lechner, M.; Hasani, R.; Zimmer, M.; Henzinger, T. A.; and Grosu, R. 2019. Designing worm-inspired neural networks for interpretable robotic control. In *2019 International Conference on Robotics and Automation (ICRA)*, 87-94. IEEE.
- Lee, G.-H.; Alvarez-Melis, D.; and Jaakkola, T. S. 2019. Towards robust, locally linear deep networks. *arXiv preprint arXiv:1907.03207*.
- Montufar, G. F.; Pascanu, R.; Cho, K.; and Bengio, Y. 2014. On the number of linear regions of deep neural networks. In *Advances in neural information processing systems*, 2924-2932.
- Mozer, M. C.; Kazakov, D.; and Lindsey, R. V. 2017. Discrete Event, Continuous Time RNNs. *arXiv preprint arXiv:1710.04110*.
- Pascanu, R.; Mikolov, T.; and Bengio, Y. 2013. リカレントニューラルネットワークの学習の困難さについて. In *International conference on machine learning*, 1310-1318.
- Pascanu, R.; Montufar, G.; and Bengio, Y. 2013. On the number of response regions of deep feed forward networks with piece-wise linear activations. *arXiv preprint arXiv:1312.6098*.
- Penny, W.; Ghahramani, Z.; and Friston, K. 2005. バイリニア力学系。 *Philosophical Transactions of the Royal Society B: Biological Sciences* 360(1457): 983-993.
- Pontryagin, L. S. 2018. *Mathematical theory of optimal processes*. Routledge.
- Poole, B.; Lahiri, S.; Raghu, M.; Sohl-Dickstein, J.; and Ganguli, S. 2016. ディープニューラルにおける指数表現力

過渡的なカオスを介してネットワーク。 *神経形成処理システムの進歩*, 3360-3368.

Press, W. H.; Teukolsky, S. A.; Vetterling, W. T.; and Flannery, B. P. 2007. *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. New York, NY, USA: Cambridge University Press, 3 edition.

Quaglino, A.; Gallieri, M.; Masci, J.; and Koutník, J. 2020. SNODE: Spectral Discretization of Neural ODEs for System Identification. In *International Conference on Learning Representations*.

Raghu, M.; Poole, B.; Kleinberg, J.; Ganguli, S.; and Dickstein, J. S. 2017. On the expressive power of deep neural networks. *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, 2847-2854. JMLR.

Rubanova, Y.; Chen, R. T.; and Duvenaud, D. 2019. 不規則にサンプリングされた時系列に対する Latent ODEs.

Rumelhart, D. E.; Hinton, G. E.; and Williams, R. J. 1986. 誤差の逆伝播による表象の学習: 533-536.

Schäfer, A. M.; and Zimmermann, H. G. 2006. リカレントニューラルネットワークは普遍的な近似器である。 In *International Conference on Artificial Neural Networks*, 632-640. Springer.

Schölkopf, B. 2019. Causality for Machine Learning. *arXiv preprint arXiv:1911.10500*.

Serra, T.; Tjandraatmadja, C.; and Ramalingam, S. 2017. 深層ニューラルネットの線形領域のバウンディングとカウント. *arXiv preprint arXiv:1711.02114*.

Shampine, L. F. 1975. 常微分方程式の計算機解法. *初期値問題*.

Todorov, E.; Erez, T.; and Tassa, Y. 2012. Mujoco: モデルベース制御のための物理エンジン. *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, 5026-5033. IEEE.

Tsagris, M.; Beneki, C.; and Hassani, H. 2014. 折り畳まれた正規分布について. *数学* 2(1): 12-28.

Wagner, P. K.; Peres, S. M.; Madeo, R. C. B.; de Moraes Lima, C. A.; and de Almeida Freitas, F. 2014. Gesture unit segmentation using spatial-temporal information and machine learning. In *The Twenty-Seventh International Flairs Conference*.

Wicks, S. R.; Roehrig, C. J.; and Rankin, C. H. 1996. 線虫タップ引き出し回路の動的ネットワークシミュレーション: 行動学的基準を用いたシナプス機能に関する予測. *Journal of Neuroscience* 16(12): 4017-4031.

Zhang, K.; and Fan, W. 2008. 歪んだ2種の確率的オゾン日の予測: 分析、解決策、そしてその先へ。

Knowledge and Information Systems 14(3): 299-326.

Zhuang, J.; Dvornek, N.; Li, X.; Tatikonda, S.; Papademetris, X.; and Duncan, J. 2020. ニューラルODEにおける勾配推定のための適応的チェックポイント・アジョイント法. *第37回Machine Learning国際会議講演論文集*. PMLR 119.

補足資料

S 1 定理1の証明

証明。ニューラルネットワークが、0と1の間で単調増加する有界シグモイド非線形性を持つと仮定する：

$$0 < f(x_j(t), v_{ij}, \mu_{ij}) < 1 \quad (S1)$$

式1の f の上界を置き換え、 f の各ニューロン i について、スケーリング重み行列 $W^{M \times 1}$ を仮定すると、次のようになる

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i + W_i x(t) + W_i A_i \quad (S2)$$

この方程式は、線形ODEに単純化される：

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i + W_i x(t) + W_i A_i \rightarrow \frac{dx_i}{dt} = -ax_i + b_i \quad (S3)$$

という形の解を持つ：

$$x_i(t) = k_1 e^{-at} + \frac{b}{a} \quad (S4)$$

この解から、システムの時定数の下界 τ^{min} を導出する：

$$\tau = \min_i \frac{1}{a} = \frac{1}{1 + \tau W_{ii}} \quad (S5)$$

式1の f の下界を置き換えることで、式は以下のように自律線形ODEに単純化される：

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i(t) \quad (S6)$$

これはシステムの時定数の上限 τ^{max} を与える：

$$\tau^{max} = \tau_i \quad (S7)$$

□

S2 定理2の証明

証明。ニューロン i の神経状態 $x_i(t)$ として $M = \max\{0, A^{max}\}$ を式1に挿入する：

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i + f(x_j(t), t, \vartheta) M + f(x_j(t), t, \vartheta) A_i \quad (S8)$$

括弧を展開すると、次のようになる。

$$\frac{dx_i}{dt} = -\frac{1}{\tau_i} x_i + \frac{f(x_j(t), t, \vartheta) M + f(x_j(t), t, \vartheta) A_i}{x_i} \quad (S9)$$

式S9の右辺は、 M の条件、正の重み、 $f(x_j)$ も正であることから負である。したがって、左辺も負でなければならず、微分項について近似を行うと、以下のようになる：

$$\frac{dx_i}{dt} \leq 0, \quad \frac{dx_i}{dt} \approx \frac{x_i(t + \Delta t) - x_i(t)}{\Delta t} \leq 0, \quad (S10)$$

$x_i(t)$ を M に代入すると、次のようになる：

従って

$$\frac{x(t+\Delta t)-M}{\Delta t} \leq 0 \rightarrow x(t+\Delta t) \leq M \tag{S11}$$

$$x_i(t) \leq \max(0, A^{max}). \tag{S12}$$

ここで、 $x_{(i)}$ を $m = \min\{0, A^{min}\}$ で置き換え、上界で用いたのと同様の手法に従えば、次のようになる：

$$\frac{x(t + \Delta t) - m}{\Delta t} \leq 0 \rightarrow x(t + \Delta t) \leq m, \quad (S13)$$

従って

$$x_i(t) \geq \min(0, A^{min}). \quad (S14)$$

□

S3定理3の証明

我々は、有限のシミュレーション時間において、与えられた n 次元の力学系は、 n 個の出力、いくつかの隠れノード、適切な初期条件を持つLTCの内部と出力の状態によって近似できることを証明する。我々の証明は、フィードフォワードニューラルネットワーク(Funahashi 1989; Cybenko 1989; Hornik, Stinchcombe, and White 1989)、リカレントニューラルネットワーク(RNN)(Funahashi 1989; Schaffer and Zimmermann 2006)、連続時間RNN(Funahashi and Nakamura 1993)の基本的な普遍近似定理(Hornik, Stinchcombe, and White 1989)に基づいている。LTCとCT-RNNの普遍的な近似能力の証明の根本的な違いは、両者のODEシステムのセマンティクスの違いにある。LTCネットワークは、式1で表される時定数モジュールに非線形入力依存項を含んでおり、これが全体の力学系をCT-RNNのそれとは変えてしまう。従って、その普遍性を証明するためには、CT-RNNと同じアプローチを取りながら、注意深い考察を調整しなければならない。まず、証明に使われる、力学系の基本的なトピックに関する予備的な記述を再確認する。

定理（基本近似定理）（船橋 1989）。 $x = (x_1, \dots, x_n)$ を n 次元ユークリッド空間 R^n とする。 $f(x)$ をシグモイド関数（ R における非定数、単調増加、有界の連続関数）とする。 K を R^n のコンパクト部分集合とし、 $f(x_1, \dots, x_n)$ を K 上の連続関数とする。任意の $\epsilon > 0$ に対して、整数 N 、実定数 $c_i, \vartheta_i (i = 1, \dots, N)$ および $w_{ij} (i = 1, \dots, N; j = 1, \dots, n)$ が存在し、次のようになる。

$$\max_{x \in K} |g(x_1, \dots, x_n) - \sum_{i=1}^N c_i f(\sum_{j=1}^n w_{ij} x_j - \vartheta_i)| < \epsilon \quad (S15)$$

