Journal of Machine Learning Research 23 (2022) 1-85 Submitted 4/21; Revised 12/21; Published 1/22

Approximation and Optimization Theory for Linear Continuous-Time Recurrent Neural Networks

Zhong Li* LI_ZHONG@PKU.EDU.CN
School of Mathematical Sciences
Peking University
Beijing, China, 100080

Jiequn Han* JIEQUNHAN@GMAIL.COM
Department of Mathematics
Princeton University
Princeton, New Jersey, USA, 08544

Weinan E WEINAN@MATH.PRINCETON.EDU
Department of Mathematics and PACM
Princeton University
Princeton, New Jersey, USA, 08544

Qianxiao Li† QIANXIAO@NUS.EDU.SG
Department of Mathematics
National University of Singapore
Singapore, 119076

Editor: Ohad Shamir

Abstract

We perform a systematic study of the approximation properties and optimization dynamics of recurrent neural networks (RNNs) when applied to learn input-output relationships in temporal data. We consider the simple but representative setting of using continuous-time linear RNNs to learn from data generated by linear relationships. On the approximation side, we prove a direct and an inverse approximation theorem of linear functionals using RNNs, which reveal the intricate connections between memory structures in the target and the corresponding approximation efficiency. In particular, we show that temporal relationships can be effectively approximated by RNNs if and only if the former possesses sufficient memory decay. On the optimization front, we perform detailed analysis of the optimization dynamics, including a precise understanding of the difficulty that may arise in learning relationships with long-term memory. The term "curse of memory" is coined to describe the uncovered phenomena, akin to the "curse of dimension" that plagues high-dimensional function approximation. These results form a relatively complete picture of the interaction of memory and recurrent structures in the linear dynamical setting.

Keywords: recurrent neural networks, dynamical systems, approximation, optimization, curse of memory

∗. Equal contribution
†. Corresponding author

Contents

1. Introduction

Recurrent neural networks (RNNs; Rumelhart et al. (1986)) are among the most frequently employed tools to build machine learning models on temporal data. Despite its ubiquitous application in many domains (Baldi et al., 1999; Graves and Schmidhuber, 2009; Graves,

2013; Graves et al., 2013; Graves and Jaitly, 2014; Gregor et al., 2015), some fundamental theoretical questions remain to be answered. Such questions come in several flavors. First, one may pose the *approximation* problem, which essentially ask what kind of input-output relationships can RNNs model to arbitrary precision. Second, one may also consider the *optimization* problem, which concerns the dynamics of training (say, by gradient descent) the RNN. While such questions can be posed for any machine learning model, the crux of the problem for RNNs is how the recurrent structure of the model and the dynamical nature of the data shape the answers to such problems. For example, it is often claimed that when there are long-term dependencies in the data (Bengio et al., 1994; Hochreiter et al., 2001), then RNN may encounter problems in learning, but such statements have rarely been put on precise mathematical footing.

In this paper, we make a step in this direction by studying the approximation and optimization properties of RNNs. Compared with their feed-forward counterparts, the key distinguishing feature of RNNs is the presence of temporal dynamics in terms of recurrent architectures and the structure of the data. Hence, to understand the influence of dynamics on learning is of fundamental importance. As is often the case, the key effects of dynamics can already be revealed in the simplest setting of linear dynamics. For this reason, we will focus our analysis on linear RNNs, i.e. those with linear recurrent activations. In this case, the RNNs serve to approximate relationships represented by sequences of linear functionals. At the first glance, the setting appears to be simple, but we show that it is very interesting and yields representative results that underlies key differences in the dynamical setting as opposed to static supervised learning problems. In fact, we show that memory, which can be made precise by the decay rates of the target linear functionals, can affect both approximation rates and optimization dynamics in a non-trivial way.

We will employ a continuous-time analysis initially studied in the context of feed-forward architectures (E, 2017; Haber and Ruthotto, 2017; Li et al., 2017; Li and Hao, 2018) and recently in recurrent settings (Ceni et al., 2020; Chang et al., 2019; Lim, 2021; Sherstinsky, 2018; Niu et al., 2019; Herrera et al., 2020; Rubanova et al., 2019) and idealize the RNN as a continuous-time dynamical system that depends on the trainable parameters. This allows us to phrase the problems under investigation in convenient analytical settings that accentuates the effect of dynamics.

The current paper is an expanded version of our conference publication (Li et al., 2021). There, we gave a precise characterization of the approximation rates using RNNs in terms of regularity and memory of the target functional. Moreover, we performed a fine-grained analysis of the optimization dynamics when training linear RNNs, and show that the training efficiency is adversely affected by the presence of long-term memories. We coin the term "curse of memory" to describe these uncovered phenomena. In the current paper, we complete the prior analysis on two fronts. On the approximation side, we prove an inverse approximation theorem, which states that only targets with sufficient memory decay can be efficiently approximated. On the optimization side, we expand upon the dynamical analysis in Li et al. (2021) to also account for the transient training dynamics, before the onset of stalling or plateauing. Together, these form a more complete description of the approximation and optimization properties of RNNs in the linear continuous-time setting.

The rest of the paper is organized as follows. We first introduce our problem setting in Section 3. The main approximation results and their implications are given in Section 4.

The optimization counter part is presented in Section 6. The detailed proofs are found in Section 5 and Section 7 respectively.

Notations. For consistency, we adhere notations used in this paper to the following. Bold-faced letters are reserved for paths, e.g. functions of time. Lower case letters can mean vectors or scalars. Matrices are denoted by capital letters. Superscript with a parenthesis denotes derivatives, i.e. $f^{(k)}(t)$ means $\frac{d^k}{dt^k}f(t)$. For any $n \in \mathbb{N}_+$, write the set $\{1, 2, \cdots, n\}$ by $[n]$. For a set $\mathcal{S}$, $|\mathcal{S}|$ represents its cardinality. For two numbers $x, y \in \mathbb{R}$, we use $x \lesssim y$ to indicate that there exists a universal constant $c_0 > 0$ such that $x \leq c_0 y$, and $x \gtrsim y$ is similarly defined. For any vector $x \in \mathbb{R}^d$, $\mathrm{Diag}(x) \in \mathbb{R}^{d\times d}$ is the diagonal matrix with the elements $x_1, x_2, \cdots, x_d$. For two vectors $x, y \in \mathbb{R}^d$, $x \succ y$ and $x \succeq y$ mean that $x_i > y_i$ and $x_i \geq y_i$ for all $i = 1, 2, \cdots, d$, respectively. We use $\circ$ to represent the common Hadamard product, i.e. $x \circ y = (x_1 y_1, x_2 y_2, \cdots, x_d y_d)$.

2. Related Work

A number of results on RNNs have been obtained in the literature. Concerning the central results in this paper, we mainly discuss on three fronts, namely approximation theory, optimization dynamics and the role of memory in learning. There are many universal approximation results for RNNs, e.g. Matthews (1993); Doya (1993); Schäfer and Zimmermann (2006, 2007) in discrete-time, and Funahashi and Nakamura (1993); Chow and Li (2000); Li et al. (2005); Maass et al. (2007); Nakamura and Nakagawa (2009) in continuous-time. Most of these perform analysis under the regime that the target relationship is generated from some underlying dynamical system (often in the form of difference or differential equations). While for the present work, the formulation of functional approximation and learning is more general. Although the results here are currently limited to the linear setting, it has been already sufficient to reveal new phenomena involving the interaction of learning and dynamics. This point will be apparent in the following, especially when we discuss for approximation rates and optimization dynamics. We also note that the functional/operator approximation has been explored in Chen and Chen (1993); Tianping Chen and Hong Chen (1995); Lu et al. (2019). However, the results therein only investigate the neural networks model and reservoir systems, not for recurrent structures, hence the derived approximation results are similar to random feature models (Gonon et al., 2020). Here we explicitly study the effect of long-term memory in target functionals on approximation using recurrent structures. This is the main difference.

On the optimization side, there are also many recent results concerning the training dynamics of RNNs using gradient-based algorithms, and most of them are positive in the sense that the trainability is proved under specific regimes, including recovering linear dynamical systems (Hardt et al., 2018), or training under over-parameterized settings (Allen-Zhu et al., 2019). Our results here consider the general setting of learning for linear functionals, which need not come from some hidden dynamics (difference/differential equations), and is also away from the over-parameterized regime. In our setting, it is discovered on the contrary that the training can become quite difficult even for this linear case, which can be understood in a quantitative way. It is shown that this difficulty again relates to the long-term memory in target functionals.

The above statements point to the practical literature regarding memory and learning. As is shown later, the dynamical analysis in this work puts the ubiquitous but heuristic observations - that the long-term memory adversely effects training efficiency (Bengio et al., 1994; Hochreiter et al., 2001) - on a concrete theoretical footing, at least under idealized settings. The theoretical analysis here may also serve as a starting point to justify and improve current heuristic methods (Tseng et al., 2016; Dieng et al., 2017; Trinh et al., 2018) developed in applications, in order to handle the difficulty in training with long-term memory. In the meantime, we also complement general results on "vanishing and explosion of gradients" (Pascanu et al., 2013; Hanin and Rolnick, 2018; Hanin, 2018) that are typically restricted to the initialization stage with more complete and precise characterizations under the dynamical regime during the whole training process.

As a supplement, we also provide related work from the time series literature. Although the long-range dependency within temporal data has been studied for a long time, its effect on learning target relationships in the input-output form is rarely covered. For example, the Hurst exponent (Hurst, 1951) is often used as a measure of long-term memory in temporal data, e.g. fractional Brownian motion (Mandelbrot and Ness, 1968). However, it only measures temporal variations and dependence within the input time series itself, which is different from the setting in this paper where memory involves the dependence of the output time series on the input. Nevertheless, motivated by much of the time series literature where the statistical properties and estimation methods of temporal data with long-range dependency are investigated (Samorodnitsky, 2006; Taqqu et al., 1995; Beran, 1992; Doukhan et al., 2002), in practice one can also combine the RNN-like architectures with these classic statistical methodologies to design hybrid models for various applications (Loukas and Öke, 2007; Diaconescu, 2008; Mohan and Gaitonde, 2018; Bukhari et al., 2020).

3. Problem Formulation

The basic problem of supervised learning on time series data is to learn a mapping from an input sequence to an output, which may be a single scalar/vector or also a temporal sequence of such values. Formally, one can think of the output as produced from the input via an unknown function that depends on the entire input sequence, at least up to the time at which the prediction is made. In the discrete-time case, one can write

$$y_k = H_k(x_0, \dots, x_{k-1}), \tag{1}$$

where $\{x_k : k = 0, 1, \dots\}$ and $\{y_k : k = 0, 1, \dots\}$ denote the input and output sequence respectively, and $\{H_k : k = 0, 1, \dots\}$ is a sequence of functions of increasing input dimension accounting for temporal evolution. The goal of supervised learning is to learn an approximation of H_K (single target setting at step K) or $\{H_k : k = 0, \dots, K\}$ (sequence to sequence setting) given observation data.

Recurrent neural networks (RNNs; Rumelhart et al. (1986)) gives a natural way to parameterize such a sequence of functions. In the simplest case, the one-layer RNN is given by

$$\begin{aligned} h_{k+1} &= \sigma(Wh_k + Ux_k), \\ \hat{y}_k &= c^\top h_k. \end{aligned} \tag{2}$$

Here, $\{h_k\}$ are the *hidden states* and its evolution is governed by a feed-forward neural network. Note that we do not include a bias term in the neural network as it can be absorbed into the hidden states. Also, the last output layer can also be nonlinear, but we will consider the simplest linear setting. For each time step k, the mapping from inputs to outputs $\{x_0, \dots, x_{k-1}\} \mapsto \hat{y}_k$ parameterizes a function $\hat{H}_k(\cdot)$ through adjustable parameters (c, W, U). Hence, for a particular choice of these parameters, a sequence of functions $\{\hat{H}_k\}$ is constructed at the same time.

The primary question one can ask is, can $\{\hat{H}_k\}$, through adjusting (c, W, U), approximate any arbitrary sequence of target functions $\{H_k\}$ using the same set of parameters? If so, what structure in the latter makes the approximation process easy or difficult? Another question one can ask is, what is the dynamics of learning (c, W, U) by gradient descent, and what properties of the system affects such dynamics? It is the purpose of this paper to investigate such questions in a precise and systematic manner.

The RNN (2) is not easy to analyze due to its discrete iterative nature. Hence, here we employ a continuous-time idealization that replaces the time-step index k by a continuous time parameter t. The key advantage of this approach is that the previously motivated questions can be investigated under a unified framework, borrowing useful tools from approximation theory, functional analysis and asymptotic analysis. Let us now introduce this framework.

3.1 Continuous-Time Formulation

Now, let us consider a sequence of inputs indexed by a real-valued variable $t \in \mathbb{R}$ instead of a discrete variable k considered previously. We will assume that the input signal is continuous in t, giving a natural input space

$$\mathcal{X} = C_0(\mathbb{R}, \mathbb{R}^d), \tag{3}$$

which is the linear space of continuous functions from $\mathbb{R}$ (time) to $\mathbb{R}^d$ that vanishes at infinity. We will equip $\mathcal{X}$ with the supremum norm

$$\|\boldsymbol{x}\|_{\mathcal{X}} := \sup_{t \in \mathbb{R}} \|x_t\|_\infty. \tag{4}$$

For the space of outputs we will take a scalar time series, i.e. the space of bounded continuous functions from $\mathbb{R}$ to $\mathbb{R}$:

$$\mathcal{Y} = C_b(\mathbb{R}, \mathbb{R}). \tag{5}$$

Vector-valued outputs can be handled by considering each output separately, and will not be explicitly treated in the following analyses. To denote paths without ambiguity, we will hereafter adopt the shorthand $\boldsymbol{x}_{s:t} := \{x_r : r \in [s, t]\}$. Similarly, we write $\boldsymbol{x}_{:s} := \{x_r : -\infty < r \le s\}$ and similarly, $\boldsymbol{x}_{s:} := \{x_r : s \le r < \infty\}$. Finally we write $\boldsymbol{x} := \{x_r : r \in \mathbb{R}\} \in \mathcal{X}$. Similar notations will be used for $\boldsymbol{y} \in \mathcal{Y}$.

To specify the target, we consider a ground truth relationship between inputs $\boldsymbol{x}$ and outputs $\boldsymbol{y}$ as

$$y_t = H_t(\boldsymbol{x}), \tag{6}$$

where for each $t \in \mathbb{R}$, H_t is a functional

$$H_t : \mathcal{X} \to \mathbb{R}. \tag{7}$$

Let us assume for the moment that the family of functionals $\{H_t : t \in \mathbb{R}\}$ satisfies the continuity condition

$$\lim_{\delta \to 0} H_{t+\delta}(\boldsymbol{x}) = H_t(\boldsymbol{x}), \quad \forall t \in \mathbb{R},\ \forall \boldsymbol{x} \in \mathcal{X}. \tag{8}$$

This ensures that $y_t = H_t(\boldsymbol{x})$ is continuous in t and so that $\boldsymbol{y} \in \mathcal{Y}$ as long as the boundedness is satisfied. Later we will show that this is a consequence of other restrictions we may wish to place on the family.

Following the continuous-time viewpoint, we can then define a continuous and residual version of (2) as a hypothesis space to model continuous-time functionals:

$$\begin{aligned} \hat{y}_t &= c^\top h_t, \\ \frac{d}{dt} h_t &= \sigma(W h_t + U x_t), \end{aligned} \tag{9}$$

where each $h_t \in \mathbb{R}^m$ denotes a hidden (latent) state with dimension m, and σ is a point-wise activation function. The dynamics then naturally defines a hypothesis space of sequences of functionals

$$\{\hat{H}_t(\boldsymbol{x}) = \hat{y}_t : t \in \mathbb{R}\}, \tag{10}$$

which can be used to approximate the target functionals $\{H_t\}$ via adjusting (c, W, U).

Remark 1 *It is worth noting that when viewed in this setting, the RNN parameterization of a family of functionals is in some sense a reverse of the Mori-Zwanzig formalism in statistical mechanics (Zwanzig, 2001). In the latter, one passes from a fully observed dynamical system, via introducing memory, to model a closed dynamics involving a subset of relevant observables. For the RNN, the reverse process occurs where one models a input-output relationship with memory by introducing a hidden, but autonomous forced dynamical system. This connection has been pointed out in Ma et al. (2018). Thus, a thorough understanding of the behavior of RNNs may also contribute towards developing practical implementations of the Mori-Zwanzig formalism for physical applications.*

Clearly, the family of functionals the RNN can represent is not arbitrary, and must possess some structure. Let us now introduce some definitions of functionals that makes these structures precise.

The first is the idea of causality, which means that each functional H_t should only depend on the input time sequence up to time t.

Definition 2 (Causal Functionals) *We call H_t a causal functional if it does not depend on the future values of $\boldsymbol{x}$. Concretely, H_t is causal if for every pair of $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}$ such that*

$$x_s = x'_s \text{ for all } s \le t, \tag{11}$$

we must have $H_t(\boldsymbol{x}) = H_t(\boldsymbol{x'})$.

Next, the primary object of study in this paper are (continuous) linear functionals, which is defined as follows.

Definition 3 (Continuous Linear Functionals) *We call H a continuous linear functional if for any $\boldsymbol{x}, \boldsymbol{x'} \in \mathcal{X}$ and $\lambda, \lambda' \in \mathbb{R}$, if*

$$H(\lambda \boldsymbol{x} + \lambda' \boldsymbol{x'}) = \lambda H(\boldsymbol{x}) + \lambda' H(\boldsymbol{x'}) \tag{12}$$

and moreover that

$$\sup_{\boldsymbol{x} \in \mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}} \leq 1} H(\boldsymbol{x}) < \infty, \tag{13}$$

in which case we can define the induced norm as

$$\|H\| := \sup_{\boldsymbol{x} \in \mathcal{X}, \|\boldsymbol{x}\|_{\mathcal{X}} \leq 1} |H(\boldsymbol{x})|, \qquad H \in \mathcal{X}^*. \tag{14}$$

We call that a family $\{H_t\}$ is continuous and linear if each H_t is continuous and linear.

We end with two other properties that functionals able to satisfy, that are especially of relevance to RNNs.

Definition 4 (Regular Functionals) *We call that a functional $H : \mathcal{X} \to \mathbb{R}$ is regular if for any sequence $\{\boldsymbol{x}(n) \in \mathcal{X}, : n \in \mathbb{N}_+\}$ such that $x(n)_t \to 0$ for almost every $t \in \mathbb{R}$ (in the sense of Lebesgue measure), then*

$$\lim_{n \to \infty} H(\boldsymbol{x}(n)) = 0. \tag{15}$$

We call that the family $\{H_t\}$ is regular if each H_t is regular.

Definition 5 (Time-Homogeneous Functionals) *We call that a family of functionals $\{H_t : t \in \mathbb{R}\}$ is time-homogeneous if for every $t, \tau \in \mathbb{R}$, we have*

$$H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}(\tau)), \tag{16}$$

where $x(\tau)_s = x_{s-\tau}$ for all s, i.e. $\boldsymbol{x}(\tau)$ is $\boldsymbol{x}$ whose time index is shifted to the right by τ.

One can think of regular functionals as those that are not determined by values of the inputs on an arbitrarily small time interval, e.g. a thin spike. Time-homogeneous functionals, on the other hand, are those where there is no special reference point in time: if the time index of both the input sequence and the functional are shifted in a coordinated way, then the output value remains the same.

4. The Problem of Approximation and Main Results

In this section we develop an approximation theory of functionals by RNNs. We first introduce the basic approximation setting. In continuous-time, the linear RNN obeys the following dynamics

$$\begin{aligned} \hat{y}_t &= c^\top h_t, \\ \frac{dh_t}{dt} &= W h_t + U x_t. \end{aligned} \tag{17}$$

Notice that in the theoretical setup, the initial time of the system goes back to $-\infty$ with $\lim_{t\to-\infty} x_t = 0$, $\forall \boldsymbol{x} \in \mathcal{X}$, thus by linearity ($H_t(\mathbf{0}) = 0$) we specify the initial condition of the hidden state $h_{-\infty} = 0$ for consistency.[1] In this case, the dynamical system (17) has the following solution

$$\hat{y}_t = \int_0^\infty c^\top e^{Ws} U x_{t-s} ds. \tag{18}$$

We will consider the stable RNNs, where $W \in \mathcal{W}_m$ with

$$\mathcal{W}_m = \{W \in \mathbb{R}^{m\times m} : \text{eigenvalues of } W \text{ have negative real parts}\}. \tag{19}$$

Owing to the representation of solutions in (18), the linear RNN defines a family of functionals

$$\begin{aligned} \hat{\mathcal{H}} &:= \cup_{m\in\mathbb{N}_+} \hat{\mathcal{H}}_m, \\ \hat{\mathcal{H}}_m &:= \Big\{ \{\hat{H}_t : t \in \mathbb{R}\} : \hat{H}_t(\boldsymbol{x}) = \int_0^\infty c^\top e^{Ws} U x_{t-s} ds, \\ &\qquad W \in \mathcal{W}_m, U \in \mathbb{R}^{m\times d}, c \in \mathbb{R}^m \Big\}. \end{aligned} \tag{20}$$

The most basic approximation problem is as follows: given some sequence of target functionals $\{H_t : t \in \mathbb{R}\}$ satisfying appropriate conditions, does there always exist a sequence of RNN functionals $\{\hat{H}_t : t \in \mathbb{R}\}$ in $\hat{\mathcal{H}}$ such that $H_t \approx \hat{H}$ for all $t \in \mathbb{R}$?

We now make an important remark with respect to the current problem formulation that differs from previous investigations in RNN approximation: we are *not* assuming that the target functionals $\{H_t : t \in \mathbb{R}\}$ are themselves generated from an underlying dynamical system. In other words, there may be no dynamical systems satisfying

$$H_t(\boldsymbol{x}) = y_t, \qquad \text{where} \quad \begin{aligned} y_t &= g(h_t), \\ \frac{d}{dt} h_t &= f(h_t, x_t) \end{aligned} \tag{21}$$

for any linear or nonlinear functions f, g. This sets apart our current setting with previous work on approximation theory of RNNs (Matthews, 1993; Nakamura and Nakagawa, 2009; Chow and Li, 2000; Li et al., 2005; Schäfer and Zimmermann, 2006, 2007; Funahashi and Nakamura, 1993), and also RNN training dynamics (Hardt et al., 2018), where it is assumed that the sequence of target functionals are indeed generated from some unknown dynamical system. In this setting, the approximation problem reduces to the approximation of the functions f, g of the underlying dynamical system by neural networks, and the obtained results often resemble those in feed-forward neural networks.

In our case, however, we consider general input-output relationships related by temporal sequences of functionals, with no necessary recourse to the mechanism from which these relationships are generated. This is an important distinction, for often in RNN applications, the time-series data may not be generated from some partially-observed Markovian process.

1. In application frameworks such as TensorFlow and PyTorch, the initial hidden state is set to zero by default.

Hence, this setting is more general, and natural for applications. Moreover, notice that in the linear case, if the target functionals $\{H_t\}$ are generated from a linear dynamical system, then the approximation question is trivial: as long as the dimension of h_t in the approximating RNN is greater than or equal to that which generates the target, we have perfect approximation. However, we will see that in the more general consideration of approximation a sequence of target functionals, this question becomes much more interesting, even in the linear regime. In fact, we will now prove precise approximation theories and characterize approximation rates that reveal intricate connections with memory effects, which may be otherwise obscured if one considers more limited settings of recovering hidden dynamical systems.

4.1 Universal Approximation Theorem

First, it is clear that the functionals in RNN hypothesis $\hat{\mathcal{H}}$ space must possess some structure, which motivated the introduction of various classes of functionals in Section 3.1. The following observation can be verified directly and its proof is immediate and hence omitted.

Proposition 6 *Let $\{\hat{H}_t : t \in \mathbb{R}\}$ be any family of functionals in $\hat{\mathcal{H}}$ (see (20)) resulting from the linear RNN dynamics (9). Then for each $t \in \mathbb{R}$,*

1. *$\hat{H}_t$ is a continuous, linear functional.*
2. *$\hat{H}_t$ is a causal functional.*
3. *$\hat{H}_t$ is a regular functional.*
4. *The family $\{\hat{H}_t\}$ is time-homogeneous.*

Our first main result of approximation is in some sense a converse of Proposition 6. In particular, we prove the following approximation theorem, which states that *any* sequence of functionals satisfying the properties in Proposition 6 can be approximated uniformly by sequences of RNN functionals in $\hat{\mathcal{H}}$ to arbitrary accuracy.

Theorem 7 (Universal Approximation for Linear RNNs) *Let $\{H_t : t \in \mathbb{R}\}$ be a family of continuous, linear, causal, regular and time-homogeneous functionals on $\mathcal{X}$. Then, for any $\epsilon > 0$ there exists $\{\hat{H}_t : t \in \mathbb{R}\} \in \hat{\mathcal{H}}$ such that*

$$\sup_{t\in\mathbb{R}} \|H_t - \hat{H}_t\| \equiv \sup_{t\in\mathbb{R}} \sup_{\|\boldsymbol{x}\|_{\mathcal{X}}\le 1} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \le \epsilon. \tag{22}$$

The proof relies on the classical Riesz-Markov-Kakutani representation theorem, which states that each linear functional H_t can be uniquely associated with a signed measure μ_t such that $H_t(\boldsymbol{x}) = \int_{\mathbb{R}} x_s^\top d\mu_t(s)$. Owing to the assumptions of Theorem 7, we can further show that the sequence of representations $\{\mu_t\}$ are related to an integrable function $\rho : [0, \infty) \to \mathbb{R}^d$, such that $\{H_t\}$ admits the common representation

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \qquad t \in \mathbb{R}, \quad \boldsymbol{x} \in \mathcal{X}. \tag{23}$$

Comparing this representation with the solution (18) of the linear continuous-time RNN, we find that the approximation property of RNNs is closely related to how well $\rho(t)$ can

be approximated by the exponential sums of the form $(c^\top e^{Wt}U)^\top$. That is to say, the functional approximation is then reduced to the function approximation in the sense of representations. Intuitively, (23) shows that each output $y_t = H_t(\boldsymbol{x})$ is simply a convolution between the input signal and the kernel ρ. Thus, the smoothness and decay of the input-output relationship is characterized by the convolution kernel ρ. Due to this observation, we will hereafter refer to $\{H_t\}$ and ρ interchangeably.

Remark 8 *Theorem 7 can be extended in several ways. Without the assumption of causality, we can use bidirectional recurrent neural networks (Schuster and Paliwal, 1997) to achieve the universal approximation. Without the assumption of time-homogeneity, we can introduce another coordinate to act as time. Without regularity assumption on $\{H_t\}$, we would not have uniform error estimate (i.e. $\sup_t$), in which case we can replace it with some L^p-estimate in time.*

Remark 9 *In the literature, there are in fact many results on the approximation properties of RNNs in discrete-time (Schäfer and Zimmermann, 2006, 2007; Matthews, 1993) and continuous-time (Funahashi and Nakamura, 1993; Nakamura and Nakagawa, 2009; Li et al., 2005; Chow and Li, 2000). However, as discussed before, most of these focus on the case where the target relationship is generated from dynamical systems. The functional approximation formulation considered here is more general, and reveals new phenomena that may not be discovered from these approaches. This will be especially apparent in the next section when it comes to approximation rates. We also note that functional/operator approximation using neural networks has been explored in Chen and Chen (1993); Tianping Chen and Hong Chen (1995); Lu et al. (2019) for non-recurrent structures and reservoir systems for which approximation results similar to random feature models are derived (Gonon et al., 2020).*

4.2 Approximation Rates and Inverse Approximation Theorem

While the previous result establishes the universal approximation property of linear RNNs for suitable classes of linear functionals, it does not reveal to us which functionals can be efficiently approximated. In the practical literature, it is often observed that when there is some long-term memory in the inputs and the outputs, the RNN becomes quite ill-behaved (Bengio et al., 1994; Hochreiter et al., 2001). It is the purpose of this section to establish results which make these heuristics statements precise. In particular, we will show that the rate at which linear functionals can be approximated by RNNs depends on the smoothness and memory properties of the former. We note that this is a much less explored area in the approximation theory of RNNs.

To characterize smoothness and decay of functionals, we may pass to investigating the properties of their actions on constant input signals. Concretely, let us denote by $e_i\,(i = 1, \ldots, d)$ the standard basis vector in $\mathbb{R}^d$, and $\boldsymbol{e}_i$ denotes a constant signal with $e_{i,t} = e_i 1_{\{t\geq 0\}}$ for all t. Then

1. smoothness is characterized by the smoothness of the maps $t \mapsto H_t(\boldsymbol{e}_i)$, $i = 1, \ldots, d$;

2. memory is characterized by the decay rate of the maps $t \mapsto H_t(\boldsymbol{e}_i)$, $i = 1, \ldots, d$.

Our second main result of approximation shows that these two properties are intimately tied with the approximation rate.

Theorem 10 (Approximation Rates of Linear RNNs) *Assuming the conditions as in Theorem 7. Consider the output of constant signal*

$$y_i(t) = H_t(\boldsymbol{e}_i), \quad i = 1, \dots, d. \tag{24}$$

Suppose there exist constants $\alpha \in \mathbb{N}_+, \beta, \gamma > 0$ such that for $i = 1, \dots, d$, $y_i(t) \in C^{(\alpha+1)}(\mathbb{R})$ and for $k = 1, \dots, \alpha + 1$,

$$e^{\beta t} y_i^{(k)}(t) = o(1) \text{ as } t \to +\infty, \tag{25}$$

$$\sup_{t \geq 0} \frac{|e^{\beta t} y_i^{(k)}(t)|}{\beta^k} \leq \gamma. \tag{26}$$

Then, there exists a universal constant $C(\alpha) > 0$ only depending on α such that for any $m \in \mathbb{N}_+$, there exists a sequence of width-m RNN functionals $\{\hat{H}_t : t \in \mathbb{R}\} \in \hat{\mathcal{H}}_m$ such that

$$\sup_{t \in \mathbb{R}} \|H_t - \hat{H}_t\| \equiv \sup_{t \in \mathbb{R}} \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \leq 1} |H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \leq \frac{C(\alpha)\gamma d}{\beta m^{\alpha}}. \tag{27}$$

Theorem 10 can be treated as a direct approximation theorem for linear functionals. The classical direct approximation theorems in approximation theory, also known as Jackson-type theorems (Jackson, 1930), provide an approximation rate of a function in terms of its smoothness properties. Following a similar spirit, Theorem 10 gives an approximation rate of a linear functional in terms of its smoothness and decay properties when using linear RNNs. Accordingly, we can also consider inverse approximation theorems, also known as Bernstein-type theorems (Bernstein, 1920), whose classical forms characterize the smoothness of target functions that can be efficiently approximated by classes of simple functions, e.g. polynomials. Our next theorem gives such a result in the context of approximating linear functionals by recurrent neural networks. In particular, we show that if a target functional can be effectively approximated by RNNs, then it must have exponentially decaying memory.

Theorem 11 (Inverse Approximation Theorem on Exponential Decay) *Assume the conditions as in Theorem 7 and consider the output of constant signal*

$$y_i(t) = H_t(\boldsymbol{e}_i) \in C^{(\alpha+1)}(\mathbb{R}), \quad i = 1, \dots, d, \ \alpha \in \mathbb{N}_+. \tag{28}$$

Suppose for each $m \in \mathbb{N}_+$, there exists a sequence of width-m RNN functionals $\{\hat{H}_t : t \in \mathbb{R}\} \in \hat{\mathcal{H}}_m$ approximating H_t in the following sense

$$\lim_{m \to \infty} \sup_{t \geq 0} |\hat{y}_{i,m}^{(k)}(t) - y_i^{(k)}(t)| = 0, \quad i = 1, \dots, d, \ k = 1, \dots, \alpha + 1, \tag{29}$$

where

$$\hat{y}_{i,m}(t) = \hat{H}_t(\boldsymbol{e}_i), \quad i = 1, \dots, d. \tag{30}$$

Define $w_m = \max_{j\in[m]} Re(\lambda_j)$*, where* $\lambda_j, j = 1, \dots, m$ *are the eigenvalues of* W *in* $\{\hat{H}_t : t \in \mathbb{R}\}$*. Assume the parameters in RNNs are uniformly bounded and there exists a constant* $\beta > 0$ *such that* $\limsup_{m\to\infty} w_m < -\beta$*, then we have*

$$e^{\beta t} y_i^{(k)}(t) = o(1) \text{ as } t \to +\infty, \quad i = 1, \dots, d, \ k = 1, \dots, \alpha + 1. \tag{31}$$

The direct and inverse approximation theorems paints a picture of what functionals are amenable to efficient approximation through linear RNNs: those, and only those, with exponentially decaying memory structures. The next section discusses what occurs when such a decay pattern is not present.