を保
持し
てい
る。

この定理は、3層のフィードフォワード・ニューラル・ネットワーク（入力-非表示層-出力）が、任意の連続写像 g をコンパクトな集合上で近似できることを示している： $R^n \rightarrow R^m$ をコンパクトな集合上で近似できることを示す。

定理（連続時間リカレントニューラルネットワークによる力学系の近似）(Funahashi and Nakamura 1993). $D \subset R^n$, $F : D \rightarrow R^n$ を自律常微分方程式、 C^1 -写像とし、 $\dot{x} = F(x)$ を D 上の力学系とする。 K を D のコンパクト部分集合とし、区間 $[0, T]$ 上の力学系の軌道を考える。そして、任意の正の ϵ については、整数 N と、 N 個の隠れユニット、 n 個の出力ユニット、および出力内部状態 $u(t) = (u_1(t), \dots, u_n(t))$ を持つリカレント・ニューラル・ネットワークが存在する：

$$\frac{du_i(t)}{dt} = -\frac{u_i(t)}{\tau_i} + \sum_{j=1}^n w_{ij} f(\mu_j(t)) + I_i(t), \quad (S16)$$

ここで、 τ_i は時定数、 w_{ij} は重み、 $I_i(t)$ は入力、 f は C^1 -シグモイド関数 ($f(x) = 1/(1 + \exp(-x))$) であり、初期値 $x(0) \in K$ を持つシステムの任意の軌跡 $x(t)$ 、 $t \in [0, T]$ 、およびネットワークの適切な初期条件に対して、以下の文が成立する：

$$\{\epsilon\} \in$$

$$\max_t |x(t) - u(t)| < \epsilon.$$

この定理は、時定数 τ が全ての隠れ状態に対して一定に保たれ、RNNに入力がない場合 ($I_i(t) = 0$) について証明された (Funahashi and Nakamura 1993)。

ここで、証明に必要な力学系の概念を再掲する。必要な場合には、定理1を証明するためにレンマの修正と拡張を行う。

リプシッツ。写像 $F: S \rightarrow \mathbb{R}^n$ (S は \mathbb{R}^n の開部分集合である) が存在するとき、 S 上でリプシッツと呼ばれる。
(リプシッツ定数) である:

$$|F(x) - F(y)| \leq L|x - y|, \quad \forall x, y \in S. \quad (\text{S17})$$

局所的にリプシッツ。 S の各点が近傍 S_0 を持ち、制限 $F|_{S_0}$ がリプシッツである場合、 F は局所的にリプシッツである。

レンマ1。写像 $F: S \rightarrow \mathbb{R}^n$ を C^1 とする。また、 $D \subset S$ がコンパクトであれば、制限 $F|_D$ はリプシッツである。(Proof in (Hirsch and Smale 1973), chapter 8, section 3).

レマ2. $F: S \rightarrow \mathbb{R}^n$ を C^1 -写像とし、 $x_0 \in S$ とする。正の α が存在し、微分方程式の一意解 $x: (a, a)$ の微分方程式

$$x' = F(x) \text{ である、} \quad (S18)$$

これは初期条件 $x(0) = x_0$ を満たす(証明は(Hirsch and Smale 1973)の第8章、第2節、定理I)。

レマ3. S を \mathbb{R}^n の開部分集合とし、 $F: S \rightarrow \mathbb{R}^n$ を C^1 -写像とする。最大区間 $J = (\alpha, \beta) \subset \mathbb{R}$ 上で、 $x(t)$ を解とする。そして、任意のコンパクトな部分集合 $D \subset S$ に対して、 $x(t) \in D$ となるいくつかの $t \in J$ が存在する(証明は(Hirsch and Smale 1973)、第8章、第5節、定理)。

レマ4. $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ が bound C^1 -mapping であるとき、微分方程式

$$x' = -\frac{x}{\tau} + F(x), \quad (S19)$$

ここで $\tau > 0$ は $[0, \infty]$ 上で一意解を持つ。(Proof in (Funahashi and Nakamura 1993), Section 4, Lemma 4)。

レマ5. $F: \mathbb{R}^n \rightarrow \mathbb{R}^n$ が有界 C^1 -写像であるとき、微分方程式

$$x' = -(1/\tau + F(x))x + AF(x), \quad (S20)$$

ここで、 τ は正の定数、 A は $0 < \alpha < +\infty$, $0 \leq \beta < +\infty$ の範囲 $[-\alpha, \beta]$ に束縛される定数係数であり、 $[0, \infty]$ 上に一意解を持つ。

証明仮定に基づき、次のような FM をとることができる。

$$0 \leq F_i(x) \leq M \quad (\forall i = 1, \dots, n) \quad (S21)$$

次の微分方程式の解を見ることによって：

$$x' = -(1/\tau + M)x + AM, \quad (S22)$$

を示すことがで
きる。

$$\tau(AM) \leq x(t) \leq \tau(AM)$$

$$\min\{|x_i(0)|, \frac{1}{1+\tau M}\} \leq x_i(t) \leq \max\{|x_i(0)|, \frac{1}{1+\tau M}\} \text{ となる、} \quad (S23)$$

\max の出力を c_{\max_i} 、 \min の出力を c_{\min_i} とし、さらに $c_1 = \min\{c_{\min_i}\}$ 、 $c_2 = \max\{c_{\max_i}\}$ とすると、解 $x(t)$ は以下を満たす。

$$\frac{1}{n}c_1 \leq x(t) \leq \frac{1}{n}c_2. \quad (S24)$$

レマ2とレマ3に基づいて、区間 $[0, +\infty]$ 上に一意解が存在する。 \square

1式S20で定義されるLTCネットワークは、 $[0, \infty]$ 上で一意な解を持つことをLemma 5は示している。

レマ6. 二つの連続写像 $F, \tilde{F}: S \rightarrow \mathbb{R}^n$ をリプシッツとし、 L を F のリプシッツ定数とする、

$$|F(x) - \tilde{F}(x)| < \epsilon, \quad (S25)$$

の解であれば、 $x(t)$ と $y(t)$ が成り立つ。

$$x' = F(x) \text{ である、} \quad (S26)$$

$$y' = \tilde{F}(x), \quad (S27)$$

$x(t_0) = y(t_0)$ となるような区間上で、次のようになる。

$$|x(t) - y(t)| \leq \frac{\epsilon}{L} (e^{L|t-t_0|} - 1). \quad (S28)$$

(Proof in (Hirsch and Smale 1973), chapter 15, section 1, Theorem 3).

S3. 1定理の証明:

証明上記の定義とレンマを用いて、LTCが普遍的近似であることを証明する。

その1。 η は $(0, \min \{L_1, 3F/5, \lambda\})$ の範囲にあり、 $\epsilon_1 > 0$ 、 λ は D と S の境界 δS との距離である。 D_η :

$$D_\eta = \{x \in \mathbb{R}^n ; \exists z \in D, |x - z| \leq \eta\}. \quad (S29)$$

D はコンパクトなので、 D_η は S のコンパクト部分集合を表す。したがって、 F は D_η 、レンマ1によりリプシッツである。 $F|_{D_\eta}$ のリプシッツ定数を L_F とすると、次のような $\epsilon_1 > 0$ を選ぶことができる。

$$\epsilon_1 < \frac{\eta L_F}{2(e^{L_F T} - 1)}. \quad (S30)$$

万能近似定理に基づき、整数 N 、
 n 次元ベクトル

$\times nN$ 個の行列 A 、

$\times Nn$ 個の行列 C が

$$\max_x |F(x) - Af(\gamma x + \mu)| < \frac{\epsilon_l}{2}. \quad (S31)$$

C^1 -mapping $F^\sim: R^n \rightarrow R^n$ を次のように定義する:

$$F^\sim(x) = -(1/\tau + W_l f(\gamma x + \mu))x + W_l f(\gamma x + \mu)A, \quad (S32)$$

であり、パラメータは $W_l = W$ で式1と一致する。

システムの時定数 τ_{sys} を次のように設定する:

$$\tau_{sys} = \frac{1}{\tau/l + \tau W_l f(\gamma x + \mu)}. \quad (S33)$$

大きな τ_{sys} を選んだ:

$$(a) \forall x \in D; \tau_{sys} < \frac{\epsilon_l}{2} \quad (S34)$$

$$(b) \left| \frac{\mu}{\tau_{sys}} \right| < \frac{\eta L_{G^\sim}}{2(e^{L_{G^\sim} T} - 1)} \text{ および } \left| \frac{1}{\tau_{sys}} \right| < \frac{L_{G^\sim}}{2}, \quad (S35)$$

ここで $L_{G^\sim}/2$ は写像 $W_l f$ のリプシッツ定数である: $R^{n+N} \rightarrow R^{n+N}$ のリプシッツ定数である。条件(a)と(b)を満たすためには、 $\tau W_l < 1$ が成立しなければならない。