4.3 The Curse of Memory in Approximation

For approximation of nonlinear functions using linear combinations of basis functions, one often suffers from the "curse of dimensionality" (Bellman, 1957), in that the number of basis functions required to achieve a certain approximation accuracy increases exponentially when the dimension d of the input space increases. In the case of Theorem 10, the bound scales linearly with d (see (27)). This is because the target functional possesses a linear structure, and hence each dimension can be approximated independently of others, resulting in an additive error estimate. Nevertheless, due to the presence of the temporal dimension, there enters another type of challenge, which we coin the *curse of memory*. Let us now discuss this point in detail.

We assume $d = 1$ and drop subscripts for simplicity. By (23) and (24), we get

$$y(t) = H_t(\mathbf{1}_{\{s\geq 0\}}) = \int_0^t \rho(s)ds, \quad t \geq 0. \tag{32}$$

Consider the example $\rho(t) \in C^{(1)}(\mathbb{R})$ and

$$\rho(t) \sim t^{-(1+\omega)} \text{ as } t \to +\infty. \tag{33}$$

Here $\omega > 0$ indicates the decay rate of the memory effects in our target functional family $\{H_t\}$. The smaller its value, the slower the decay and the longer the system memory. Notice that $y^{(1)}(t) = \rho(t)$ and in this case there exists no $\beta > 0$ making (25) true, and no rate estimate can be deduced from it.

A natural way to circumvent this obstacle is to introduce a truncation in time. With $T \gg 1$ we can define $\tilde{\rho}(t) \in C^{(1)}(\mathbb{R})$, such that $\tilde{\rho}(t) \equiv \rho(t)$ for $t \leq T$, $\tilde{\rho}(t) \equiv 0$ for $t \geq T + 1$, and $\tilde{\rho}(t)$ is monotonically decreasing for $T \leq t \leq T + 1$. Considering the linear functional

$$\tilde{H}_t(\boldsymbol{x}) := \int_0^t x_{t-s}\tilde{\rho}(s)ds, \tag{34}$$

we have the truncation error estimate

$$|H_t(\boldsymbol{x}) - \tilde{H}_t(\boldsymbol{x})| \leq \|\boldsymbol{x}\|_{\mathcal{X}} \left(\int_T^\infty |\rho(s)|ds \right) \sim \|\boldsymbol{x}\|_{\mathcal{X}} T^{-\omega}. \tag{35}$$

Now Theorem 10 is applicable to the truncated $\{\tilde{H}_t\}$ with $\alpha = 1$, and we have for any $\beta > 0$, there is a linear RNN (i.e. there exists parameters (c, W, U)) such that the associated functionals $\{\hat{H}_t\} \in \hat{\mathcal{H}}_m$ satisfy

$$\sup_{t\in\mathbb{R}} \|\tilde{H}_t - \hat{H}_t\| \leq \frac{C\gamma}{\beta m} := \frac{C}{\beta m} \sup_{t\geq 0} \frac{|e^{\beta t} y^{(1)}(t)|}{\beta} = \frac{C\omega}{m} \frac{e^{\beta T}}{\beta^2 T^{\omega+1}}. \tag{36}$$

It is straightforward to verify that when $\beta = 2/T$, the right-hand side of (36) achieves the minimum, which gives

$$\sup_{t\in\mathbb{R}} \|\tilde{H}_t - \hat{H}_t\| \leq \frac{C\omega}{m} T^{1-\omega}. \tag{37}$$

Combining (35) and (37) gives

$$\sup_{t\in\mathbb{R}} \|H_t - \hat{H}_t\| \leq C \left(T^{-\omega} + \frac{\omega}{m} T^{1-\omega} \right). \tag{38}$$

In order to achieve an error tolerance ϵ, we require $T \sim \epsilon^{-\frac{1}{\omega}}$ according to the first term above, and then according to second term, we have

$$m = \mathcal{O}\left(\frac{\omega T^{1-\omega}}{\epsilon} \right) = \mathcal{O}\left(\omega \epsilon^{-\frac{1}{\omega}} \right). \tag{39}$$

This estimate gives us a quantitative relationship between the degree of freedom needed and the decay speed. When $\omega > 0$ is small, i.e. the system has long-term memory, the size of the RNN model required grows exponentially. This is akin to the curse of dimensionality, but this time on memory, which manifests itself even in the simplest linear settings.

Remark 12 *Here, the curse of memory is on approximation properties, in that functionals with long-term memory are hard to approximate by the RNN architecture. This is unrelated to the commonly quoted idea of "vanishing and explosion of gradients" (Pascanu et al., 2013; Hanin and Rolnick, 2018; Hanin, 2018) that plagues RNNs (and indeed many deep architectures). In fact, the curse of memory for approximation is inherent in the architecture itself, without reference to any training algorithm. At the same time, this is also specific to RNNs when viewed as approximators of functionals.*

Remark 13 *The result here also highlights the importance of considering the approximation of general sequences of functionals, instead of those generated by underlying dynamical systems. In the latter case, approximation theory reduces to the function approximation regime of feed-forward neural networks, and the approximation rates one obtains may not capture the dynamical aspect of the problem and reveal the curse of memory associated.*

5. Proofs of Approximation Results

We first present the proof of Theorem 7. A key simplification of considering linear functionals is due to the classical representation result below, which allows us to pass from the approximation of functionals to the approximation of functions.

Theorem 14 (Riesz-Markov-Kakutani Representation Theorem) *Let $H : \mathcal{X} \to \mathbb{R}$ be a continuous linear functional. Then, there exists a unique, vector-valued, regular, countably additive signed measure μ on $\mathbb{R}$ such that*

$$H(\boldsymbol{x}) = \int_{\mathbb{R}} x_s^\top d\mu(s) = \sum_{i=1}^{d} \int_{\mathbb{R}} x_{s,i} d\mu_i(s). \tag{40}$$

Moreover, we have

$$\|H\| := \sup_{\|\boldsymbol{x}\|_{\mathcal{X}} \leq 1} |H(\boldsymbol{x})| = \|\mu\|_1(\mathbb{R}) := \sum_i |\mu_i|(\mathbb{R}). \tag{41}$$

Proof Well-known. See e.g. Bogachev (2007), CH 7.10.4. ■

We will use the representation theorem to prove Theorem 7. First, we prove some lemmas.

Lemma 15 *Let $\{H_t\}$ be a family of continuous, linear, regular, causal and time-homogeneous functionals on $\mathcal{X}$. Then, there exists a measurable function $\rho : [0, \infty) \to \mathbb{R}^d$ that is integrable, i.e.*

$$\|\rho\|_{L^1([0,\infty))} := \sum_{i=1}^{d} \int_0^\infty |\rho_i(s)| ds < \infty \tag{42}$$

and

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \quad t \in \mathbb{R}. \tag{43}$$

In particular, $\{H_t\}$ is uniformly bounded with $\sup_t \|H_t\| = \|\rho\|_{L^1([0,\infty))}$ and $t \mapsto H_t(\boldsymbol{x})$ is continuous for all $\boldsymbol{x} \in \mathcal{X}$.

Proof By the Riesz-Markov-Kakutani representation theorem (Theorem 14), for each t there is a unique regular signed Borel measure μ_t such that

$$H_t(\boldsymbol{x}) = \int_{\mathbb{R}} x_s^\top d\mu_t(s), \tag{44}$$

and $\sum_i |\mu_{t,i}|(\mathbb{R}) = \|H_t\|$. Since $\{H_t\}$ is causal, we must have $\int_t^\infty x_s^\top d\mu_t(s) = 0$ for any $\boldsymbol{x}$, thus

$$H_t(\boldsymbol{x}) = \int_{-\infty}^{t} x_s^\top d\mu_t(s). \tag{45}$$

Now, by time homogeneity we have

$$\int_{-\infty}^{t} x_s^\top d\mu_t(s) = H_t(\boldsymbol{x}) = H_{t+\tau}(\boldsymbol{x}^{(\tau)}) = \int_{-\infty}^{t+\tau} x_{s-\tau}^\top d\mu_{t+\tau}(s). \tag{46}$$

Take $\tau = -t$ and set $\mu = -\mu_0$ to get

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top d\mu(s). \tag{47}$$

Note that we have $\|\mu\|_1([0,\infty)) = \|\mu_0\|_1([0,\infty)) = \|H_0\| = \|H_t\|$, and continuity follows from the fact that

$$\begin{aligned} |H_{t+\delta}(\boldsymbol{x}) - H_t(\boldsymbol{x})| &= \left| \int_0^\infty (x_{t+\delta-s} - x_{t-s})^\top d\mu(s) \right| \\ &\le \sum_i \int_0^\infty \|x_{t+\delta-s} - x_{t-s}\|_\infty d|\mu_i|(s), \end{aligned} \tag{48}$$

which converges to 0 as $\delta \to 0$ by Lebesgue's dominated convergence theorem. Finally, we will show that each μ_i is absolutely continuous with respect to λ (Lebesgue measure). Take a measurable $E \subset [0,\infty)$ such that $\lambda(E) = 0$ and set $E' = [0,\infty) \setminus E$. For each $n \in \mathbb{N}_+$, set $K_n \subset E, K_n' \subset E'$ where K_n, K_n' are closed and $\mu_i(E \setminus K_n) \le 1/n$, $\mu_i(E' \setminus K_n') \le 1/n$. For a fixed $i \in [d]$, define $\boldsymbol{x}^{(n)}$ to be such that $x_{t-s,j}^{(n)} = 0$ for all $j \ne i$ and all s. For $j = i$, we set $x_{t-s,i}^{(n)} = 1$ if $s \in K_n$ and 0 if $s \in K_n'$, which can then be continuously extended to $[0,\infty)$. Observe that by construction, $x_{t-s}^{(n)} \to 0$ for λ-a.e. s, thus by Lebesgue's dominated convergence theorem

$$0 = \lim_{n\to\infty} H_t(\boldsymbol{x}^{(n)}) = \mu_i(E). \tag{49}$$

This shows that μ_i is absolutely continuous with respect to λ, and by the Radon-Nikodym theorem, there exists a measurable function $\rho_i : [0,\infty) \to \mathbb{R}$ such that for any measurable $A \subset \mathbb{R}$, we have

$$\int_A d\mu_i(s) = \int_A \rho_i(s) ds, \quad i = 1, \ldots, d. \tag{50}$$

Hence, we get

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \tag{51}$$

with $\|\rho\|_{L^1([0,\infty))} = \sum_i \int_0^\infty |\rho_i(s)| ds = \|\mu\|_1([0,\infty)) < \infty$. The proof is completed. ■

Lemma 16 *Let $\rho : [0,\infty) \to \mathbb{R}$ be a Lebesgue integrable function, i.e. $\|\rho\|_{L^1([0,\infty))} < \infty$. Then, for any $\epsilon > 0$, there exists a polynomial p with $p(0) = 0$ such that*

$$\left\| \rho - p(e^{-\cdot}) \right\|_{L^1([0,\infty))} = \int_0^\infty |\rho(t) - p(e^{-t})| dt \le \epsilon. \tag{52}$$

Proof The approach here is similar to that of the approximation of functions using exponential sums (Kammler, 1976; Braess, 1986). Alternatively, one may also appeal to the

density of phase type distributions (He and Zhang, 2007; O'Cinneide, 1990) in the space of positive distributions, and generalizing them to signed measures.

Fix any $\epsilon > 0$. Define

$$R(u) = \begin{cases} \frac{1}{u}\rho(-\log u), & u \in (0,1], \\ 0, & u = 0. \end{cases} \tag{53}$$

Then, we can check that

$$\|R\|_{L^1([0,1])} = \|\rho\|_{L^1([0,\infty))} < \infty. \tag{54}$$

By density of continuous functions in L^1, there exists a continuous function $\tilde{R}$ on $[0,1]$ with $\tilde{R}(0) = 0$ such that

$$\|R - \tilde{R}\|_{L^1([0,1])} \leq \epsilon/2. \tag{55}$$

By Müntz-Szász theorem (Müntz, 1914; Szász, 1916), there exists a polynomial p with $p(0) = 0$ such that

$$\|q - \tilde{R}\|_{L^1([0,1])} \leq \epsilon/2, \tag{56}$$

and $q(u) := p(u)/u$ is also a polynomial. Therefore, we have

$$\begin{aligned} \left\|\rho - p(e^{-\cdot})\right\|_{L^1([0,\infty))} &= \int_0^1 |R(u) - p(u)/u| du \\ &\leq \int_0^1 |R(u) - \tilde{R}(u)| du + \int_0^1 |\tilde{R}(u) - p(u)/u| du \leq \epsilon. \end{aligned} \tag{57}$$

The proof is completed. ■

We are now ready to present the proofs of Theorem 7, Theorem 10 and Theorem 11.

Proof [Proof of Theorem 7] By (18), for each $\{\hat{H}_t\} \in \hat{\mathcal{H}}$ we can write

$$\hat{H}_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top (U^\top [e^{Ws}]^\top c) ds. \tag{58}$$

By Lemma 15, we have

$$H_t(\boldsymbol{x}) = \int_0^\infty x_{t-s}^\top \rho(s) ds, \tag{59}$$

where ρ is integrable. Thus, we can apply Lemma 16 to conclude that there exists polynomials $p_i, i = 1, \ldots, d$ with $p_i(0) = 0$, such that

$$\sum_i \|\rho_i - p_i(e^{-\cdot})\|_{L^1([0,\infty))} \leq \epsilon. \tag{60}$$

Notice that we can write each $p_i(u) = \sum_{j=1}^m \alpha_{ij} u^j$ for some $m \in \mathbb{N}_+$ equaling the maximal degree of $\{p_i\}_{i=1}^d$. Taking $W = \mathrm{diag}(-1, \ldots, -m)$, $c = (1, \ldots, 1)$ and $U_{ij} = \alpha_{ji}$, we get

$$(U^\top [e^{Ws}]^\top c)_i = p_i(e^{-s}), \quad i = 1, \ldots, d. \tag{61}$$

Consequently, we have for any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_\infty \le 1$,

$$\begin{aligned}|H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| =& \left|\int_0^\infty x_{t-s}^\top \rho(s) ds - \int_0^\infty x_{t-s}^\top p(e^{-s}) ds\right| \\ \le& \sum_i \int_0^\infty |x_{t-s,i}| \left|\rho_i(s) - p_i(e^{-s})\right| ds \le \sum_i \|\rho_i - p_i(e^{-\cdot})\|_{L^1([0,\infty))} \\ \le& \epsilon.\end{aligned} \tag{62}$$

The proof is completed. ■

Proof [Proof of Theorem 10] We fix $i \in [d]$ below until the last part of the proof. By Lemma 15, there exists $\rho_i(t) \in C^\alpha[0,\infty)$ such that

$$y_i(t) = H_t(\boldsymbol{e}_i) = \int_0^t \rho_i(r) dr, \quad t \ge 0. \tag{63}$$

By the assumption,

$$\rho_i^{(k)}(t) = o(e^{-\beta t}) \text{ as } t \to \infty, \quad k = 0, \dots, \alpha. \tag{64}$$

Consider the transform

$$q_i(s) = \begin{cases} 0, & s = 0, \\ \frac{\rho_i\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s}, & s \in (0,1]. \end{cases} \tag{65}$$

For any $k \in \{0, 1, \dots, \alpha\}$, one can prove by induction that

$$q_i^{(k)}(s) = (-1)^k \sum_{j=0}^k c(j,k) \left(\frac{\alpha+1}{\beta}\right)^j \frac{\rho_i^{(j)}\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s^{k+1}}, \tag{66}$$

where $c(j,k)$ are some integer constants. Together with the assumption, we have

$$\left|q_i^{(k)}(e^{-\frac{\beta}{\alpha+1}t})\right| = \left|\sum_{j=0}^k c(j,k) \left(\frac{\alpha+1}{\beta}\right)^j \frac{\rho_i^{(j)}(t)}{e^{-\frac{(k+1)\beta}{\alpha+1}t}}\right| \le \sum_{j=0}^k \left|c(j,k)|(\alpha+1)^j\right| \gamma \le C(\alpha)\gamma, \tag{67}$$

where $C(\alpha) > 0$ is a universal constant only depending on α. Note that for $j = 0, 1, \dots, \alpha$,

$$\lim_{s \to 0^+} \frac{\rho_i^{(j)}\left(\frac{-(\alpha+1)\log s}{\beta}\right)}{s^{k+1}} = \lim_{t \to \infty} \frac{\rho_i^{(j)}(t)}{e^{-\frac{(k+1)\beta}{\alpha+1}t}} = \lim_{t \to \infty} \frac{\rho_i^{(j)}(t)}{e^{-\beta t}} e^{-\frac{(\alpha-k)\beta}{\alpha+1}t} = 0, \tag{68}$$

hence $q_i(s) \in C^\alpha[0,1]$ with $q_i(0) = q_i^{(1)}(0) = \cdots = q_i^{(\alpha)}(0) = 0$. By Jackson's theorem (Jackson, 1930), for any $m \in \mathbb{N}_+$, there exists a polynomial $Q_{i,m}$ of degree $m-1$ such that

$$\|q_i - Q_{i,m}\|_{L^\infty([0,1])} \le \frac{C(\alpha)\gamma}{m^\alpha}. \tag{69}$$

Denote the polynomial $Q_{i,m}$ as

$$Q_{i,m}(s) = \sum_{j=0}^{m-1} \alpha_{i,j} s^j, \tag{70}$$

and define

$$\phi_{i,m}(t) = e^{-\frac{\beta}{\alpha+1}t} Q_{i,m}(e^{-\frac{\beta}{\alpha+1}t}). \tag{71}$$

Then we have

$$\phi_{i,m}(t) = c^\top e^{Wt} u_i, \tag{72}$$

where

$$c = (1, 1, \ldots, 1), \tag{73}$$

$$W = \begin{bmatrix} -\frac{\beta}{\alpha+1} & & & \\ & -\frac{2\beta}{\alpha+1} & & \\ & & \ddots & \\ & & & -\frac{m\beta}{\alpha+1} \end{bmatrix}, \tag{74}$$

$$u_i = (\alpha_{i,0}, \alpha_{i,1}, \ldots, \alpha_{i,m-1}). \tag{75}$$

With the change of variables $s = e^{-\frac{\beta}{\alpha+1}t}$, we have the estimate

$$\begin{aligned} \|\rho_i - \phi_{i,m}\|_{L^1([0,\infty))} &= \int_0^\infty |\rho_i(t) - \phi_{i,m}(t)| dt \\ &= \int_0^1 \left| \rho_i\left(\frac{-(\alpha+1)\log s}{\beta}\right) - s Q_{i,m}(s) \right| \frac{\alpha+1}{\beta s} ds \\ &= \frac{\alpha+1}{\beta} \int_0^1 |q_i(s) - Q_{i,m}(s)| ds \\ &\leq \frac{C(\alpha)\gamma}{\beta m^\alpha}. \end{aligned} \tag{76}$$

Finally, define $U = [u_1, \ldots, u_d] \in \mathbb{R}^{m \times d}$, we have

$$c^\top e^{Wt} U = (\phi_{1,m}(t), \ldots, \phi_{d,m}(t)). \tag{77}$$

The parameters (c, W, U) together determine the dynamical system (9). Similar to the arguments in the proof of Theorem 7, for any $\boldsymbol{x}$ with $\|\boldsymbol{x}\|_\infty \leq 1$ and t, we have

$$|H_t(\boldsymbol{x}) - \hat{H}_t(\boldsymbol{x})| \leq \sum_i \|\rho_i - \phi_{i,m}\|_{L^1([0,\infty))} \leq \frac{C(\alpha)\gamma d}{\beta m^\alpha}. \tag{78}$$

The proof is completed. ■

Proof [Proof of Theorem 11] For each $m \in \mathbb{N}_+$, by (18) and the assumption, we know there exist $c \in \mathbb{R}^m, W \in \mathbb{R}^{m\times m}, U = [u_1, \ldots, u_d] \in \mathbb{R}^{m\times d}$ such that

$$\hat{y}_{i,m}(t) = \int_0^t c^\top e^{Ws} u_i ds, \quad i = 1, \ldots, d. \tag{79}$$

Accordingly, for $k = 1, \ldots, \alpha + 1$,

$$\hat{y}^{(k)}_{i,m}(t) = c^\top W^{k-1} e^{Wt} u_i, \quad i = 1, \ldots, d. \tag{80}$$

Given a function $f \in C[0, \infty)$, we again consider the transformation $\mathcal{T}f : [0, 1] \mapsto \mathbb{R}$ defined as

$$(\mathcal{T}f)(s) = \begin{cases} 0, & s = 0, \\ \frac{f\left(\frac{-\log s}{\beta}\right)}{s}, & s \in (0, 1]. \end{cases} \tag{81}$$

Under the change of variables $s = e^{-\beta t}$, we have

$$f(t) = e^{-\beta t}(\mathcal{T}f)(e^{-\beta t}), \quad t \geq 0. \tag{82}$$

By the assumption on the uniform bound of the eigenvalues of W, we have

$$\lim_{s\to 0^+} (\mathcal{T}\hat{y}^{(k)}_{i,m})(s) = \lim_{t\to\infty} \frac{\hat{y}^{(k)}_{i,m}(t)}{e^{-\beta t}} = 0, \tag{83}$$

which implies $\mathcal{T}\hat{y}^{(k)}_{i,m} \in C([0, 1])$. Let $\delta = -\beta - \limsup_{m\to\infty} w_m$. By the assumption, we have $\delta > 0$, and

$$\begin{aligned} &\sup_{s\in[0,1]} \left|(\mathcal{T}\hat{y}^{(k)}_{i,m_1})(s) - (\mathcal{T}\hat{y}^{(k)}_{i,m_2})(s)\right| \\ &= \sup_{t\geq 0} \left|\frac{\hat{y}^{(k)}_{i,m_1}(t)}{e^{-\beta t}} - \frac{\hat{y}^{(k)}_{i,m_2}(t)}{e^{-\beta t}}\right| \\ &\leq \max\left\{ \sup_{0\leq t\leq T_0} \left|\frac{\hat{y}^{(k)}_{i,m_1}(t)}{e^{-\beta t}} - \frac{\hat{y}^{(k)}_{i,m_2}(t)}{e^{-\beta t}}\right|, c_0 e^{-\delta T_0} \right\} \\ &\leq \max\left\{ e^{\beta T_0} \sup_{0\leq t\leq T_0} \left|\hat{y}^{(k)}_{i,m_1}(t) - \hat{y}^{(k)}_{i,m_2}(t)\right|, c_0 e^{-\delta T_0} \right\}, \end{aligned} \tag{84}$$

where $c_0 > 0$ is a universal constant. Since $\{\hat{y}^{(k)}_{i,m}\}_m$ is a Cauchy sequence in $C([0, \infty))$ equipped with the sup-norm, using the estimate (84), we know $\{\mathcal{T}\hat{y}^{(k)}_{i,m}\}_m$ is a Cauchy sequence in $C([0, 1])$ equipped with the sup-norm. By the completeness of $C([0, 1])$, there exists $f^* \in C([0, 1])$ with $f^*(0) = 0$ such that

$$\lim_{m\to\infty} \sup_{s\in[0,1]} |(\mathcal{T}\hat{y}^{(k)}_{i,m})(s) - f^*(s)| = 0. \tag{85}$$

Given any $s > 0$, we have

$$f^*(s) = \lim_{m\to\infty} (\mathcal{T}\hat{y}_{i,m}^{(k)})(s) = (\mathcal{T}y_i^{(k)})(s), \tag{86}$$

hence

$$\lim_{t\to\infty} e^{\beta t} y_i^{(k)}(t) = \lim_{s\to 0^+} (\mathcal{T}y_i^{(k)})(s) = f^*(0) = 0, \tag{87}$$

which completes the proof. ■

6. The Problem of Optimization and Main Results

In the previous section, we gave a general characterization of the approximation of linear functionals using linear RNNs. It is revealed that memory plays an important role in determining the approximation rates, and vice versa. The result therein only depends on the architecture, and does not concern the actual training dynamics. In this section, we turn our attention to the optimization problem and perform a fine-grained analysis of the dynamics of the training process when applying linear RNNs to learn linear functionals. In this case, we again find an interesting interaction between memory and learning dynamics. These results then put the ubiquitous but heuristic observations - that long-term memory negatively impacts training efficiency (Bengio et al., 1994; Hochreiter et al., 2001) - on concrete theoretical footing, at least in idealized settings. At the same time, we also complement general results on "vanishing and explosion of gradients" (Pascanu et al., 2013; Hanin and Rolnick, 2018; Hanin, 2018) that are typically restricted to initialization settings with more precise characterizations in the dynamical regime during the whole training process.

6.1 Optimization Problem Formulation

We first define the loss function for training. We use the squared difference between the target functional and the (linear) RNN model at some terminal time $T > 0$ averaged over input distributions, which can be written as

$$\mathbb{E}_{\boldsymbol{x}} J_T(\boldsymbol{x}; c, W, U) := \mathbb{E}_{\boldsymbol{x}} |\hat{H}_T(\boldsymbol{x}) - H_T(\boldsymbol{x})|^2 = \mathbb{E}_{\boldsymbol{x}} \left| \int_0^T [c^\top e^{Wt} U - \rho(t)^\top] x_{T-t} dt \right|^2. \tag{88}$$

Without loss of generality, here the input time series $\boldsymbol{x}$ is assumed to be finitely cut off at zero, i.e. $x_t = 0$ for any $t \le 0$ almost surely. Training the RNN amounts to optimizing $\mathbb{E}_{\boldsymbol{x}} J_T$ with respect to the parameters (c, W, U). The most commonly applied method in practice is gradient descent (GD) or its stochastic variants (say SGD), which updates the parameters in the steepest descent direction.

6.1.1 Motivating Numerical Examples

We first show numerically that the gradient descent training dynamics of $\mathbb{E}_{\boldsymbol{x}} J_T$ exhibits very interesting and different behaviors depending on the form of target functionals. Take the input dimension $d = 1$ and recall the target functional $H_T(\boldsymbol{x}) = \int_0^T \rho(t) x_{T-t} dt$, we take

the input $\boldsymbol{x}$ as white noise, while the representation ρ is selected as the exponential sum or the (scaled) Airy function:

1. Exponential sum: $\rho(t) = [c^{*\top}]e^{W^* t}b^*$, where $c^*, b^* \in \mathbb{R}^{m^*}$ are standard normal random vectors, and $W = -I - Z^\top Z$ with $Z \in \mathbb{R}^{m^* \times m^*}$ is a Gaussian random matrix with i.i.d. entries having variance $1/m^*$.

2. Airy function: $\rho(t) = \mathrm{Ai}(s_0[t - t_0])$, where $\mathrm{Ai}(t)$ denotes the Airy function of the first kind, given by the improper integral

$$\mathrm{Ai}(t) = \frac{1}{\pi} \lim_{\xi \to \infty} \int_0^\xi \cos\left(\frac{u^3}{3} + tu\right) du. \tag{89}$$

Note that in the first example, the memory of target functional decays quickly. While for the second example, the effective rate of decay is controlled by the parameter t_0 and s_0: for $t \le t_0$, the Airy function is oscillatory, hence a large amount of memory is present in the target for large t_0.

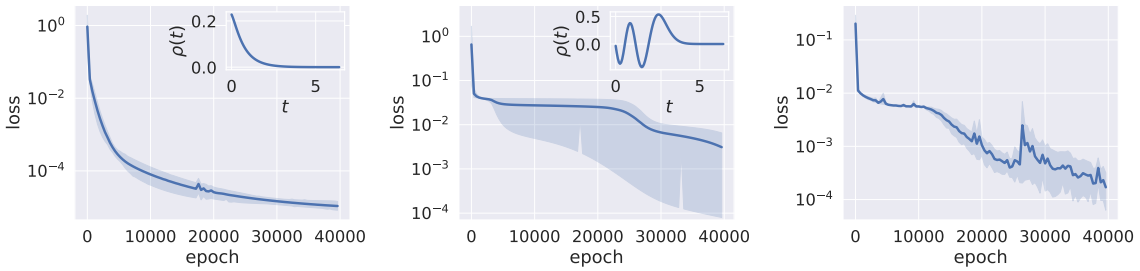
We now show via numerical experiments that the long-term memory adversely affects the optimization process with gradient descent. In Figure 1 (a) (b), we plot the GD dynamics on training linear RNNs (discretized using Euler method, hence equivalent to residual RNNs). We observe that the training proceeds efficiently for the exponential sum target. However, for the Airy function target, there are interesting "plateauing" behaviors of the loss function, where the training slows down significantly after some initial decrements. The plateauing is sustained for a long time before further decrements are observed. This causes a severe slow down of training. This effect gets worse as t_0 increases, which corresponds to a more complex Airy function with more memory effects.

As a further demonstration that this plateauing behavior may be generic, a nonlinear forced dynamical system is also investigated. That is, the Lorenz 96 system (Lorenz, 1996):

$$\begin{aligned} \dot{y} &= -y + x + \sum_{k=1}^{K} z_k/K, \\ \dot{z}_k &= 2[z_{k+1}(z_{k-1} - z_{k+2}) - z_k + y], \quad k = 1, \cdots, K. \end{aligned} \tag{90}$$

Here, x is an external stochastic noise, and the variables z_k are with cyclic indices: $z_{k+K} = z_k$ for $k \in [K]$. When the unresolved variables z_k are unknown, the dynamics of the resolved variable y driven by x is a nonlinear dynamical system with memory effects. We use a standard nonlinear RNN model (with the tanh activation) to learn the sequence-to-sequence mapping $\boldsymbol{x}_{0:T} \mapsto \boldsymbol{y}_{0:T}$ with the Adam optimizer. Figure 1 (c) shows that the training of the Lorenz 96 system with presence of memory also exhibits the interesting plateauing phenomenon.

The results in Figure 1 hint at the fact that there are certainly some functionals which are much harder to learn than others. It is the purpose of the remaining analysis to understand precisely when and why such difficulties occur, at least in simplified but representative settings. In particular, we will again relate this in a precise manner to memory effects in the target functional, which shows yet another facet of the *curse of memory* when it comes to optimization.



(a) Exponential sum (b) Airy function (c) Lorenz 96 dynamical system

Figure 1: Comparison of training dynamics on different types of functionals. (a) and (b): using the linear RNN model with the GD optimizer; (c): using the nonlinear RNN model (with tanh activation) with the Adam optimizer. The shaded region depicts the mean $\pm$ the standard deviation in 10 independent runs using randomized initialization. Here, we set $m^* = 8$ for the exponential sum target, and $t_0 = 3$, $s_0 = 2.25$ for the Airy function target. In all cases, the trained RNN has the hidden dimension 16 and the total length of the path is $T = 6.4$. The continuous-time RNNs are discretized using the Euler method with step size 0.1. Observe that learning complex functionals (Airy, Lorenz) suffers from slow downs in the form of long plateaus.

Remark 17 *There are a number of recent results concerning the training of RNNs using gradient methods (Hardt et al., 2018; Allen-Zhu et al., 2019). They are mostly positive in the sense that trainability is proved under specific settings, including recovering linear dynamics (Hardt et al., 2018) or over-parameterized settings (Allen-Zhu et al., 2019). Our result here concerns the general setting of learning linear functionals, which need not come from certain underlying dynamics, and may not be in the over-parameterized regime. In this setting, we discover on the contrary that the training can become very difficult even in the linear case. This can be understood in a quantitative way, as we will show later.*

6.1.2 Simplifications

Motivated by numerical examples presented above, it is our goal now to precisely analyze the observed difficulty in training (i.e. the plateauing behavior) when the target functional possesses certain structures (with long-term memories), as shown in Section 6.1.1. To make the theoretical analysis amenable, we make the following simplifications.