そして、式 (S31) と式 (S32) によって証明できる:

$$\max_{x \in D} |F(x) - F^\sim(x)| < \epsilon_l \quad (S36)$$

初期状態 $x(0) = x \sim(0) = x_0 \in D$ を持つ $x(t)$ と $\sim x(t)$ を以下の方程式の解とする:

$$\dot{x} = F(x) \text{ である、} \quad (S37)$$

$$\dot{\sim x} = F^\sim(\sim x). \quad (S38)$$

任意の $t \in [0, \infty)$ に対して、レマ6

に基づく、

$$|x(t) - \sim x(t)| \leq \frac{\epsilon_l}{L_F} (e^{L_F t} - 1) \quad (S39)$$

$$\leq \frac{\epsilon_l}{L_F} (e^{L_F T} - 1). \quad (S40)$$

$$\max_{t \in [0, \infty)} |x(t) - \sim x(t)| < \frac{\epsilon_l}{2} \quad (S41)$$

したがって、 ϵ の条件に基づい

ている、

$$t \in [0, \infty) \quad (S42)$$

第2部第1部の F^\sim で定義された以下の力学系を考えてみよう:

$$\dot{\sim x} = -\frac{1}{\tau_{sys}} \sim x + W_l f(\gamma \sim x + \mu)A. \quad (S42)$$

$y = \gamma \sim x + \mu$ とする:

$$\dot{\sim y} = \gamma \dot{\sim x} = -\frac{1}{\tau_{sys}} \sim y + E f(\sim y) + \frac{\mu}{\tau_{sys}}, \quad (S43)$$

ここで $E = \gamma W_l A$ は $N \times N$ の行列である。こ

こで

$$\sim z = (\sim x_1, \dots, \sim x_n, \sim y_1, \dots, \sim y_n), \quad (S44)$$

として、写像 $G^{\sim} : \mathbb{R}^{n+N} \rightarrow \mathbb{R}^{n+N}$ を設定する：

$$G^{\sim}(\sim \mathbf{z}) = -\frac{1}{\tau_{\text{sys}}} \sim \mathbf{z} + W f(\sim \mathbf{z}) + \frac{\mu_1}{\tau_{\text{sys}}} \text{ ,} \tag{S45}$$

どこ
だ？

$$W^{(n+N) \times (n+N)} = \begin{pmatrix} 0 & A \\ 0 & E \end{pmatrix}, \quad (S46)$$

$$\mu_1^{n+N} = \begin{pmatrix} 0 \\ \mu \end{pmatrix}. \quad (S47)$$

ここでレンマ2を用いると、以下の力学系の解を示すことができる：

$$\begin{aligned} \tilde{z} &= \tilde{G}(\tilde{z}), & y \sim(0) \\ &= \gamma x \sim(0) + \mu, & (S48) \end{aligned}$$

は式S42の解と等価である。

新しい力学系 $G: \mathbb{R}^{n+N} \rightarrow \mathbb{R}^{n+N}$ を以下のように定義しよう：

$$G(z) = -\frac{1}{\tau_{\text{sys}}} z + Wf(z), \quad (S49)$$

ここで、 $z = (x_1, \dots, x_n, y_1, \dots, y_n)$ とすると、以下の力学系は

$$\dot{z} = -\frac{1}{\tau_{\text{sys}}} z + Wf(z), \quad (S50)$$

は、 $h(t) = (h_1(t), \dots, h_N(t))$ を隠れ状態とし、 $u(t) = (u_1(t), \dots, u_n(t))$ をシステムの出力状態とすると、LTCによって実現できる。 \tilde{G} と G はともに C^1 -写像であり、 $f'(x)$ は束縛されているので、写像 $\tilde{z} \mapsto Wf(\tilde{z})$ は \mathbb{R}^{n+N} 上でリプシッツ定数 $L_{\tilde{G}}/2$ を持つ。 $L_{\tilde{G}}/2$ は τ_{sys} 上の条件(b)により $-z \sim / \tau_{\text{sys}}$ に対してリプシッツ定数であるので、 $L_{\tilde{G}}$ は \tilde{G} のリプシッツ定数である。

式S45、式S49、および τ_{sys} の条件(b)から、以下のように導ける：

$$|G \sim(z) - G(z)| = \left| \frac{\mu}{\tau_{\text{sys}}} \right| < \frac{\eta L_{\tilde{G}}}{2(e^{L_{\tilde{G}} T} - 1)}. \quad (S51)$$

Accordingly, we can set $\tilde{z}(t)$ and $z(t)$, solutions of the dynamical systems:

$$\begin{aligned} \dot{\tilde{z}} &= \tilde{G}(\tilde{z}), & x \sim(0) &= x_0 \in \end{aligned} \quad (S52)$$

$$y \sim(0) = \gamma x_0 + \mu$$

$$\dot{z} = G(z), & u(0) = x_0 \in \quad (S53)$$

$$\begin{aligned} D h \sim(0) &= \\ \gamma x_0 + \mu & \end{aligned}$$

レンマ6により、次の

$$\max_t |\tilde{z}(t) - z(t)| < \frac{\eta}{2}, \quad (S54)$$

ようになる。

$$t \leq T$$

したがって、こんな

$$\max_{t \leq T} |\tilde{x}(t) - x(t)| < \frac{\eta}{2} \quad (S55)$$

る：

$$t \leq T$$

Part3. さて、式S41と式S55を用いると、正?の場合、内部力学的状態 $z(t)$ を持つLTCを設計することができる。

τ_{sys} と W 。 $x' = F(x)$ を満たす $x(t)$ に対して、 $u(0) = x(0)$ 、 $h(0) = \gamma x(0) + \mu$ でネットワークを初期化すると、次のようになる：

$$\max_{t \leq T} |x(t) - u(t)| < \frac{\eta}{2} = \eta < ?_0. \quad (S56)$$

□

備考LTCは隠れ層の要素が互いにリカレント接続を持つことを可能にする。しかし、これは隠れノードから出力ユニットへのフィードフォワード接続の流れを仮定している。我々はシステムへの入力を仮定せず、隠れノードと出力ユニットが自律的な力学系の有限軌道を近似できることを示した。

S4定理4の証明

このセクションでは、数学的な概念を説明し、証明に必要な概念を再確認する。時間連続ニューラルネットワークの表現力に関する我々の理論的結果の主な状態は、主に(Raghu et al. 2017)で静的なディープニューラルネットワークに対して導入されたex-pressivity measureである軌跡長に基づいて構築されている。したがって、モデルの連続的な性質に起因する、慎重な検討を伴う同様のステップをたどることは直感的である。

S4.1 表記

ニューラルネットワークアーキテクチャ- $f_{n,k}(x(t), l(t, \theta))$ によってニューラルネットワークアーキテクチャを決定する。

k 、ニューロンの総数 $N = n \times k$.

神経状態、 $x(t)$ - ネットワーク f の層 d について $x^{(d)}(t)$ は層の神経状態を表し、以下のサイズの行列である。

$x \in \mathbb{R}^{k \times m}$ 、 m は入力時系列のサイズ。

入力、 $I(t)$ - は、 $t \in [0, t_{max}]$ の2次元軌跡を含む2次元行列である。

ネットワーク・パラメータ θ には、各層 d の重み行列 $W^{(d)} \sim N(0, \sigma^2/k)$ とバイアス・ベクトルが含まれる。
 $b^{(d)} \sim N(0, \sigma^2)$ 。CT-RNNの場合、ベクトル・パラメータ $\tau^{(d)}$ も $\sim N(0, \sigma^2)$ からサンプリングされる。

垂直成分と平行成分- あるベクトル x と y に対して、それぞれのベクトルを $y = y_{\parallel} + y_{\perp}$ のように分解することができる。すなわち、 y_{\parallel} は x に平行な y の成分を表し、 y_{\perp} は x に垂直な成分である。

重み行列分解- (Raghu et al. 2017)は、ゼロでないベクトル x と y が与えられ、フルランク行列

W の行列分解を x と y に関して次のように書くことができる： $W = W_{\parallel}^{\parallel} W + W_{\perp}^{\perp} W_{\perp}^{\perp}$ ，このようなものである。

$W_{\perp}^{\perp} x = 0, W_{\perp}^{\perp} x = 0, y^{\top} W_{\parallel} = 0, y^{\top} W_{\perp}^{\perp} = 0$ 。この表記では、左の分解上付き添え字は次のようになる。
また、 x に関する W_{\perp}^{\perp} ：

$W_{\perp}^{\perp} = W - W_{\parallel}^{\parallel}$ (Raghu et al. 2017)。

レンマ7. 射影の独立性 (Raghu et al.) $N(0, \sigma^2)$ の形で描かれた iid エントリを持つ行列 W が与えられたとき、 x に関してその分解行列 W_{\perp}^{\perp} と $W_{\parallel}^{\parallel}$ は独立な確率変数である。

証明は(Raghu et al. 2017), Appendix, Lemma 2にある。

レンマ8. ガウス・ベクトルのノルム (Raghu et al. 2017)。ガウスベクトル $x \in \mathbb{R}^k$ のノルムは、そのエントリーが iid でサンプリングされた
 $\sim N(0, \sigma^2)$ は次式で与えられる：

$$E[\|x\|] = \sigma \sqrt{2} \frac{\Gamma((k+1)/2)}{\Gamma(k/2)} \quad (S57)$$

証明は(Raghu et al. 2017), Appendix, Lemma 3にある。

レンマ9. 投影のノルム (Raghu et al. 2017)。レンマ8の条件を持つ $W^{k \times k}$ 、2つのベクトル x と y について、 x に垂直な非ゼロベクトルである x_{\perp} 、以下が成り立つ：

$$E[\|x_{\perp} W x_{\perp}\|] = \|x_{\perp}\| \sigma \sqrt{2} \frac{\Gamma((k)/2)}{\Gamma((k-1)/2)} \geq \|x_{\perp}\| \sigma \sqrt{2} \left(\frac{k-1}{2}\right)^{\frac{1}{2}} \quad (S58)$$