- Take the input data $\boldsymbol{x}$ to be the white noise, so that

$$x_{T-t}dt \overset{\text{in distribution}}{=} dB_t, \tag{91}$$

  where B_t is the standard d-dimensional Wiener process. As a consequence, simplifying (88) via Itô's isometry gives

$$J_T(c, W, U) := \mathbb{E}_{\boldsymbol{x}} J_T(\boldsymbol{x}; c, W, U) = \int_0^T \left\| c^\top e^{Wt} U - \rho(t)^\top \right\|_2^2 dt. \tag{92}$$

- We focus on the temporal dimension and take the spatial dimension $d = 1$ in (92).[2] Moreover, to investigate the effect of long-term memory, it is necessary to consider the training on large time horizons. Hence, we take $T \to \infty$ to get

$$J_\infty(c, W, b) := \int_0^\infty \left(c^\top e^{Wt} b - \rho(t)\right)^2 dt, \tag{93}$$

  where b is the sole column of U in (92) and $\rho(t)$ becomes a scalar-valued target. This corresponds to the so-called single-input-single-output (SISO) system.

- Due to the difficulty of directly analyzing $\nabla_W e^{Wt}$ and $\nabla_W^2 e^{Wt}$, we consider a further simplified ansatz. Assume that W is a diagonal matrix with negative entries (to guarantee the stability of the model). That is, $W = -\mathrm{diag}(w)$ with $w \in \mathbb{R}_+^m$. Then we can combine $a = b \circ c$ and rewrite the model as

$$\hat{\rho}(t; c, W, b) := c^\top e^{Wt} b = \sum_{i=1}^m a_i e^{-w_i t} \triangleq a^\top e^{-wt} \triangleq \hat{\rho}(t; a, w). \tag{94}$$

  The optimization problem (93) becomes

$$\min_{(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m} J(a, w) := \int_0^\infty \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right)^2 dt. \tag{95}$$

  Here we omit the subscript ∞.[3]

- We apply a continuous-time idealization of the gradient descent dynamics by considering the gradient flow with respect to $J(a, w)$. That is,

$$\begin{cases} a'(\tau) = -\nabla_a J(a(\tau), w(\tau)), \\ w'(\tau) = -\nabla_w J(a(\tau), w(\tau)), \end{cases} \tag{96}$$

  with some initial value $a(0) = a_0 \in \mathbb{R}^m$, $w(0) = w_0 \in \mathbb{R}_+^m$.

As we will show later, applying the training dynamics (96) to optimize (95) is able to serve as a starting point in the fine-grained dynamical analysis, since this still preserves the plateauing behavior observed in the optimization process (Section 6.1.1), provided additional structures related to memories (see details in Section 6.1.3) on the target $\rho \in L^2([0, \infty)) \cap C^2([0, \infty))$ are imposed, as discussed next.

2. One can observe that the spatial dimension plays little role in the previous approximation analysis (see the proof of Theorem 10 and Theorem 11), since each spatial dimension can be handled separately.
3. The time horizon is always taken as ∞ in the whole analysis. Note that here we also omit an index m (width of the network, relating to the model capacity), since it remains unchanged in the following content if not specified.

6.1.3 Heuristic Insights

We start with some informal discussion on probable reasons behind the plateauing behavior. A straightforward computation shows that, for $k = 1, 2, \dots, m$,

$$\frac{\partial J}{\partial a_k}(a, w) = 2 \int_0^\infty e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t) \right) dt, \tag{97}$$

$$\frac{\partial J}{\partial w_k}(a, w) = 2a_k \int_0^\infty (-t) e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t) \right) dt. \tag{98}$$

To construct conditions when the dynamics (96) shows plateauing behaviors, one can consider as follows. When there are plateaus in the training dynamics, the loss J must be large while the gradient norm $\|\nabla J\|_2$ must be small. To make both $\frac{\partial J}{\partial a_k}$ and $\frac{\partial J}{\partial w_k}$ small, we need to require that the multiplier, $e^{-w_k t}$ or $te^{-w_k t}$, is nearly orthogonal to the residual $\hat{\rho}(t; a, w) - \rho(t) = a^\top e^{-wt} - \rho(t)$. Since the large loss also implies a large residual, and notice that multipliers are exponentially decayed, one can construct a sufficient condition: the plateauing behavior occurs if the residual is large only for large t. That is to say, the learned functional differs from the target functional only at large times. This again relates to the long-term memory.

Based on this observation, we build the memory effect explicitly into the target functional. Concretely, we consider a ground truth ρ with the form

$$\rho(t) = \bar{\rho}(t) + \rho_{0,\omega}(t), \tag{99}$$

where $\bar{\rho}$ is a function which can be well-approximated by $\hat{\rho}(t; a, w)$, e.g. $\bar{\rho}$ is also an exponential sum (or integral), which appears the short-term memory; while $\rho_{0,\omega}$ is a bounded function with light tails and the long-term memory (parameterized by $\omega > 0$), e.g. a spike function with a finite support at an increasing time position as $\omega \to 0^+$. In this case, one can investigate the landscape around the area $\hat{\rho}(t; a, w) \approx \bar{\rho}(t)$:

- For small t, the residual is small;
- For large t, the residual is not small (but remains bounded), while $e^{-w_k t}$ will be small.

Both of them give small gradients, but the overall residual (and hence loss) is large. This then leads to the plateauing behavior observed in Section 6.1.1. More importantly, by taking a target like (99), one can get an increasing plateauing time as the memory becomes longer, since $\hat{\rho}$ tends to first fit the short-term memory part $\bar{\rho}$, and then fit the long-term memory part $\rho_{0,\omega}$.[4] A simple example of such target functionals is

$$\rho(t) = a^* e^{-w^* t} + c e^{-\frac{(t-1/\omega)^2}{2\sigma^2}}. \tag{100}$$

where $w^* > 0$ and $a^*, c, \sigma \neq 0$. Observe that as $\omega \to 0^+$, the memory of the sequence of functionals corresponding to ρ increases. We numerically verify that this simple target (100) gives rise to the plateauing behavior (see Figure 2) as expected, just like Figure 1 for the Airy function and the Lorenz 96 system. We will subsequently quantify this behavior in the following section.

4. This is a claim based on the observation of numerical experiments, which are omitted here. The theoretical characterization will be given later.

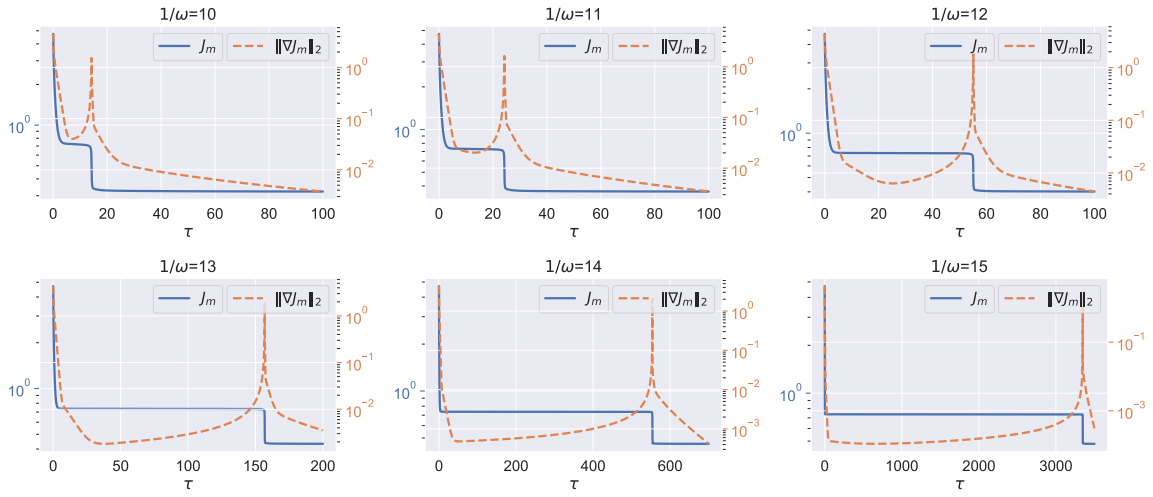


Figure 2: The training dynamics of the target functional defined by (100) using the model (94). Here we take the width $m = 2$ in $\hat{\rho}$. The corresponding gradient flow (96) is numerically solved by the Adams-Bashforth-Moulton method. Observe that the plateauing time increases rapidly as the memory becomes longer ($\omega > 0$ decreases).

Remark 18 *As with the approximation results, we emphasize that it is not obvious at all if target functionals with the representation (99) can be generated by autonomous or forced differential equations. Hence, it is interesting and necessary to consider the general case of learning/approximating sequences of functionals as is done here, as opposed to previous approaches of recovering some underlying dynamical systems using RNNs (Hardt et al., 2018).*

6.2 Main Results

In this section, we present the main optimization results theoretically and numerically.

6.2.1 Theoretical Results

Let us implement the insights proposed above in a precise manner and quantify the plateauing dynamics. Recall the discussion in Section 6.1.2 and Section 6.1.3, i.e. (95) and (99), (100), we aim to analyze the optimization problem

$$\min_{(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m} J_\omega(a,w) := \|\hat{\rho}(t;a,w) - \rho_\omega(t)\|^2_{L^2[0,\infty)}\,, \tag{101}$$

where $\hat{\rho}(t;a,w) := \sum_{i=1}^m a_i e^{-w_i t} \triangleq a^\top e^{-wt}$ is the exponential sum model to be trained (i.e. the kernel of linear RNNs); $\rho_\omega(t)$ denotes the target and is set to be two-part and memory-dependent, as defined in (99) and motivated by (100):

$$\rho_\omega(t) := \bar{\rho}(t) + \rho_{0,\omega}(t) = \int_{w_l}^{w_r} a^*(w) e^{-wt} d\pi_0(w) + \rho_0(t - 1/\omega). \tag{102}$$

Here we take the short-term memory part $\bar{\rho}$ to be an exponential sum, where $w_r > w_l > 0$ are two fixed constants, $a^* : [w_l, w_r] \mapsto \mathbb{R}$ is a fixed bounded function, and π_0 is a fixed probability distribution defined on $[w_l, w_r]$. For the long-term memory part $\rho_{0,\omega}$, $\omega > 0$ controls the quantity of memory, and $\rho_0 : \mathbb{R} \mapsto \mathbb{R}$ is a fixed template function which satisfies the following assumptions.

Assumptions on ρ_0. (i) $\rho_0(t) \not\equiv 0$; (ii) $\rho_0 \in L^2(\mathbb{R}) \cap C^2(\mathbb{R})$; (iii) ρ_0 is bounded on $\mathbb{R}$, i.e. $\|\rho_0\|_{L^\infty(\mathbb{R})} < \infty$; (iv) $\lim_{t\to-\infty} \rho_0(t) = 0$.

Remark 19 *The above assumptions (i), (ii) and (iii) are rather natural, and (iv) only restricts the single side tail of* ρ_0 *to be zero. In the following analysis, we further focus on* ρ_0 *with light tails, e.g. the sub-Gaussian tails*

$$|\rho_0(t)| \le c_0 e^{-c_1 t^2}, \quad \forall t : |t| \ge t_0 \tag{103}$$

for some fixed positive constants c_0, c_1, t_0. *Obviously, the Gaussian densities and continuous functions with compact supports possess the sub-Gaussian tails.*

The ultimate goal is to analyze the gradient flow training dynamics of the loss J_ω:

$$\frac{d}{d\tau}\theta_\omega(\tau) = -\nabla J_\omega(\theta_\omega(\tau)), \quad \theta_\omega(0) = \theta_0, \tag{104}$$

where $\theta_\omega(\tau) := (a_\omega(\tau), w_\omega(\tau)) \in \mathbb{R}^{2m}$ for any $\tau \ge 0$, and the initialization $\theta_0 := (a_0, w_0)$.

In Li et al. (2021), we proved that the *curse of memory* occurs in the optimization of linear RNNs when applied to linear functional targets. That is, the learning can get slow downs with the exponentially large timescale as the target memory increases.

Theorem 20 (Exponential Timescale of Plateauing) *Take* $d\pi_0(w)/dw = \sum_{j=1}^{m^*} \delta(w - w_j^*)$ *with* $m^* < m$.[5] *For any* $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 = (a_0, w_0) \in \mathbb{R}^m \times \mathbb{R}_+^m$ *and* $\delta > 0$, *define the hitting time*

$$\tau_0 = \tau_0(\delta; \omega, m, \theta_0) := \inf\left\{\tau \ge 0 : \|\theta_\omega(\tau) - \theta_0\|_2 > \delta\right\}, \tag{105}$$

$$\tau_0' = \tau_0'(\delta; \omega, m, \theta_0) := \inf\left\{\tau \ge 0 : |J_\omega(\theta_\omega(\tau)) - J_\omega(\theta_0)| > \delta\right\}. \tag{106}$$

Assume that the initialization satisfies $\hat{\rho}(t; \theta_0) \approx \bar{\rho}(t)$. *Then there exist universal constants* $C(\rho_0), C'(\rho_0) > 0$ *only depending on* ρ_0, *such that*

$$J_\omega(\theta_0) \gtrsim C(\rho_0) > 0, \quad \forall \omega \in (0, C'(\rho_0)), \tag{107}$$

and

$$\lim_{\omega\to 0^+} \tau_0(\delta; \omega, m, \theta_0) = \lim_{\omega\to 0^+} \tau_0'(\delta; \omega, m, \theta_0) = +\infty. \tag{108}$$

5. Here $\delta(\cdot)$ denotes the common Dirac function. That is, assume that the distribution π_0 is discrete with particles $\{w_j^*\}_{j=1}^{m^*}$, which gives $\bar{\rho}(t) = \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t}$ with $a_j^* := a^*(w_j^*)$. Without loss of generality, here we set the non-degenerate conditions: $a_j^* \ne 0$, $w_j^* > 0$ and $w_i^* \ne w_j^*$ for any $i \ne j$, $i, j \in [m^*]$. The condition $m^* < m$ is to ensure that $\bar{\rho}$ can be well-approximated by $\hat{\rho}$, and also introduces degeneracy to simplify the analysis and help to derive a more concise bound.

In particular, if ρ_0 has the sub-Gaussian tails (103), and the initialization is bounded as $(a_0, w_0) \in [a_l^0, a_r^0]^m \times [w_l^0, w_r^0]^m$ with constants $a_l^0 < a_r^0$, $0 < w_l^0 < w_r^0$, we further have

$$\tau_0'(\delta; \omega, m, \theta_0) \geq \tau_0(\delta; \omega, m, \theta_0) \gtrsim \omega^2 e^{\frac{w_l^0}{\omega}} \min\left\{\frac{\delta}{\sqrt{m}}, \ln(1+\delta)\right\} \tag{109}$$

for any $\omega \in (0, \min\{1/2, 1/t_0, 2c_1/w_r^0\})$ sufficiently small, where $\delta \ll \min\{1, C(\rho_0)\}$, and $\gtrsim$ hides universal positive constants only depending on a_l^0, a_r^0, w_r^0 and ρ_0, t_0, c_0, c_1.

Remark 21 *The proof of Theorem 20 follows from $\hat{\rho}(t; \theta_0) \equiv \bar{\rho}(t)$ and continuity. According to the linear independence of exponential functions (Lemma 28), we can obtain a further relation between parameters of $\hat{\rho}(\cdot; \theta_0)$ and $\bar{\rho}(\cdot)$ (see Definition 29 and Lemma 30). This further implies a permutation symmetry in the parameter space, which leads to a factorial number of equivalent initialization regions. All the details are found in Definition 29, Lemma 30, Remark 31 and Remark 36 in Section 7.1.*

In the current work, we expand significantly the setting of Theorem 20 (i.e. π_0 is discrete and the initialization condition $\hat{\rho} \approx \bar{\rho}$), and perform a complete dynamical analysis under a general regime: π_0 can be either discrete or continuous, and the initialization is generally defined as follows.

Bounded Initialization. Denote the set of initializations by

$$\begin{aligned}\Theta_0 := \big\{\theta_0 = (a_0, w_0) \in \mathbb{R}^m \times \mathbb{R}_+^m : w_0 \in [w_l^0, w_r^0]^m, \\ a_0 = a_0' m^{-\beta} \text{ with } a_0' \in [a_l^0, a_r^0]^m\big\},\end{aligned} \tag{110}$$

where $0 < w_l^0 < w_r^0$, $a_l^0 < a_r^0$ and $\beta \geq 1$ are fixed constants.

Remark 22 *Here we take the bounded initialization for convenience. There are two points that need explanation:*

- *notice that the exponential sum $\hat{\rho}(t; a, w) = a^\top e^{-wt}$ is (square) integrable on $[0, \infty)$ if and only if $w \succ \mathbf{0}_m$, it is natural to set the non-degenerate condition $w_0 \succeq w_l^0 \mathbf{1}_m$ to ensure a finite initial loss;*
- *the normalization factor $m^{-\beta}$ ($\beta \geq 1$) on a_0 is to ensure a bounded initial loss with respect to the network width $m \in \mathbb{N}_+$. It is due to $\|\hat{\rho}(t; a_0, w_0)\|_{L^2[0,\infty)} = \mathcal{O}(\|a_0\|_1)$, which gives $J_\omega(a_0, w_0) = \mathcal{O}(\|a_0\|_1 + 1)$ (see details in the proof of Lemma 38 and Remark 39).*

The bounded initialization (110) also implies *stability* of the training dynamics (104). The details are found in Lemma 40 (and Remark 41, Lemma 50), which states that there exists a universal constant $\tau_1 > 0$, such that for any $\omega > 0$, $m \in \mathbb{N}_+$ and $\theta_0 \in \Theta_0$, each component of $w_\omega(\tau)$ is *uniformly* bounded away from zero within the time horizon $[0, \tau_1]$.

To summarize, it is shown that as $\omega \to 0^+$, the gradient flow training dynamics (104) becomes a two-stage process with the timescale separation:

- Stage I: $\hat{\rho}(\cdot; \theta_\omega(\tau))$ first learns $\bar{\rho}(\cdot)$ within an $\mathcal{O}(1)$ time;

- Stage II: the dynamical system (104) then gets stuck for a long time, which can be exponentially large as a function of $1/\omega$ (the target memory).

Now we state the precise results regarding both stages. For Stage II, we have the following estimate for the timescale of plateauing under general settings.

Theorem 23 (Extensions on Plateauing Time) *For any $\omega > 0$, $m \in \mathbb{N}_+$ and $\theta_0 \in \Theta_0$, assume*

$$\left\| [a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t) \right\|_{L^2[0,\infty)}^2 \le \bar{c}\bar{\epsilon}, \quad \forall \tau \in [\bar{\tau}/2, \bar{\tau}] \tag{111}$$

with $\bar{\epsilon} = \bar{\epsilon}(\omega, m)$, $\bar{\tau} = \bar{\tau}(\omega, m)$, and $\bar{c} > 0$ is a universal constant independent of ω, m and θ_0.[6] *Then there exist universal constants $C(\rho_0), C'(\rho_0) > 0$ only depending on ρ_0, such that for any $\omega \in (0, C'(\rho_0))$, $m \in \mathbb{N}_+$ with $\bar{\epsilon} = \bar{\epsilon}(\omega, m) < \frac{C(\rho_0)}{4\bar{c}}$ and any $\theta_0 \in \Theta_0$, we have*

$$J_\omega(\theta_\omega(\tau)) \gtrsim C(\rho_0) > 0, \quad \forall \tau \in [\bar{\tau}/2, \bar{\tau}]. \tag{112}$$

In addition, for any $\delta > 0$, $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \Theta_0$ and some $\tau' \in [\bar{\tau}/2, \bar{\tau}]$, define the hitting time

$$\tau_0 = \tau_0(\delta; \omega, m, \theta_0) := \inf\left\{\tau \ge \tau' : \|\theta_\omega(\tau) - \theta_\omega(\tau')\|_2 > \delta\right\}, \tag{113}$$

$$\tau_0' = \tau_0'(\delta; \omega, m, \theta_0) := \inf\left\{\tau \ge \tau' : |J_\omega(\theta_\omega(\tau)) - J_\omega(\theta_\omega(\tau'))| > \delta\right\}. \tag{114}$$

Then for any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l^0\})$ sufficiently small, $m \in \mathbb{N}_+$ appropriately large such that $\bar{\tau} = \bar{\tau}(\omega, m) \le \tau_1$, $\bar{\epsilon} = \bar{\epsilon}(\omega, m)$ and $\bar{\epsilon}/\bar{\tau}$ sufficiently small, and any $\theta_0 \in \Theta_0$, there exists $\tau' = \tau'(\omega, m, \theta_0) \in [\bar{\tau}/2, \bar{\tau}]$, such that

$$\tau_0' - \tau' \ge \tau_0 - \tau' \gtrsim \begin{cases} \min\left\{ \frac{\delta}{\sqrt{m\bar{\epsilon}} + \sqrt{m}\omega^{-1} e^{-\frac{w_l^0}{2\omega}}}, \ \frac{\ln(1+\delta)}{mC^{-m} + \sqrt{\bar{\epsilon}} + \omega^{-2} e^{-\frac{w_l^0}{2\omega}}} \right\}, & \bar{\tau} = 0, \\ \min\left\{ \frac{\delta}{\sqrt{\bar{\epsilon}/\bar{\tau}} + \sqrt{m}\omega^{-1} e^{-\frac{w_l^0}{4\omega}}}, \ \frac{\ln(1+\delta)}{mC^{-m} + \sqrt{\bar{\epsilon}} + \omega^{-2} e^{-\frac{w_l^0}{2\omega}}} \right\}, & \bar{\tau} \ne 0, \end{cases} \tag{115}$$

where $\delta \ll \min\{1, C(\rho_0)\}$, $C > 1$ and $\gtrsim$ hide universal positive constants only depending on a_l^0, a_r^0, w_l^0, w_r^0, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1.

Remark 24 *One can view Theorem 23 as an extension of Theorem 20. In fact, by taking $\bar{\epsilon} = \bar{\tau} = 0$ and noticing that the term mC^{-m} is eliminated due to simpler analysis,*[7] *we can recover Theorem 20 based on Theorem 23. In addition, Theorem 23 implies that an exponentially small $\bar{\epsilon}$ in the target memory $1/\omega$ (in particular, $\bar{\epsilon} = 0$) with an appropriately large $m \in \mathbb{N}_+$ leads to an exponentially large lower bound for the timescale of plateauing. That is to say, the learning to the short-term memory part of the target results in the curse of memory in optimization.*

6. That is, the gradient flow training dynamics (104) achieves an error tolerance $\bar{\epsilon}$ to the short-term memory part $\bar{\rho}(\cdot)$ within a timescale $\mathcal{O}(\bar{\tau})$.
7. Compare the proofs of Proposition 34 and Proposition 46, we have that the term mC^{-m} can be eliminated by introducing the degeneracy (the condition $m^* < m$ in Theorem 20) and applying the "parameter" arguments.

Now we derive conditions for dynamics to display the behavior of Stage I, i.e. $\hat{\rho}(\cdot;\theta_\omega(\tau))$ learns $\bar{\rho}(\cdot)$ within an $\mathcal{O}(1)$-time, and also to illustrate the conditions in Theorem 23 in a sufficient sense. In particular, we consider random bounded initialization for sufficiently wide RNNs.

Random Bounded Initialization. Denote the set of random initializations by

$$\begin{aligned}\tilde{\Theta}_0 :=\big\{&\theta_0=(a_0,w_0)\in\mathbb{R}^m\times\mathbb{R}_+^m:\ \text{the components of } \theta_0 \text{ are i.i.d. sampled,}\\ &w_0\sim\pi_0^m([w_l,w_r]),\ a_0=a_0'm^{-\beta}\ \text{with } a_0'\sim\pi_1^m([a_l,a_r])\big\},\end{aligned}\tag{116}$$

where $0<w_l<w_r$, $a_l<a_r$ and $\beta\geq 1$ are fixed constants, and π_1 is a fixed probability distribution defined on $[a_l,a_r]$.

Theorem 25 (Stage-I Dynamics) *Set $p\in(1/3,1)$ and assume that $m\geq 1/\tau_1^{1/p}$ with τ_1 defined above (after Remark 22). Fix any $\omega\in(0,\min\{1/2,1/t_0,4c_1/w_l\})$ and $m\in\mathbb{N}_+$. For any $\delta_0>0$, with probability of at least $1-\delta_0$ over the choice of $\theta_0\in\tilde{\Theta}_0$, we have*

$$\begin{aligned}&\left\|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t}-\bar{\rho}(t)\right\|_{L^2[0,\infty)}^2\\ \lesssim\ &\frac{1}{m^{1-p}}+\ln(6/\delta_0)\left(\frac{1}{m^{6p-2}}+\frac{1}{m}\right)+m^{2(1-p)}e^{-\frac{w_l}{2\omega}},\quad\forall\tau\in[m^{-p}/2,m^{-p}],\end{aligned}\tag{117}$$

where $\lesssim$ hides universal positive constants only depending on a_l, a_r, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1.

According to Theorem 25, when there is sufficient model capacity, the gradient flow training dynamics of linear RNNs learns a solution close to the short-term memory part of target rapidly with high probabilities, even when the target with long-term memory.

Remark 26 *By Theorem 25, let*

$$\begin{aligned}\bar{\epsilon}=\bar{\epsilon}(\omega,m)&=\frac{1}{m^{1-p}}+\ln(6/\delta_0)\left(\frac{1}{m^{6p-2}}+\frac{1}{m}\right)+m^{2(1-p)}e^{-\frac{w_l}{2\omega}},\\ \bar{\tau}=\bar{\tau}(\omega,m)&=m^{-p},\quad \bar{c}=\bar{c}(a_l,a_r,w_l,w_r,a^*,\rho_0,t_0,c_0,c_1),\end{aligned}$$

we get

$$\bar{\epsilon}(\omega,m)/\bar{\tau}(\omega,m)=\frac{1}{m^{1-2p}}+\ln(6/\delta_0)\left(\frac{1}{m^{5p-2}}+\frac{1}{m^{1-p}}\right)+m^{2-p}e^{-\frac{w_l}{2\omega}}.$$

Further restrict $p\in(2/5,1/2)$, we have

$$\lim_{m\to\infty}\lim_{\omega\to0^+}\bar{\epsilon}(\omega,m)=\lim_{m\to\infty}\lim_{\omega\to0^+}\bar{\tau}(\omega,m)=\lim_{m\to\infty}\lim_{\omega\to0^+}\bar{\epsilon}(\omega,m)/\bar{\tau}(\omega,m)=0,$$

which implies that $\bar{\epsilon}$, $\bar{\tau}$ and $\bar{c}$ satisfy all the conditions in Theorem 23 for any $\omega>0$ sufficiently small and $m\in\mathbb{N}_+$ sufficiently large. Applying Theorem 25 to Theorem 23 gives that for any $\delta_0>0$, with the probability of at least $1-\delta_0$ over the initialization $\theta_0\in\tilde{\Theta}_0$,

the plateauing timescale $\tau_0(\delta;\omega,m,\theta_0), \tau_0'(\delta;\omega,m,\theta_0)$ *defined in* (113), (114) *can be lower bounded by*

$$\min\left\{\frac{\delta}{m^{p-\frac{1}{2}}+\frac{\sqrt{\ln(6/\delta_0)}}{m^{\frac{5}{2}p-1}}+m^{1-\frac{p}{2}}\omega^{-1}e^{-\frac{w_l}{4\omega}}}, \frac{\ln(1+\delta)}{mC^{-m}+m^{\frac{p-1}{2}}+\frac{\sqrt{\ln(6/\delta_0)}}{m^{3p-1}}+m^{1-p}\omega^{-2}e^{-\frac{w_l}{4\omega}}}\right\}. \tag{118}$$

This lower bound goes to infinity in the iterated limit when $\omega \to 0^+$ *first and then* $m \to \infty$.

Connecting Theorem 25 with Theorem 23, we conclude that the gradient flow training dynamics for learning with linear RNNs appears a typical *two-stage process with timescale separation*. That is, when there are long-term memories in targets, the model first learns the short-term memory part of target rapidly, then becomes stuck for a long time. Increasing the model capacity cannot alleviate this issue. This is because, even if we increase the width m of linear RNNs, a sufficiently long memory term ($\omega > 0$ sufficiently small that depends on m) can make the lower bound in (118) arbitrarily large.

6.2.2 Numerical Verifications

(1) Timescale Estimates

We numerically verify the timescale proved in Theorem 20. That is, we verify that the timescale of plateauing ($|J_\omega(\theta_\omega(\tau)) - J_\omega(\theta_0)|$, see (106)) and parameter separation ($\|\theta_\omega(\tau) - \theta_0\|_2$, see (105)) are exponentially large as the memory $1/\omega \to +\infty$. The results are shown in Figure 3, where we observe good agreement with the predicted scaling.

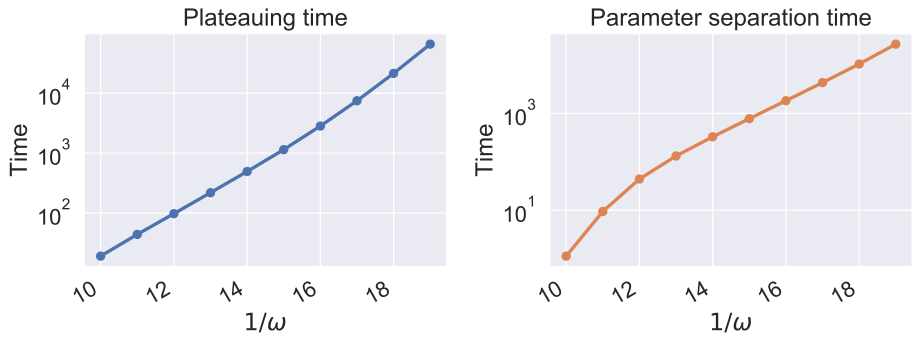


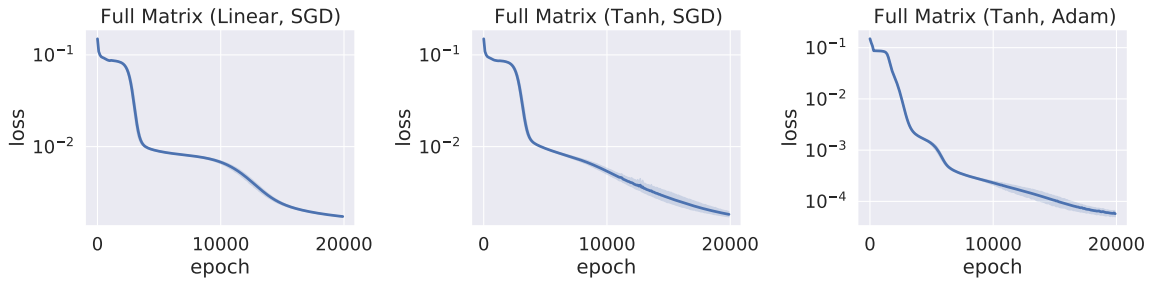
Figure 3: The timescale of plateauing and parameter separation. Here the model and target are both selected the same as Figure 2, but with a larger width $m = 10$. We observe that the logarithm of plateauing time and parameter separation is almost linear to the memory $1/\omega$.

(2) General Settings

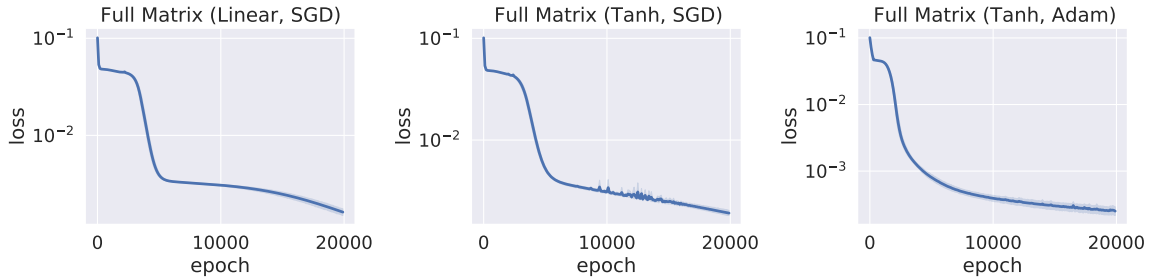
To facilitate the mathematical tractability, the results in Section 6.2.1 is derived under the restrictive conditions of the diagonal W (recurrent kernel) with negative entries, linear activations and the gradient flow training dynamics. However, we show here that the plateauing behavior - which we now understand as a generic feature of long-term memory

of the target functional and its interaction with the optimization dynamics - is present even for more general settings, and hence our simplified analytical setting is representative of general situations.