また、(Raghu et al. 2017)で示されている：“if I_A is a identity matrix with non-zero diagonal entry i if $i \in A \subset [k]$ で、 $|A| \geq 2$ である：

$$E[\|I_A W x_{\perp}\|] = \|x_{\perp}\| \sigma \sqrt{2} \frac{\Gamma(|A|/2)}{\Gamma((|A|-1)/2)} \geq \|x_{\perp}\| \sigma \sqrt{2} \left(\frac{|A|-1}{2}\right)^{\frac{1}{2}} \quad (S59)$$

証明は(Raghu et al. 2017), Appendix, Lemma 4にある。

レンマ10. ノルムと変換 (Raghu et al.) X がゼロ平均の多変量ガウスで、対角共分散行列を持ち、 μ が定数のベクトルであるとき、次が成り立つ：

$$E[\|X - \mu\|] \geq E[\|X\|] \quad (S60)$$

証明は(Raghu et al. 2017), Appendix, Lemma 5にある。

S4.2定理4の証明の 始まり

我々はまずニューラルODEの下界を確立し、その結果をCT-RNNの下界に拡張する。

証明ニューラルODEの連続層 $d+1$ の場合、 $t+\delta t$ と t の状態間の勾配、 $x^{d+1}(t+\delta t)$ と $x^{d+1}(t)$ によって決定される：

$$\frac{dx^{d+1}}{dt}$$

dt

$$= f(h^{(d)}), \quad h^{(d)} = W x^{(d)(d)} + b^{(d)}. \quad (\text{S61})$$

したがって、 $z^{(d+1)}(t)$ で示される潜在表現（隠れた状態 $x^{(d+1)}$ の最初の2つの主成分）については、この勾配は次式で求めることができる：

$$\frac{dz^{(d+1)}}{dt} = f(h^{(d)}), \quad h^{(d)} = W z^{(d)(d)} + b^{(d)} \quad (\text{S62})$$

バイアスがゼロの場合について説明を続け、ゼロでない場合については後で説明しよう。

$W^{(d)} = W_{\parallel}^{(d)} + W_{\perp}^{(d)}$ として、 $W^{(d)}$ を $z^{(d)}$ に関して分解する。この分解の場合、隠れ状態 $h = W$ となる。^(d+1)

$W_{\perp}^{(d)}$ 垂直成分は $z^{(d)}$ をゼロにマップする。

f がハード・タン活性化によって定義されているかのように、勾配状態が飽和していないインデックスの集合を決定する：

$$A_{W_{\parallel}^{(d)}} = \{i : i \in [k], |\hat{h}_i^{(d+1)}| < 1\}. \quad (S63)$$

$W^{(d)}$ の分解成分は独立な確率変数であるため、レンマ9に基づき、期待値を構築することができる。勾配状態を以下のように変換する：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] = E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[f(Wz)^{(d)} \right]. \quad (S64)$$

ここで、 $W_{\parallel}^{(d)}$ を条件とすれば、右辺のノルムを非飽和指数 $A_{W_{\parallel}^{(d)}}$ 以下の通りである：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] = E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \sum_i \frac{z^{(d)} + (W_{\parallel}^{(d)})^T z^{(d)}}{(W_{\perp}^{(d)})^T z^{(d)}}. \quad (S65)$$

式S65の漸化式を導出する必要がある。そのために、勾配状態を $z^{(d)}$ に関して次のように分解する。

$$\frac{dz^{(d)}}{dt} = \frac{dz_{\parallel}^{(d)}}{dt} + \frac{dz_{\perp}^{(d)}}{dt}. \quad (S66)$$

$h^{(d+1)}$ 、すべての不飽和ユニットの潜在勾配ベクトル、およびゼロ化された飽和ユニットである。また

$z \sim^{(d+1)}$ に関して重み行列の列空間を次のように分解する： $w^{(d)} = w^{\perp(d)} + w^{(d)}$ 。すると、定義により以下の式が得られる：

$$\frac{dz^{(d+1)}}{dt} = Wz \mathbf{1}^{(d)} A - \langle Wz \mathbf{1}^{(d)} A, \hat{z}^{(d+1)} \rangle \hat{z}^{(d+1)}, \quad (S66)$$

$$W^{\perp(d)} z^{(d)} = W^{(d)} z^{(d)} - \langle W^{(d)} z^{(d)}, \hat{z}^{(d+1)} \rangle \hat{z}^{(d+1)} \quad (S67)$$

式S66と式S67を見ると、提供された定義に基づき、これらの右辺はどのような場合でも互いに等しい。したがって、これらの左辺も等価である。より正確には

$$\frac{dz^{(d+1)}}{dt} \cdot \mathbf{1}_A = Wz^{\perp(d)} \cdot \mathbf{1}_A. \quad (S68)$$

式S68の記述により、以下の不等式を決定することができ、再帰の最初のステップを構築することができる：

$$\frac{dz^{(d+1)}}{dt} \geq \frac{dz^{(d+1)}}{dt} \cdot \mathbf{1}_A. \quad (S69)$$

ここでS65式に戻り、以下の分解をプラグインしてみよう：

$$\frac{dz^{(d)}}{dt} = \frac{dz_{\parallel}^{(d)}}{dt} + \frac{dz_{\perp}^{(d)}}{dt} \quad (S70)$$

$$W_{\perp}^{(d)} = \|W_{\perp}^{(d)}\| W_{\perp}^{(d)} \quad W_{\parallel}^{(d)} = \|W_{\parallel}^{(d)}\| W_{\parallel}^{(d)}, \quad (S71)$$

我々は

持って

いる：

$$E_{\parallel} E_{\perp} \left[\frac{dz^{(d+1)}}{dt} \right] = E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \sum_i \frac{z^{(d)} + (W_{\parallel}^{(d)})^T z^{(d)}}{(W_{\perp}^{(d)})^T z^{(d)}} A_{W_{\parallel}^{(d)}} \quad (S72)$$

$E_{W^{(d)}}$

$$\frac{d}{dt} \dots = \dots \tag{S72}$$

$$\perp \dots \perp^{(d)} \rangle_i z_{\perp}^{(d)} + (W + W^{(d)} \perp^{(d)} \rangle_i z_{\perp}^{(d)} \frac{2}{1/2} \# \tag{S73}$$

$$\parallel \parallel \parallel \parallel$$

定理4で述べたように、我々は入力をその垂直成分で条件付けた。そこで、平行成分である $\|w_{\perp}^{(d)}$ と $\|w_{\parallel}^{(d)}$ を削除し、式 S69 を用いて、それらの垂直成分についても状態の再帰を以下のように記述する：

$$E_{w^{(d)}} \frac{dz^{(d+1)}}{dt} \geq E_{w_{\parallel}^{(d)}} E_{w_{\perp}^{(d)}} \sum_{i \in \mathcal{A}(w_{\parallel}^{(d)})} \left(\frac{z_{\perp}^{(d)}}{W_{\perp}^{(d)}} \right)_i z_{\perp}^{(d)} + \left(\frac{z_{\parallel}^{(d)}}{W_{\parallel}^{(d)}} \right)_i z_{\parallel}^{(d)} \quad (S74)$$

この項 $W_{\perp}^{(d)}$ は、内的期待値が $w^{(d)}$ を条件としているので、一定である。ここで、レンマ10を用いると、次のようになる：

$$E_{w_{\perp}^{(d)}} \sum_{i \in \mathcal{A}(w_{\parallel}^{(d)})} \left(\frac{z_{\perp}^{(d)}}{W_{\perp}^{(d)}} \right)_i z_{\perp}^{(d)} + \left(\frac{z_{\parallel}^{(d)}}{W_{\parallel}^{(d)}} \right)_i z_{\parallel}^{(d)} \geq \quad (S75)$$

$$E_{w_{\perp}^{(d)}} \sum_{i \in \mathcal{A}(w_{\parallel}^{(d)})} \left(\frac{z_{\perp}^{(d)}}{W_{\perp}^{(d)}} \right)_i z_{\perp}^{(d)} \geq \quad (S76)$$

レンマ9を適用するところ
なる：

$$E_{w_{\perp}^{(d)}} \sum_{i \in \mathcal{A}(w_{\parallel}^{(d)})} \left(\frac{z_{\perp}^{(d)}}{W_{\perp}^{(d)}} \right)_i z_{\perp}^{(d)} \geq \frac{\sigma}{\sqrt{k}} \sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} E_{w_{\perp}^{(d)}} z_{\perp}^{(d)} \quad (S77)$$

$p = P(|h^{(d+1)}| < 1)$ 、条件 $|A_{w_{\parallel}^{(d)}}| \geq 2$ のハード・タン活性化関数を選択したので、次のようになる。

$$\sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} \geq \frac{1}{2} \sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} \quad \text{したがってこうなる：}$$

$$E_{w^{(d)}} \frac{dz^{(d+1)}}{dt} \geq \frac{1}{\sqrt{2}} \sum_{j=2}^k \frac{p^j (1-p)^{k-j}}{\sqrt{k}} \frac{\sigma}{\sqrt{k}} \sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} E_{w_{\perp}^{(d)}} z_{\perp}^{(d)} \quad (S78)$$

ここで、 $|A_{w_{\parallel}^{(d)}}|$ を j としていることに留意されたい。ここで、和によって表される二項分布を考慮することによって、 \sqrt{k} 項を束縛する必要がある。その結果、式S78の和を次のように書き換えることができる：

$$\begin{aligned} \sum_{j=2}^k \frac{p^j (1-p)^{k-j}}{\sqrt{k}} \frac{\sigma}{\sqrt{k}} \sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} &= \sum_{j=2}^k \frac{p^j (1-p)^{k-j}}{\sqrt{k}} \frac{\sigma}{\sqrt{k}} \sqrt{\frac{2|A_{w_{\parallel}^{(d)}}| - 3}{2}} \\ &= -\sigma_w \sqrt{k} p (1-p)^{k-1} + k p \frac{\sigma}{\sqrt{k}} \sum_{j=2}^k \frac{1}{j-1} (1-p)^{k-j} \end{aligned}$$

であり、 $1/\sqrt{x}$ によるJensenの不等式を利用することで、 XT は二項分布 $(k-1, p)$ の期待値であるため、以下のように単純化できる(Raghu et al. 2017)：

$$\sum_{j=2}^k \frac{1}{j-1} (1-p)^{k-j} \geq \frac{1}{\sum_{j=2}^k (j-1) (1-p)^{k-j}} = \frac{1}{\sqrt{k}} \quad \text{従って} \quad \sum_{j=2}^k \frac{1}{j-1} (1-p)^{k-j} \geq \frac{1}{\sqrt{k}}$$

$$\begin{matrix} k - 1 \\ j - 1 \end{matrix}$$

$$p^{j-1}(1 - p)$$

$$\sum_{i=0}^{k-1} p^i$$

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq \frac{1}{2} - \sigma_w \sqrt{\frac{1}{k} \left(1 - \frac{1}{\sigma_w} \right)^{k-1}} + \sigma_w \sqrt{\frac{kp}{(k-1)p+1}} E \left[\frac{dz^{(d)}}{dt} \right] \quad (S79)$$

ここで、 p の範囲を見つける必要がある。(Raghu et al. 2017) は、Hard-tanh活性化について、入力引数 $i/A \sim N(0, \sigma^2)$ に対して、 $h^{(d+1)}$ が σ_w よりも小さい分散を持つ確率変数であるという事実が与えられれば、 $p = P(|h^{(d+1)}| < 1)$ を下界にできることを示した、

以下の通

りである

$$p = P(|h_i| < 1) \geq P(|A| < 1) \geq \frac{1}{\sqrt{2\pi}\sigma} \int_{-1}^1 e^{-\frac{t^2}{2\sigma^2}} dt \geq 1, \quad (S80)$$

を計算し、 $\frac{1}{2}$ に等しい上限を求める (Raghu et al. 2017)。したがって、式は次のようになる：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq \frac{1}{2} - \sigma_w \sqrt{\frac{1}{k} \left(1 - \frac{1}{\sigma_w} \right)^{k-1}} + \sigma_w \sqrt{\frac{kp}{(k-1)p+1}} E \left[\frac{dz^{(d)}}{dt} \right] \quad (S81)$$

そして、いくつかの簡略化を加えた：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq \frac{1}{2} - \sqrt{k} \left(1 - \frac{1}{\sigma_w} \right)^{k-1} + (2\pi)^{1/4} \frac{\sqrt{k\sigma_w}}{(k-1) + \sqrt{2\pi}\sigma_w} E \left[\frac{dz^{(d)}}{dt} \right] \quad (S82)$$

ここで、式S82をロールバックして入力にたどり着きたい。そのためには、右辺の期待項を次のように置き換える：

$$E \left[\frac{dz^{(d)}}{dt} \right] = E \left[\int_t^{\infty} \frac{dz^{(d)}}{dt} dt \right] \quad (S83)$$

命題 1. $f: \mathbb{R} \rightarrow \mathbb{S}$ をバナッハ空間 \mathbb{S} 上の積分可能な関数とする：

$$\int_t^\infty \|f(t)\| dt \geq \int_t^\infty f(t) dt. \quad (S84)$$

証明: $x = \int_t^\infty f(t) dt \in \mathbb{S}, \Lambda \in \mathbb{S}^*$ with $\|\Lambda\| = 1$ とする。すると次のようになる：

$$\Lambda x = \int_t^\infty \Lambda f(t) dt \leq \|\Lambda\|_{\mathbb{S}^*} \int_t^\infty \|f(t)\|_{\mathbb{S}} dt = \int_t^\infty \|f(t)\| dt. \quad (S85)$$

今、ハーン・バナッハに基づく次のようになる □

： $\|x\| \leq \int_t^\infty \|f(t)\| dt$. 命題1と式S83に基づき、次式が得

られる：

$$E \left[\int_t^\infty \frac{dz^{(d)}}{dt} dt \right] \geq E \left[\int_t^\infty \frac{dz^{(d)}}{dt} dt \right] = I(z^{(d)}(t)). \quad (S86)$$

ここで、式S82を再帰的に展開して入力 $I(t)$ を求め、 $c_1 = I(I^{(t)})$ とすると、次のようになる：

$$E \left[\frac{dz^{(d+1)}}{dt} \right] \geq \frac{1}{2} - \sqrt{k} \left(1 - \frac{1}{\sigma_w} \right)^{k-1} + (2\pi)^{1/4} \frac{\sqrt{k\sigma_w}}{(k-1) + \sqrt{2\pi}\sigma_w} E \left[\frac{dz^{(d)}}{dt} \right]$$

$$E_{W^{(d)}} \frac{d^2 z^{(d+1)}}{dt^2} \geq \frac{\sqrt{2}}{2} \left(k \left(1 - \frac{1}{\sigma_w} \right)^{k-1} + (2\pi)^{-1/4} \sqrt{\frac{w}{(k-1) + 2\pi\sigma_w}} \right) c_1 l(l(t)) \tag{S87}$$

最後に、境界の漸近形と、連続する時点に直交する入力軌道について $c_1 = 1$ を考慮すると、次のようになる：

$$E_{W^{(d)}} \frac{d^2 z^{(d+1)}}{dt^2} \geq O \left(\sqrt{\frac{kq_w}{k + \sigma_w}} \right) l(t) . \tag{S88}$$

式S88は、ニューラルODEアーキテクチャの隠れ状態（原理的には構成要素の状態、 z ）の長さの無限小数ごとの下界を示している。その結果、全体の軌跡の長さは以下ようになる：

$$\mathbb{E} \|z^{(d)}(t)\| \leq O\left(\sqrt{\frac{k\sigma_w}{k+\sigma_w}}\right)^{d \times L} \|l(t)\|, \quad (\text{S89})$$

L は ODE ステップ数である。最後に、バイアスがゼロでない場合を考える：

表記法のセクションで述べたように、ネットワーク・パラメータは $W^{(d)} \sim N(0, \sigma^2/k)$ 、バイアス・ベクトルは $b^{(d)} \sim N(0, \sigma_b^2)$ として設定される。したがって、 $h^{(d+1)}$ の分散は $\sigma_i^2 + \sigma_w^2$ よりも小さくなる。したがって、我々は次のようになる（Raghu et al. 2017）^b：

$$p = P(|h_i^{(d+1)}| < 1) \geq \frac{1}{2\pi\sqrt{\sigma_w^2 + \sigma_b^2}} \quad (\text{S90})$$

これを式S79に置き換え、さらに単純化するとこうなる：

$$\mathbb{E} \|z^{(d)}(t)\| \geq O\left(\frac{\sigma_w}{\sqrt{\sigma_w^2 + \sigma_b^2 + k}}\right)^{d \times L} \|l(t)\|, \quad (\text{S91})$$

の定理4が得られる。

CT-RNNの軌跡長下界の導出 CT-RNNの連続する層 $d+1$ について、 $t + \delta t$ と t 、 $x^{(d+1)}(t + \delta t)$ と $x^{(d+1)}(t)$ の状態間の勾配は次式で決定される：

$$\frac{dx^{(d+1)}}{dt} = -W_\tau^{(d+1)} x^{(d+1)} + f(h^{(d)}), \quad h^{(d)} = W^{(d)} x^{(d)} + b^{(d)}. \quad (\text{S92})$$

$W_\tau^{(d+1)}$ はパラメータ・ベクトル¹を表し、厳密に正であることが条件となる。したがって、潜在表現（隠れた状態 x の最初の2つの主成分 $^{(d+1)}$ ）、これは $z^{(d+1)}(t)$ で示され、この勾配は次式で求めることができる：

$$\frac{dz^{(d+1)}}{dt} = -W_\tau^{(d+1)} z^{(d+1)} + f(h^{(d)}), \quad h^{(d)} = W^{(d)} z^{(d)} + b^{(d)} \quad (\text{S93})$$