In Figure 4, we still take the target functional as defined in (100), but apply more general models to learn it, including using the RNN with full (non-diagonal) recurrent kernel W with no restrictions on entries, using the nonlinear activation (tanh) and using the Adam optimizer (Kingma and Ba, 2015). Furthermore, to be consistent with practice, we use the actual input sample paths of finite time horizons, instead of taking the Itô isometry simplification. We observe that the plateauing behaviors are present in all cases. Moreover, in the last case of applying the Adam optimizer (which can be viewed as a momentum-based optimization method), the plateauing behavior is somewhat alleviated, although the separation of timescales is still present. This is consistent with our supplemental analysis in Appendix B, where we show that momentum-based methods will speed up training based on the dynamical analysis of plateauing given in Section 7.1.



(a) White noise inputs



(b) Cosine noise inputs

Figure 4: Numerical verifications of the plateauing behavior under general settings, with the non-diagonal recurrent kernel, the nonlinear activation (tanh), and the Adam (momentum-based) optimizer. Here we use the target functional the same as Figure 2 with the memory $1/\omega = 20$. The time horizon is chosen as $T = 32$, and 128 input samples are generated from a standard white noise. The learning rate is 1.0 for GD and 0.001 for Adam. 10 initializations are sampled and trained for each experiment. We consider two possible input distributions: (a) white noise inputs; (b) inputs of the form $x_t = \sum_{j=1}^{J} \alpha_j \cos(\lambda_j t)$, where $\lambda_j \sim \mathcal{U}[0, 10]$ and $\alpha_j \sim \mathcal{N}(0, 1)$. We observe that plateaus occur in all cases, and the momentum generally improves the situation but still not resolve the difficulty.

6.2.3 The Curse of Memory in Optimization

Let us summarize the findings from Section 6.2.1 and Section 6.2.2. The form of target functional defined in (102) captures the long-term memory in a concrete way. Note that when $\omega \to 0^+$, it corresponds to the case where the influence of target functional H_t does not decay, much like that considered in the curse of memory in approximation (see Section 4.3). However, different from the approximation part where an exponentially large number of neurons/hidden states is required to achieve a given tolerance, here in optimization the adverse effect of memories comes with (possibly exponential) slow downs of the gradient flow training dynamics. In other words, the plateauing time can be exponential in memory, as shown in Theorem 20 and numerically verified in Figure 3. Under more general settings, this adverse effect of memories is still quantified, even increasing the model capacity can not help (Theorem 23 and Theorem 25). While these results are proved under sensible but restrictive conditions, we show numerically in Figure 4 that they are representative of the general settings.

In the literature, a number of results have been obtained pertaining to the analysis of training dynamics of RNNs. A positive result for training with GD is established in Hardt et al. (2018), but that is in the setting of identifying dynamical systems. That is, the target functional analyzed there is generated by a hidden linear dynamical system, hence it must possess exponential memory decay properties that is consistent with the RNN, provided that the hidden linear dynamics is stable. The results here show that if one makes no assumption that the target linear functional is generated from a linear dynamical system, then the situation is different. Another example of positive result is when the RNN is sufficiently over-parameterized in empirical risk minimization (i.e. the number of model parameters are much more than the number of samples; see Allen-Zhu et al. (2019)). Here, we consider population risk minimization, and we show that in this case, even when the model size is large, the risk minimization problem can become difficult. These results provide an alternative analysis of a setting that is representative of the difficulties one may encounter in practice. In particular, the curse of memory that we quantified here is consistent with the difficulty in the training of RNNs often observed in practical applications, where heuristic attributions to "memory" are often alluded to (Hu et al., 2018; Campos et al., 2018; Talathi and Vartak, 2015; Li et al., 2018). The analysis here makes the connection between target memories and optimization difficulties precise, and forms a basis for the principled development of means to overcome such difficulties in applications.

7. Proofs of Optimization Results

In this section, proofs of the theoretical results shown in Section 6.2.1 are given.

7.1 Exponential Timescale of Plateauing

We prove Theorem 20 in this section. The basic insight is, by adding (appropriate) long-term memories in targets, one can increase the loss with little effect on the gradient and Hessian, which leads to a significant slow down of the gradient flow training dynamics near the short-term memory part of target. Therefore, Theorem 20 is proved subsequently in the following procedure:

1. We prove that J_ω has a large value but small gradient when $\hat{\rho}(t;a,w) \equiv \bar{\rho}(t)$;
2. We prove that when $\hat{\rho}(t;a,w) \equiv \bar{\rho}(t)$, the Hessian $\nabla^2 J_\omega$ is positive semi-definite for $\omega = 0$, but for finite, small $\omega > 0$, $\nabla^2 J_\omega$ has $\mathcal{O}(1)$ positive eigenvalues and multiple $o(1)$ eigenvalues;
3. Based on these results, we perform a local linearization analysis on the gradient flow (104) initialized by $\hat{\rho}(t;a_0,w_0) \equiv \bar{\rho}(t)$, from which and by continuity the timescale of plateauing is derived.

(1) Preliminary Results

Recall the assumptions on targets in Section 6.2.1. Since $d\pi_0(w)/dw = \sum_{j=1}^{m^*} \delta(w - w_j^*)$, we get

$$\begin{aligned}\rho_\omega(t) = \bar{\rho}(t) + \rho_{0,\omega}(t) &= \int_{w_l}^{w_r} a^*(w) e^{-wt} d\pi_0(w) + \rho_0(t - 1/\omega) \\ &= \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t} + \rho_0(t - 1/\omega).\end{aligned}$$

Here, $a_j^* := a^*(w_j^*) \neq 0$, $w_j^* > 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j \in [m^*]$, and $m^* < m$. The former requirements are just non-degenerate conditions, and the last requirement ensures that the model can perfectly represent the short-term memory part of target, $\bar{\rho}(t)$. The memory in target is controlled by $\rho_{0,\omega}(t) = \rho_0(t - 1/\omega)$, with ρ_0 as a fixed template function. As $\omega \to 0^+$, the support of function shifts towards large times, modeling the dominance of long-term memories. Recall the assumptions on ρ_0: (i) $\rho_0(t) \not\equiv 0$; (ii) $\rho_0 \in L^2(\mathbb{R}) \cap C^2(\mathbb{R})$; (iii) $\|\rho_0\|_{L^\infty(\mathbb{R})} < \infty$; (iv) $\lim_{t \to -\infty} \rho_0(t) = 0$. For quantitative analysis, we mainly focus on ρ_0 with the sub-Gaussian tails (103).

We begin by the following preliminary estimate that is used throughout the subsequent analysis.

Lemma 27 *For any $n \in \mathbb{N}$, $\omega > 0$ and $w > 0$, let*

$$\Delta_{n,\omega}(w) := \int_0^\infty t^n e^{-wt} \rho_{0,\omega}(t) dt, \tag{119}$$

$$\Delta_{n,\omega}^+(w) := \int_0^\infty t^n e^{-wt} |\rho_{0,\omega}(t)| dt. \tag{120}$$

Then

- *$\Delta_{n,\omega}^+(w)$ is monotonically decreasing on $(0, \infty)$;*
- *$\lim_{\omega \to 0^+} \Delta_{n,\omega}(w) = \lim_{\omega \to 0^+} \Delta_{n,\omega}^+(w) = 0$;*
- *In particular, if ρ_0 is sub-Gaussian, we further have*

$$|\Delta_{n,\omega}(w)| \le \Delta_{n,\omega}^+(w) \lesssim \omega^{-n} e^{-w/\omega} \left(c_2^{w^2} + c_3^w \right), \quad \omega \in (0, \min\{1/2, 1/t_0, 2c_1/w\}). \tag{121}$$

Here $c_2 = e^{\frac{1}{4c_1}} > 1$, $c_3 = e^{t_0} > 1$, and $\lesssim$ hides universal positive constants only depending on n and ρ_0, t_0, c_0, c_1.

Proof (i) Obviously $\Delta^+_{n,\omega}(w_1) \le \Delta^+_{n,\omega}(w_2)$ for any $w_1 > w_2 > 0$.

(ii) Obviously $|\Delta_{n,\omega}(w)| \le \Delta^+_{n,\omega}(w)$, we only need to show $\lim\limits_{\omega\to 0^+} \Delta^+_{n,\omega}(w) = 0$. By the assumptions on ρ_0, we get

$$\lim_{\omega\to 0^+} |\rho_{0,\omega}(t)| = \lim_{\omega\to 0^+} |\rho_0(t-1/\omega)| = \lim_{s\to-\infty} |\rho_0(s)| = 0, \quad \forall t \ge 0,$$

and $M_0 := \|\rho_0\|_{L^\infty(\mathbb{R})} < +\infty$, which gives $t^n e^{-wt}|\rho_{0,\omega}(t)| \le M_0 t^n e^{-wt} \in L^1([0,\infty))$ for any $n \in \mathbb{N}$, $\omega > 0$ and $w > 0$. By Lebesgue's dominant convergence theorem, we have

$$\lim_{\omega\to 0^+} \Delta^+_{n,\omega}(w) = \int_0^\infty t^n e^{-wt} \cdot \lim_{\omega\to 0^+} |\rho_{0,\omega}(t)| dt = 0, \quad \forall n \in \mathbb{N},\ \forall w > 0. \tag{122}$$

(iii) Now we estimate $\Delta^+_{n,\omega}(w)$ under the sub-Gaussian condition (103). Suppose $0 < \omega < 1/t_0$, we have

$$\begin{aligned}
\Delta^+_{n,\omega}(w) &= \int_0^\infty t^n e^{-wt}|\rho_0(t-1/\omega)|dt \\
&= \int_{1/\omega-t_0}^{1/\omega+t_0} t^n e^{-wt}|\rho_0(t-1/\omega)|dt + \int_0^{1/\omega-t_0} t^n e^{-wt}|\rho_0(t-1/\omega)|dt \\
&\quad + \int_{1/\omega+t_0}^\infty t^n e^{-wt}|\rho_0(t-1/\omega)|dt \triangleq I_1 + I_2 + I_3.
\end{aligned}$$

Then we bound I_1, I_2 and I_3 respectively:

$$\begin{aligned}
I_1 &\le M_0 \int_{1/\omega-t_0}^{1/\omega+t_0} t^n e^{-wt} dt \le M_0 e^{-w(1/\omega-t_0)} \int_{1/\omega-t_0}^{1/\omega+t_0} t^n dt \\
&= M_0 e^{wt_0} \cdot e^{-w/\omega} \cdot \frac{(1+\omega t_0)^{n+1} - (1-\omega t_0)^{n+1}}{(n+1)\omega^{n+1}} \\
&\lesssim M_0 e^{wt_0} \omega^{-n} e^{-w/\omega}(t_0+\omega),
\end{aligned}$$

where $\omega \in (0, 1/2)$, and $\lesssim$ hides universal positive constants only related to n and t_0. Let $1/c_1 := 2\sigma^2$, we have

$$I_2 = e^{-w/\omega} \int_{-1/\omega}^{-t_0} (s+1/\omega)^n e^{-ws}|\rho_0(s)|ds \le e^{-w/\omega} \int_{-1/\omega}^{-t_0} (s+1/\omega)^n e^{-ws} \cdot c_0 e^{-c_1 s^2} ds,$$

where

$$\begin{aligned}
\int_{-1/\omega}^{-t_0} (s+1/\omega)^n e^{-ws} e^{-c_1 s^2} ds &= e^{\frac{w^2}{4c_1}} \int_{-1/\omega}^{-t_0} (s+1/\omega)^n e^{-c_1(s+\frac{w}{2c_1})^2} ds \\
&\le e^{\sigma^2 w^2/2} \int_{\mathbb{R}} (|t| + |1/\omega - \sigma^2 w|)^n \cdot e^{-\frac{t^2}{2\sigma^2}} dt \\
&= e^{\sigma^2 w^2/2} \sum_{k=0}^n C_n^k (1/\omega - \sigma^2 w)^{n-k} \cdot 2 \int_0^\infty t^k e^{-\frac{t^2}{2\sigma^2}} dt \\
&\le e^{\sigma^2 w^2/2} \sum_{k=0}^n C_n^k (1/\omega)^{n-k} \big(\sqrt{2}\sigma\big)^{k+1} \Gamma\left(\frac{k+1}{2}\right)
\end{aligned}$$

holds for any $\omega \in (0, 2c_1/w)$. Here the last inequality is due to the Mellin Transform of absolute moments of the Gaussian density (see Dytso et al. (2018), Proposition 1). The argument is similar for I_3, which gives the same bound as I_2. Combining all the estimates gives the desired conclusion. The proof is completed. ■

The main idea to analyze plateauing behaviors is to investigate the local dynamics of the gradient flow (104) when $\hat{\rho} = \bar{\rho}$, then extend the results to the setting $\hat{\rho} \approx \bar{\rho}$ by continuity. Recall that both of them are exponential sums, we can obtain the relation of parameters between $\hat{\rho}$ and $\bar{\rho}$, according to the following lemma.

Lemma 28 *For any $m \in \mathbb{N}_+$, let $\lambda = (\lambda_1, \cdots, \lambda_m)$ with $\lambda_i \neq \lambda_j$ for any $i \neq j$, $i, j \in [m]$. Then the series of functions $\left\{e^{\lambda_i t}\right\}_{i=1}^m$ is linear independent on any interval $I \subset \mathbb{R}$.*

Proof The aim is to show

$$\sum_{i=1}^m c_i e^{\lambda_i t} = 0,\ t \in I \Rightarrow c_i = 0,\ \forall i \in [m]. \tag{123}$$

(123) holds trivially for $m = 1$. Assume that (123) holds for $m - 1$, then

$$\begin{aligned} \sum_{i=1}^m c_i e^{\lambda_i t} = 0,\ t \in I &\Rightarrow \sum_{i=1}^{m-1} c_i e^{(\lambda_i - \lambda_m)t} + c_m = 0,\ t \in I \\ &\Rightarrow \sum_{i=1}^{m-1} c_i (\lambda_i - \lambda_m) e^{(\lambda_i - \lambda_m)t} = 0,\ t \in I. \end{aligned} \tag{124}$$

By induction, we get $c_i(\lambda_i - \lambda_m) = 0$ for any $i = 1, \cdots, m-1$. Since $\lambda_1, \cdots, \lambda_m$ are distinct, we have $c_i = 0$ for any $i = 1, \cdots, m-1$. Together with (124), we get $c_m = 0$, which completes the proof. ■

Definition 29 *Let $m \geq m^*$. For any partition $\mathcal{P}$: $[m] = \cup_{j=0}^{m^*} \mathcal{I}_j$ with $\mathcal{I}_{j_1} \cap \mathcal{I}_{j_2} = \varnothing$ for any $j_1 \neq j_2$, $j_1, j_2 \in \{0\} \cup [m^*]$, and $\mathcal{I}_0 = \cup_{r=1}^{i_0} \mathcal{I}_{0,r}$ with $\mathcal{I}_{0,r_1} \cap \mathcal{I}_{0,r_2} = \varnothing$ for any $r_1 \neq r_2$, $r_1, r_2 \in [i_0]$, where $\mathcal{I}_j \neq \varnothing$ for any $j \in [m^*]$ and $\mathcal{I}_{0,r} \neq \varnothing$ for any $r \in [i_0]$ (if $\mathcal{I}_0 \neq \varnothing$), define the affine space (with respect to $\mathcal{P}$):*

$$\begin{aligned} \mathcal{M}_{\mathcal{P}}^* := \Bigg\{ &(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m : \sum_{i \in \mathcal{I}_j} a_i = a_j^*,\ w_i = w_j^* \text{ for any } i \in \mathcal{I}_j,\ j \in [m^*]; \\ &\sum_{i \in \mathcal{I}_{0,r}} a_i = 0,\ w_i = v_r \neq w_j^* \text{ for any } i \in \mathcal{I}_{0,r},\ r \in [i_0] \text{ and } j \in [m^*] \Bigg\}. \end{aligned}$$

Denote the collection of all such affine spaces by $\mathcal{M}^* := \bigcup_{\mathcal{P}} \mathcal{M}_{\mathcal{P}}^*$.

The following lemma characterizes the relation of parameters, by showing that $\mathcal{M}^*$ is exactly the set of equivalent points to (a^*, w^*) for the purpose of representation via exponential sums.

Lemma 30 *For any* $(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m$, $\hat{\rho}(t;a,w)\equiv\bar{\rho}(t)\Leftrightarrow(a,w)\in\mathcal{M}^*$.