このODEを明示的にオイラー離散化するとこうなる：

$$z^{(d+1)}(t + \delta t) = (1 - \delta t W_\tau^{(d+1)}) z^{(d+1)} + \delta t f(h^{(d)}), \quad h^{(d)} = W z^{(d)} + b^{(d)}. \quad (\text{S94})$$

と同じ離散化モデルをニューラル ODEs に適用すると、次のようになる：

$$z^{(d+1)}(t + \delta t) = z^{(d+1)} + \delta t f(h^{(d)}), \quad h^{(d)} = W z^{(d)} + b^{(d)}. \quad (\text{S95})$$

2つの表現の違いは、 $z^{(d+1)}$ の前の $-\delta t W_\tau^{(d+1)}$ の項だけである。^(d+1)これは、折りたたまれた正規分布 $N(|x|; \mu_Y, \sigma_Y)$ からサンプリングされた厳密に正の確率変数であり、^τ平均 $\mu_Y = \frac{q}{\pi} e^{\frac{-\mu^2 + 2\sigma^2}{2}}$ and variance $\sigma_Y^2 = \mu^2 + \sigma^2 - \mu^2_Y$ (Tsagris, Beneki, and Hassani 2014). μ と σ はは確率変数 x 上の正規分布の平均と分散、 Φ は正規累積分布関数である。分散が σ^2 のゼロ平均正規分布では、次式が得られる。

$$N(|W|; \sigma, \frac{2}{\pi}, (1 - \frac{2}{\pi})\sigma_b^2). \quad (\text{S96})$$

従って、CT-RNNの下界を近似し、簡略化された漸近形とする：

$$\mathbb{E} \|z^{(d)}(t)\| \geq O\left(\frac{(\sigma_w - \sigma_b)k}{\sigma^2 + \sigma^2 + k}\right)^{d \times L} \|l(t)\| \text{である、} \quad (\text{S97})$$

$$w \quad b \quad w \quad b$$

□

これによりCT-RNNの定理が導かれる。

定理5の証明

LTCのパラメータ分布

各層 d の重み行列 $W^{(d)} \sim N(0, \sigma^2/k)$ 。バイアスベクトル $b^{(d)} \sim N(0, \sigma^2)$ 。ベクトルパラメータ $W_{\tau}^{(d+1)}$ は厳密に正であり、折り畳まれた正規分布からサンプリングされる (Tsagris, Beneki, and Hassani 2014)。

$N(|W_{\tau}|; \sigma_{\tau}^2, (1 - \frac{2}{\pi})\sigma^2)$ 。パラメータはニューロンの時定数の逆数を表す¹。パラメータ $A_{\tau}^{(d)}$

は $N(0, \sigma^2/k)$ からサンプリングされた重み行列である。

証明 LTCネットワークの連続する層 $d+1$ の場合、 $t + \delta t$ と t の状態間の勾配 $x^{(d+1)}(t + \delta t)$ と $x^{(d+1)}(t)$ によって決定される：

$$\frac{dx^{(d+1)}}{dt} = -(w_{\tau}^{(d+1)} + f(h^{(d)}))x^{(d+1)} + A^{(d)}f(h^{(d)}), \quad h^{(d)} = W^{(d)}x^{(d)} + b^{(d)}. \quad (S98)$$

したがって、 $z^{(d+1)}(t)$ で示される潜在表現（隠れた状態 $x^{(d+1)}$ の最初の2つの主成分）については、この勾配は次式で求めることができる：

$$\frac{dz^{(d+1)}}{dt} = -(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} + A^{(d)}f(h^{(d)}), \quad h^{(d)} = W^{(d)}z^{(d)} + b^{(d)}. \quad (S99)$$

まず、式S99の両側からノルムの期待値をとる。一方、式S64と同様に、レンマ9に基づき、ウェイト行列 $W^{(d)}$ の平行成分と直交成分に対する期待値を以下のように分解する：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] = E_{W_{\parallel}^{(d)}} \left[-(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} + A^{(d)}f(h^{(d)}) \right]. \quad (S100)$$

ここで、差の規範対規範の差について、以下の不等式を導くことができる：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] = \quad (S101)$$

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[A^{(d)}f(h^{(d)}) - (w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} \right] \geq 0. \quad (S102)$$

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[A^{(d)}f(h^{(d)}) \right] - \left[(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} \right] \geq 0. \quad (S103)$$

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[A^{(d)}f(h^{(d)}) \right] - E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} \right]. \quad (S104)$$

まず式S104の**右式**に注目しよう。ノルムは次のように積のノルムに分割できる：

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} \right]. \quad (S105)$$

ここで、 $E[XY] = E[X]E[Y]$ というルールで期待値を条件付けると、次のようになる：

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \left[(w_{\tau}^{(d+1)} + f(h^{(d)}))z^{(d+1)} \right] = E_{W_{\parallel}^{(d)}} \left[(w_{\tau}^{(d+1)} + f(h^{(d)})) \right] E_{W_{\perp}^{(d)}} \left[z^{(d+1)} \right]. \quad (S106)$$

が飽和しないインデックスの集合を決定し、それがハード・タン活性化によって定義されると仮定する：

$$A_{W_{\parallel}^{(d)}} = \{i : i \in [k], |h_i^{(d+1)}| < 1\}. \quad (S107)$$

ここで、 $W_{\parallel}^{(d)}$ を条件とすれば、第一ノルムを以下のように非飽和インデックス上の和 $A_{\parallel}^{(d)}$ で置き換えることができる：

$$E_{W_{\parallel}^{(d)}} E_{W_{\perp}^{(d)}} \sum_{i \in A_{\parallel}^{(d)}} \left(W_{\perp}^{(d)} + \frac{w_{\tau}^{(d+1)}}{|A|} \right) z_i^{(d)} + \left(W_{\parallel}^{(d)} + \frac{w_{\tau}^{(d+1)}}{|A|} \right) z_i^{(d)2} \quad (S108)$$

式 (S108) において τ

定数の重みの平均効果を決定する。

は、各状態の計算における時

$|A|$ は非飽和状態の数である。さて、式S65から式S77まで同様のステップを踏み、式S108にレンマ9を適用すると、次のようになる：

$$\begin{aligned} E_{W_{\perp}^{(d)}} \sum_i \left(\frac{w_{\tau}^{(d+1)}}{|A_{W_{\parallel}^{(d)}}|} \right) z_{\perp}^{(d)2} E_{W^{(d)}} z_{\perp}^{(d+1)} &\geq \\ \sqrt{\frac{\sigma_w^2}{k/A} + \frac{\sigma_b^2}{|W_{\parallel}^{(d)}|^2}} \sqrt{\frac{2|A_{W_{\parallel}^{(d)}}|}{2}} E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} & \end{aligned} \quad (S109)$$

$p = P(|h_i^{(d+1)}| < 1)$ 、条件 $|A_{W_{\parallel}^{(d)}}| \geq 2$ のハード・タン活性化関数を選択したので、次のようになる。

$\sqrt{\frac{1}{2}} \geq \sqrt{\frac{1}{2}} \frac{\sigma_w}{|A_{W_{\parallel}^{(d)}}|}$ したがってさらに単純化できる：

$$\begin{aligned} E_{W_{\perp}^{(d)}} \sum_i \left(\frac{w_{\tau}^{(d+1)}}{|A_{W_{\parallel}^{(d)}}|} \right) z_{\perp}^{(d)2} E_{W^{(d)}} z_{\perp}^{(d+1)} &\geq \\ \sqrt{\frac{1}{2}} \frac{\sigma_w}{|A_{W_{\parallel}^{(d)}}|} \sqrt{\frac{2|A_{W_{\parallel}^{(d)}}|}{2}} E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} & \end{aligned} \quad (S110)$$

最後に

$$E_{W^{(d)}} \left(\frac{w_{\tau}^{(d+1)}}{|A_{W_{\parallel}^{(d)}}|} \right) z_{\perp}^{(d+1)} \geq \frac{\sigma_w}{2} \frac{1}{|A_{W_{\parallel}^{(d)}}|} E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} \quad (S111)$$

ここで、式S78からS79までの計算ステップを踏むと、次のようになる：

$$\begin{aligned} E_{W^{(d)}} \left(\frac{w_{\tau}^{(d+1)}}{|A_{W_{\parallel}^{(d)}}|} \right) z_{\perp}^{(d+1)} &\geq \frac{\sigma_w}{2} \frac{1}{|A_{W_{\parallel}^{(d)}}|} E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} \\ \frac{1}{\sqrt{2}} - \sigma_w k p (1 - p) + \sigma_w \sqrt{\frac{1}{(k-1)p+1}} & E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} \end{aligned} \quad (S112)$$

前述のように、ネットワーク・パラメータは $W^{(d)} \sim N(0, \sigma^2/k)$ 、バイアス・ベクトルは $b^{(d)} \sim N(0, \sigma^2)$ として設定される。したがって $h^{(d+1)}$ の分散は $\sigma_w^2 + \sigma_b^2$ よりも小さくなる。したがって、我々は次のようになる (Raghu et al. 2017)：

$$p = P(|h_i^{(d+1)}| < 1) \geq \frac{1}{\sqrt{2\pi} \sqrt{\sigma_w^2 + \sigma_b^2}} \quad (S113)$$

これにより、式S104の右式について以下のような漸近的境界が得られる：

$$\begin{aligned} E_{W_{\perp}^{(d)}} \left(\frac{w_{\tau}^{(d+1)}}{|A_{W_{\parallel}^{(d)}}|} \right) z_{\perp}^{(d+1)} &\geq \frac{\sigma_w}{\sqrt{\sigma_w^2 + \sigma_b^2 + k}} \sqrt{\frac{1}{(k-1)p+1}} E_{W^{(d)}} z_{\perp}^{(d)} E_{W^{(d+1)}} z_{\perp}^{(d+1)} \end{aligned} \quad (S114)$$

ここで、式 (S104) の左式に取り組んでみよう

：

$$E_{W_{\perp}^{(d)}} E_{W_{\perp}^{(d)}} \left(\frac{h}{A^{(d)}} f(h^{(d)}) \right) \quad (S115)$$

A は定数として機能するので、ノルムと期待値から外すことができる。結果として得られるノルムの期待値は、Hard-tanhの活性を持つディープニューラルネットワーク f を正確に表現し、それに対して (Raghu et al. 2017) は以下のように束縛できることを示した：

$$\left| A_{(d)} \left| E_{W_{\perp}}^{(d)} E_{W_{\perp}}^{(d)} \right| \right| \geq O \left(\frac{\sigma_w \sqrt{k}}{\sqrt{\sigma_w^2 + \sigma_b^2 + k} \sqrt{\sigma_w^2 + \sigma_b^2}} \right) \left| A_{(d)} \left| E_{\perp}^{(d)} \right| \right| \quad (\text{S116})$$

また、 $A \sim N(0, \sigma^2)$ であるから、境界は以下のように計算できる：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq 0 \quad \sigma_w \sqrt{k} \quad E_{Z^{(d)}} \quad (S117)$$

したがって、隠れた状態の勾配の垂直方向の区画については、次のようになる：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq 0 \quad \sigma_w \sqrt{k} \quad E_{Z^{(d)}} \quad E_{Z^{(d+1)}} \quad (S118)$$

さらに単純化して、システムのダイナミクスの無限小値 δt ごとに再帰を形成していることを考慮すると、次のような漸近的境界が得られる：

$$E_{W^{(d)}} \left[\frac{dz^{(d+1)}}{dt} \right] \geq 0 \quad \sigma_w \sqrt{k} \quad E_{Z^{(d)}} \quad \sigma_w + \frac{z^{(d+1)}}{\min(\delta t, L)} \quad (S119)$$

さて、先ほどと同様に、 n 層のニューラルネットワーク f を再帰的に展開して入力に到達させ、 $c_1 = l(I(t)) \approx 1$ とする、

を計算し、長さ T の入力シーケンスに対する境界を確立すると、ネットワークの層 d について次のようになる：

$$E_{f^{(d)}}(t) \geq 0 \quad \sigma_w \sqrt{k} \quad \sigma_w + \frac{z^{(d)}}{\min(\delta t, L)} \quad l(I(t)) \quad (S120)$$

式S120は、定理の記述を与える。

□

S5実験セットアップ-セクション6

ここでは、表3、4、5、6で取り上げたタスクの実験セットアップについて説明する。

各実験において、75:10:15の割合で訓練-検証-テストを行った。各トレーニングエポックの後、検証メトリックを評価した。トレーニングの全過程で最高の検証メトリックを達成した構成のネットワーク重みのバックアップを保持した。訓練プロセスの最後に、バックアップした重みを復元し、テストセットでネットワークを評価した。重みの初期化を変えてこの手順を5回繰り返し、平均と標準偏差を表3、4、5、6に報告した。ハイパーパラメーターを表S1に示す。

各RNNは32の隠れユニットから構成される。タスクごとに異なる数の出力ユニットを必要とするため、RNNの出力は、必要な次元に出力を投影するために学習可能な線形層を介して供給された。我々の実験セットアップの目的は、最高の予測モデルを構築することではなく、様々なRNNモデルの表現力と汎化能力を経験的に比較することであることに注意。

すべてのRNNモデルをTensorFlow1.14で実装した。再現性を高めるため、すべてのコードとデータを投稿と一緒に提出し、受理され次第公開する予定である。

ODEソルバー 微分方程式をシミュレートするために、CT-RNNには陽解法オイラー法、Neural ODEには(Chen et al. 2018)で提案されている4次ルンゲクッタ法、LTCには我々の融合ODEソルバーを使用した。すべてのODEソルバーは固定ステップソルバーであった。時間ステップは入力サンプリング周波数の1/6に設定され、すなわち、

各RNNステップは6つのODEソルバーステップで構成される。

ハンドジェスチャーのセグメンテーション 実験はハンドジェスチャーの時間的セグメンテーションに関するものである。データセットは、一連の手のジェスチャーを行う個人の7つの録画から構成される（Wagner et al.）各タイムステップの入力特徴は、動き検出センサーから記録された32のデータポイントで構成される。各時間ステップにおける出力は、5つの可能なハンドジェスチャー（静止位置、準備、ストローク、保持、後退）のうちの1つを表す。目的は、モーションデータからハンドジェスチャーを検出する分類器を学習することである。

7つの録音をそれぞれ、ちょうど32タイムステップの重複するサブシーケンスに切り分けた。すべてのサブシーケンスをランダムに、重複しないトレーニングセット（75%）、検証セット（10%）、テストセット（15%）に分けた。入力特徴量は、平均がゼロ、標準偏差が単位となるように正規化した。性能指標としてカテゴリ分類精度を用いた。

部屋の占有状況 目的は、温度、湿度、CO2濃度センサーなど、5つの物理センサーストリームから記録された観測によって、部屋が占有されているかどうかを検出することである（Candanedo and Feldheim 2016）。入力データとバイナリ・ラベルは1分間隔でサンプリングされる。

オリジナルのデータセットは、あらかじめ定義されたトレーニングセットとテストセットから構成される。パフォーマンス指標として2値分類精度を用いた。つのセットのそれぞれのシーケンスを、ちょうど32タイムステップの重複するサブシーケンスからなる訓練セットとテストセットに切り分けた。この過程で、テスト・セットからトレーニング・セットにアイテムが漏れることはない。すべてのデータの入力特徴量は、訓練セットの平均と標準偏差によって正規化され、訓練セットの平均と標準偏差がゼロになるようにした。訓練セットの10%を検証セットとして選択する。

人間の活動認識 このタスクでは、ユーザーのスマートフォンの慣性測定値から、歩く、座る、立つなどの人間の活動を認識する（Anguita et al.）データは、6つの可能なカテゴリからなる活動を行う30人のボランティアの記録から構成される。入力変数はフィルタリングされ、各時間ステップで561項目の特徴列を得るために前処理される。

出力変数は、各時間ステップにおける6つの活動カテゴリの1つを表す。性能指標としてカテゴリー分類精度を採用した。元データは既にトレーニングセットとテストセットに分割され、時間フィルタによって前処理されている。加速度センサーとジャイロセンサーのデータは、各タイムステップで合計561の特徴量に変換された。トレーニングセットとテストセットのシーケンスを、ちょうど32タイムステップのオーバーラップするサブシーケンスに整列させた。トレーニングセットの10%を検証セットとして選択する。

シーケンシャルMNIST 我々はMNISTにも取り組んだ。オリジナルのMNISTはコンピュータビジョンの分類問題であるが、我々はこのデータセットをシーケンス分類タスクに変換した。特に、各サンプルは長さ28の28次元時系列として符号化される。さらに、全ての入力特徴量を[0,1]の範囲にダウンスケールする。トレーニングセットの10%を除外し、検証セットとして使用する。

交通量の予測 この実験の目的は、ミネアポリスとセントポールを結ぶ高速道路USインターステート94の1時間ごとの西行き交通量を予測することである。入力特徴量は、気象データと、現地時間や週末、国、地域の祝日の有無を示すフラグなどの日付情報から構成される。出力変数は1時間ごとの交通量である。

元データは、ミネソタ州交通局とOpenWeatherMapから提供された2012年10月から2018年10月までの1時間ごとの記録である。データの7つの列を入力特徴として選択した：1.現在の日が休日であるかどうかを示すフラグ、2.年平均で正規化したケルビン単位の気温、3.降雨量、4.降雪量、5.パーセント単位の雲量、6.現在の日が平日であるかどうかを示すフラグ、7.午前0時の不連続を避けるために正弦関数で事前処理した時刻。出力変数は、平均がゼロ、標準偏差が単位となるように正規化した。平均二乗誤差を学習損失および評価指標として使用した。データを32時間の部分的に重複するシーケンスに分割した。すべてのシーケンスをランダムに非重複トレーニングセット（75%）、検証セット（10%）、テストセット（15%）に分けた。

電力 UCI機械学習リポジトリ（Dua and Graff 2017）の "Individual household electric power consumption Data Set" を使用した。このタスクの目的は、家庭の1時間当たりの有効電力消費量を予測することである。入力特徴は、無効電力引き込みやサブメーターなどの二次測定である。全測定値の約1.25%が欠落しており、同じ特徴量の最新の測定値で上書きする。特徴ごとの白色化を適用し、データセットを32タイムステップの長さの重複しないサブシーケンスに分割する。予測変数（アクティブ消費電力）も白色化する。最適化の損失と評価指標として二乗誤差を使用する。

オゾン日の予測 このタスクの目的は、オゾン日、すなわち、地域のオゾン濃度が限界レベルを超える日を予測

することである。入力機能は、風、天候、日射量からなる。

元のデータセット "Ozone Level Detection Data Set" は、UCIのリポジトリ (Dua and Graff 2017) から取得したもので、テキサス州環境品質委員会 (TCEQ) が収集した毎日のデータポイントで構成されている。6年間の期間を32日間の重複したシーケンスに分割した。少なくとも8時間、オゾンへの曝露が10億分の80を超えた場合、その日はオゾン日とラベル付けされた。入力は、風、気温、日射データを含む73の特徴量からなる。バイナリ予測変数は6.31%の事前分布を持ち、1:15の不均衡を表現する。学習手順では、ラベルに応じて各日のクロスエントロピー損失を重み付けした。オゾンの日を表すラベルは、オゾンでない日の15倍の重みを割り当てた。さらに、標準的な精度ではなく、 F_1 -スコアを報告した (スコアが高いほど良い)。

全サンプルの約27%において、入力特徴の一部が欠落していた。収集したデータの連続性を乱さないために、すべての欠落した特徴をゼロに設定した。このような一部の入力特徴のゼロ化は、非リカレントアプローチや欠損データのフィルタリングと比較して、RNNモデルの性能に悪影響を及ぼす可能性があることに注意してください。その結果、アンサンブル手法やモデルベースのアプローチ、すなわちドメイン知識を活用する手法 (Zhang and Fan 2008) は、実験で研究したエンドツーエンドのRNNを上回ることができる。すべてのサブシーケンスをランダムにトレーニング(75%)、検証(10%)、テスト(15%)セットに分割した。

Person Activity - 1st Setting この設定では、(Rubanova, Chen, and Duvenaud 2019)で説明されている "Human Activity" データセットを使用した。しかし、我々は、訓練-検証-テストの分割に異なるランダムな種を使用し、異なる入力表現を使用するため、我々の結果は、現在の設定で、(Rubanova, Chen, and Duvenaud 2019)によって得られたものに直接移行することはできない。

このデータセットは、人間の様々な身体活動の25の記録から構成されている。

伏せたり、歩いたり、地面に座ったり。参加者には4種類のセンサーが装着され、それぞれ211ミリ秒の周期でサンプリングされた。

Rubanova, Chen, and Duvenaud 2019) と同様に、11の活動カテゴリーを7つのクラスに分類した。入力特徴には正規化を適用しない。25個のシーケンスは、長さ32タイムステップの部分的に重複するサブシーケンスに分割された。

Rubanovaら (Rubanova, Chen, and Duvenaud 2019) と異なり、我々は入力時系列を7次元特徴ベクトルとして表現し、最初の4エントリはセンサーID、最後の3エントリはセンサー値を指定した。サンプリング周波数が高いため、すべてのタイミング情報を破棄した。

結果を表4に報告する。

人物アクティビティ - 2つ目の設定 上記の人物アクティビティタスクと同じデータセットに基づいて、2つ目の実験設定を行った。最初の設定とは対照的に、結果を直接比較できるように、トレーニングセットとテストセットが(Rubanova, Chen, and Duvenaud 2019)と同等であることを確認した。しかし、我々は以前の実験と同じ前処理を適用する。特に、パディングとマスキングを用いて、時間的にも次元的にも不規則にサンプリングされたデータセットを表現し、その結果24次元の入力ベクトルが得られる。一方、時間情報をすべて破棄し、7次元ベクトルとして入力データを投入する。データの形式が異なるだけで、データは同じであることに注意。

(Rubanova, Chen, and Duvenaud 2019)のトレーニング-テスト分割に基づき、トレーニングセットの10%を検証セットとして選択する。さらに、モデルを400エポック訓練し、検証セットで最良の結果を達成したエポックチェックポイントを選択する。このモデルは、(Rubanova, Chen, and Duvenaud 2019)によって提供されたテストセットでテストされる。結果を表5に報告する。

Half-Cheetahキネマティックモデリング このタスクは、RNNがキネマティックダイナミクスのモデリングにどの程度適しているかを評価したChenらの物理シミュレーション実験 (Rubanova, Chen, and Duvenaud 2019) に触発されたものである。私たちの実験では、HalfCheetah-v2ジム環境 (Brockman et al. 2016) 用の事前に訓練されたコントローラの25ロールアウトを収集した。各ロールアウトは、MuJoCo物理エンジン (Todorov, Erez, and Tassa 2012) によって生成された1000個の17次元観測ベクトルの系列で構成されている。タスクは、観測空間の時系列を自己回帰的にフィットさせることである。難易度を上げるために、事前に訓練されたコントローラによって生成されたアクションの5%をランダムなアクションで上書きした。データを2:2:1の比率で訓練、テスト、検証セットに分割した。訓練損失とテスト指標は平均二乗誤差 (MSE) とした。結果は表6に報告されている。

S6ハイパーパラメータとパラメータ・カウント-表3、4、6

表S1: 実験評価に使用したハイパーパラメータ

パラメータ	価値	説明
隠しユニット数	32	入力サンプリング周期に対する
ミニバッチサイズ	16	
学習率	0.001 - 0.02	
ODEソルバーステップ	1/6	
オブティマイザー	アダム (キングマ、バ2014)	
β_1	0.9	アダムのパラメーター
β_2	0.999	
ϵ	1e-08	
BPTTの長さ	32	パラメーター アダムのパラメーター 時間長によるバックプロパゲーション

検証評価間隔	1	時間ステップで x 番目のエポックごとに 指標が評価される
トレーニング・エポック	200	

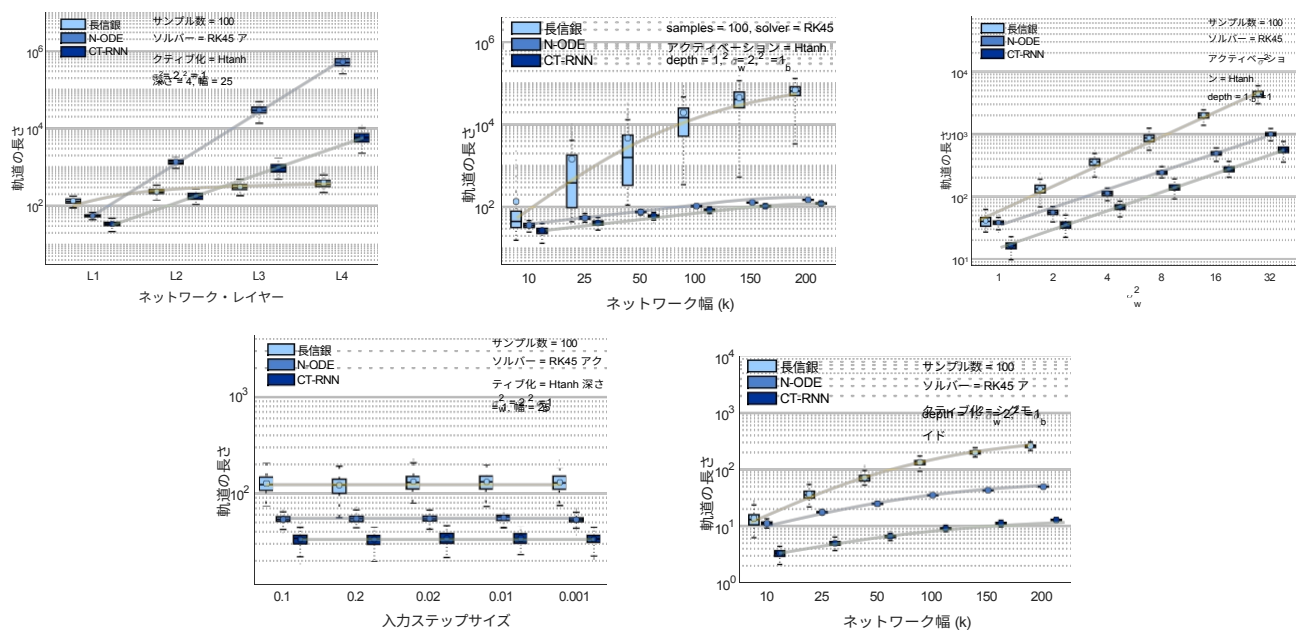
表S2: RNNの幅 k 、隠れ層の数 n 、および減衰スロットの数 m に対する様々なRNNモデルのパラメータの数。

モデル	パラメータ数 (漸近的)	パラメータ数 (正確)
CT-RNN	$O(nk)^2$	$nk^2 + 2nk$
ODE-RNN	$O(nk)^2$	$nk^2 + nk$
エルエステ イーエム	$O(nk)^2$	$4nk^2 + 4nk$
CT-GRU	$O(mk)^2$	$2mk^2 + 2mk + k^2 + k$
長信銀	$O(nk)^2$	$4nk^2 + 3nk$

S7追加の軌道空間表現:

提供された結果の軌跡空間表現は次のサイトで見ることができる: https://www.dropbox.com/s/ly6my34mbvsfi6k/追加 LTC_neurIPS 2020.zip?dl=0

S8 軌道長結果



図S1: 軌跡の長さの追加結果。

S 9Codeとデータの可用性

すべてのコードとデータは、https://github.com/raminmh/liquid_time_constant_networksで公開されている。