Proof (i) ($\Leftarrow$) Since $(a,w)\in\mathcal{M}^*$, there exists $\mathcal{P}$ such that $(a,w)\in\mathcal{M}_{\mathcal{P}}^*$. Then for any $t\geq 0$,

$$\begin{aligned}\hat{\rho}(t;a,w)&=\sum_{i=1}^{m}a_ie^{-w_it}=\sum_{j=0}^{m^*}\sum_{i\in\mathcal{I}_j}a_ie^{-w_it}=\sum_{r=1}^{i_0}\sum_{i\in\mathcal{I}_{0,r}}a_ie^{-w_it}+\sum_{j=1}^{m^*}\sum_{i\in\mathcal{I}_j}a_ie^{-w_it}\\&=\sum_{r=1}^{i_0}\left(\sum_{i\in\mathcal{I}_{0,r}}a_i\right)e^{-v_rt}+\sum_{j=1}^{m^*}\left(\sum_{i\in\mathcal{I}_j}a_i\right)e^{-w_j^*t}=\sum_{j=1}^{m^*}a_j^*e^{-w_j^*t}=\bar{\rho}(t).\end{aligned}$$

(ii) ($\Rightarrow$) Let $\mathcal{I}_j=\left\{i\in[m]:w_i=w_j^*\right\}$ for any $j\in[m^*]$, and $\mathcal{I}_0=\left\{i\in[m]:w_i\neq w_j^*\text{ for any }j\in[m^*]\right\}$. Recall that $\bar{\rho}(t)=\sum_{j=1}^{m^*}a_j^*e^{-w_j^*t}$ is non-degenerate: $a_j^*\neq 0$, $w_j^*>0$ and $w_i^*\neq w_j^*$ for any $i\neq j$, $i,j\in[m^*]$, we get $[m]=\cup_{j=0}^{m^*}\mathcal{I}_j$, $\mathcal{I}_{j_1}\cap\mathcal{I}_{j_2}=\varnothing$ for any $j_1\neq j_2$, $j_1,j_2\in\{0\}\cup[m^*]$. Combining Lemma 28 and the non-degeneracy of $\bar{\rho}$, $\mathcal{I}_j\neq\varnothing$ for any $j\in[m^*]$. Assume that there are i_0 different components in $(w_i)_{i\in\mathcal{I}_0}$, say $v_1,\cdots,v_{i_0}$, then $v_r\neq w_j^*$ for any $r\in[i_0]$ and $j\in[m^*]$. Let $\mathcal{I}_{0,r}=\left\{i\in\mathcal{I}_0:w_i=v_r\right\}$ for any $r\in[i_0]$, we get $\mathcal{I}_{0,r}\neq\varnothing$ for any $r\in[i_0]$ (if $\mathcal{I}_0\neq\varnothing$), and $\mathcal{I}_0=\cup_{r=1}^{i_0}\mathcal{I}_{0,r}$, and $\mathcal{I}_{0,r_1}\cap\mathcal{I}_{0,r_2}=\varnothing$ for any $r_1\neq r_2$, $r_1,r_2\in[i_0]$. Hence $[m]=\cup_{j=0}^{m^*}\mathcal{I}_j$ with $\mathcal{I}_0=\cup_{r=1}^{i_0}\mathcal{I}_{0,r}$ forms a $\mathcal{P}$ defined in Definition 29, and

$$\begin{aligned}0\equiv\hat{\rho}(t;a,w)-\bar{\rho}(t)&=\sum_{i=1}^{m}a_ie^{-w_it}-\sum_{j=1}^{m^*}a_j^*e^{-w_j^*t}\\&=\sum_{j=0}^{m^*}\sum_{i\in\mathcal{I}_j}a_ie^{-w_it}-\sum_{j=1}^{m^*}a_j^*e^{-w_j^*t}\\&=\sum_{r=1}^{i_0}\sum_{i\in\mathcal{I}_{0,r}}a_ie^{-w_it}+\left(\sum_{j=1}^{m^*}\sum_{i\in\mathcal{I}_j}a_ie^{-w_it}-\sum_{j=1}^{m^*}a_j^*e^{-w_j^*t}\right)\\&=\sum_{r=1}^{i_0}\left(\sum_{i\in\mathcal{I}_{0,r}}a_i\right)e^{-v_rt}+\sum_{j=1}^{m^*}\left(\sum_{i\in\mathcal{I}_j}a_i-a_j^*\right)e^{-w_j^*t}.\end{aligned}$$

Again by Lemma 28, we have $\sum_{i\in\mathcal{I}_j}a_i=a_j^*$ for any $j\in[m^*]$ and $\sum_{i\in\mathcal{I}_{0,r}}a_i=0$ for any $r\in[i_0]$, which gives $(a,w)\in\mathcal{M}_{\mathcal{P}}^*$. The proof is completed. ■

Remark 31 *Let* $\mathcal{I}_0=\varnothing$.[8] *It is straightforward to check that for any partition* $\mathcal{P}$, *the dimension of* $\mathcal{M}_{\mathcal{P}}^*$ *is* $\sum_{j=1}^{m^*}(|\mathcal{I}_j|-1)=m-m^*$. *In addition, it can be verified that the cardinality of* $\mathcal{M}^*$ *is* $m^*!\left\{{m\atop m^*}\right\}$, *where* $\left\{{m\atop m^*}\right\}$ *is the Stirling number of the second kind.*[9]

8. That is, the non-degenerate case. Obviously, $\mathcal{I}_0\neq\varnothing$ implies an uncountable $\mathcal{M}_{\mathcal{P}}^*$, but they are all degenerate.
9. The result follows from basic knowledge of combinatorics. See details in the proof of Theorem 57.

(2) Loss

Proposition 32 *There exist universal constants $C(\rho_0), C'(\rho_0) > 0$ only depending on ρ_0, such that for any $(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m$ satisfying $\hat{\rho}(t; a, w) \equiv \bar{\rho}(t)$, we have*

$$J_\omega(a, w) \geq C(\rho_0) > 0, \quad \forall \omega \in (0, C'(\rho_0)). \tag{125}$$

That is, the loss is lower bounded away from zero uniformly in sufficiently small $\omega > 0$.

Proof Recall the assumptions on ρ_0, we have $\rho_0(t) \not\equiv 0$ and $\rho_0(t) \in C(\mathbb{R})$. Let $t_1 \in \mathbb{R}$ satisfying $\rho_0(t_1) \neq 0$. By continuity, there exists $\delta_0 > 0$ such that $|\rho_0(t)| \geq |\rho_0(t_1)|/2$ for any $t \in [t_1 - \delta_0, t_1 + \delta_0]$. Hence, for any $\omega > 0$ satisfying $-1/\omega < t_1 - \delta_0$, we have

$$\begin{aligned} \|\rho_{0,\omega}\|_{L^2[0,\infty)}^2 &= \int_0^\infty \rho_0^2(t - 1/\omega) dt = \int_{-\frac{1}{\omega}}^\infty \rho_0^2(t) dt \\ &\geq \int_{t_1 - \delta_0}^{t_1 + \delta_0} \rho_0^2(t) dt \geq \frac{1}{2} \delta_0 |\rho_0(t_1)|^2 > 0. \end{aligned} \tag{126}$$

Let $C(\rho_0) = \delta_0 |\rho_0(t_1)|^2/2$ and $C'(\rho_0) = 1/|t_1 - \delta_0|$. Then for any $(\hat{a}, \hat{w}) \in \mathbb{R}^m \times \mathbb{R}_+^m$ such that $\hat{\rho}(t; \hat{a}, \hat{w}) \equiv \bar{\rho}(t)$, and any $\omega \in (0, C'(\rho_0))$, we get

$$J_\omega(\hat{a}, \hat{w}) = \|\hat{\rho}(t; \hat{a}, \hat{w}) - \bar{\rho}(t) - \rho_{0,\omega}(t)\|_{L^2[0,\infty)}^2 = \|\rho_{0,\omega}\|_{L^2[0,\infty)}^2 \geq C(\rho_0) > 0,$$

which completes the proof. ∎

(3) Gradient

Proposition 33 *For any $(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m$ satisfying $\hat{\rho}(t; a, w) \equiv \bar{\rho}(t)$, we have*

$$\lim_{\omega \to 0^+} \|\nabla J_\omega(a, w)\|_2 = 0. \tag{127}$$

In particular, if ρ_0 has the sub-Gaussian tails (103), the estimate

$$\|\nabla J_\omega(a, w)\|_2 \lesssim \sqrt{m} \omega^{-1} e^{-w_{min}/\omega} \left(c_2^{w_{min}^2} + c_3^{w_{min}} \right) (1 + \|a\|_\infty) \tag{128}$$

holds for any $\omega \in (0, \min\{1/2, 1/t_0, 2c_1/w_{min}\})$. Here $w_{min} := \min_{i \in [m]} w_i > 0$, $c_2, c_3 > 1$ are constants only related to c_1, t_0, and $\lesssim$ hides universal positive constants only depending on ρ_0, t_0, c_0, c_1.

Proof A straightforward computation shows, for $k = 1, 2, \cdots, m$,

$$\frac{\partial J_\omega}{\partial a_k}(a, w) = 2 \left[\sum_{i=1}^m \frac{a_i}{w_k + w_i} - \sum_{j=1}^{m^*} \frac{a_j^*}{w_k + w_j^*} \right] - 2\Delta_{0,\omega}(w_k), \tag{129}$$

$$\frac{\partial J_\omega}{\partial w_k}(a, w) = -2a_k \left[\sum_{i=1}^m \frac{a_i}{(w_k + w_i)^2} - \sum_{j=1}^{m^*} \frac{a_j^*}{(w_k + w_j^*)^2} \right] + 2a_k \Delta_{1,\omega}(w_k). \tag{130}$$

Fix any $(\hat{a},\hat{w})\in\mathbb{R}^m\times\mathbb{R}_+^m$ satisfying $\hat{\rho}(t;\hat{a},\hat{w})\equiv\bar{\rho}(t)$. By Lemma 30, we have $(\hat{a},\hat{w})\in\mathcal{M}^*$. Recall Definition 29, there exists a partition $\mathcal{P}$: $[m]=\cup_{j=0}^{m^*}\mathcal{I}_j$ with $\mathcal{I}_0=\cup_{r=1}^{i_0}\mathcal{I}_{0,r}$, where $\mathcal{I}_j\neq\varnothing$ for any $j\in[m^*]$ and $\mathcal{I}_{0,r}\neq\varnothing$ for any $r\in[i_0]$ (if $\mathcal{I}_0\neq\varnothing$), such that $(\hat{a},\hat{w})\in\mathcal{M}_{\mathcal{P}}^*$, which gives that $\sum_{i\in\mathcal{I}_j}\hat{a}_i=a_j^*$, $\hat{w}_i=w_j^*$ for any $i\in\mathcal{I}_j$ and $j\in[m^*]$; $\sum_{i\in\mathcal{I}_{0,r}}\hat{a}_i=0$, $\hat{w}_i=v_r\neq w_j^*$ for any $i\in\mathcal{I}_{0,r}$, $r\in[i_0]$ and $j\in[m^*]$. Therefore, for any $n\in\mathbb{N}_+$, we have

$$\begin{aligned}
&\sum_{i=1}^{m}\frac{\hat{a}_i}{(\hat{w}_k+\hat{w}_i)^n}-\sum_{j=1}^{m^*}\frac{a_j^*}{(\hat{w}_k+w_j^*)^n}\\
=&\sum_{r=1}^{i_0}\sum_{i\in\mathcal{I}_{0,r}}\frac{\hat{a}_i}{(\hat{w}_k+\hat{w}_i)^n}+\sum_{j=1}^{m^*}\sum_{i\in\mathcal{I}_j}\frac{\hat{a}_i}{(\hat{w}_k+\hat{w}_i)^n}-\sum_{j=1}^{m^*}\frac{a_j^*}{(\hat{w}_k+w_j^*)^n}\\
=&\sum_{r=1}^{i_0}\frac{\sum_{i\in\mathcal{I}_{0,r}}\hat{a}_i}{(\hat{w}_k+v_r)^n}+\sum_{j=1}^{m^*}\frac{\sum_{i\in\mathcal{I}_j}\hat{a}_i}{(\hat{w}_k+w_j^*)^n}-\sum_{j=1}^{m^*}\frac{a_j^*}{(\hat{w}_k+w_j^*)^n}=0.
\end{aligned}\tag{131}$$

This yields

$$\frac{\partial J_\omega}{\partial a_k}(\hat{a},\hat{w})=-2\Delta_{0,\omega}(\hat{w}_k),\qquad\frac{\partial J_\omega}{\partial w_k}(\hat{a},\hat{w})=2\hat{a}_k\Delta_{1,\omega}(\hat{w}_k),$$

and hence

$$\|\nabla J_\omega(\hat{a},\hat{w})\|_2^2=4\sum_{k=1}^{m}\left[\Delta_{0,\omega}^2(\hat{w}_k)+\hat{a}_k^2\Delta_{1,\omega}^2(\hat{w}_k)\right].$$

By Lemma 27, we get $\lim_{\omega\to0^+}\|\nabla J_\omega(\hat{a},\hat{w})\|_2=0$.

If ρ_0 has the sub-Gaussian tails (103), again by Lemma 27, the estimate

$$|\Delta_{n,\omega}(\hat{w}_k)|\le\Delta_{n,\omega}^+(\hat{w}_k)\le\Delta_{n,\omega}^+(\hat{w}_{\min})\lesssim\omega^{-n}e^{-\hat{w}_{\min}/\omega}\left(c_2^{\hat{w}_{\min}^2}+c_3^{\hat{w}_{\min}}\right)\tag{132}$$

holds for any $n\in\mathbb{N}$, $\omega\in(0,\min\{1/2,1/t_0,2c_1/\hat{w}_{\min}\})$ and $k\in[m]$. Here $\hat{w}_{\min}:=\min_{i\in[m]}\hat{w}_i>0$, $c_2,c_3>1$ are constants only related to c_1,t_0, and $\lesssim$ hides universal positive constants only depending on n and ρ_0, t_0, c_0, c_1. Therefore

$$\|\nabla J_\omega(\hat{a},\hat{w})\|_2\lesssim\sqrt{m}\omega^{-1}e^{-\hat{w}_{\min}/\omega}\left(c_2^{\hat{w}_{\min}^2}+c_3^{\hat{w}_{\min}}\right)(1+\|\hat{a}\|_\infty),\quad\omega\in(0,1].$$

The proof is completed. ■

(4) Eigenvalues of Hessian

Proposition 34 *For any $(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m$ satisfying $\hat{\rho}(t;a,w)\equiv\bar{\rho}(t)$, denote the eigenvalues of $\nabla^2J_\omega(a,w)$ by $\lambda_1(\omega)\ge\lambda_2(\omega)\ge\cdots\ge\lambda_{2m}(\omega)$. If $m>m^*$, we have*

$$\lambda_k(\omega)>0,\quad k=1,2,\cdots,m',\tag{133}$$

$$\lim_{\omega\to0^+}\lambda_k(\omega)=0,\quad k=m'+1,m'+2,\cdots,2m\tag{134}$$

for $\omega > 0$ sufficiently small, where $m' \le 2m^* + |\mathcal{I}_0| \le m + m^*$. In particular, if ρ_0 has the sub-Gaussian tails (103), the estimate

$$|\lambda_k(\omega)| \lesssim \omega^{-2} e^{-w_{min}/\omega} \left(c_2^{w_{min}^2} + c_3^{w_{min}} \right) (1 + \|a\|_\infty) \quad k = m' + 1, m' + 2, \cdots, 2m \tag{135}$$

holds for any $\omega \in (0, \min\{1/2, 1/t_0, 2c_1/w_{min}\})$. Here $w_{min} := \min_{i\in[m]} w_i > 0$, $c_2, c_3 > 1$ are constants only related to c_1, t_0, and $\lesssim$ hides universal positive constants only depending on ρ_0, t_0, c_0, c_1.

Proof A straightforward computation shows, for $k, j = 1, 2, \cdots, m$,

$$\frac{\partial^2 J_\omega}{\partial a_k \partial a_j}(a, w) = \frac{2}{w_k + w_j}, \tag{136}$$

$$\frac{\partial^2 J_\omega}{\partial a_k \partial w_j}(a, w) = \frac{-2a_j}{(w_k + w_j)^2}, \quad k \neq j, \tag{137}$$

$$\frac{\partial^2 J_\omega}{\partial a_k \partial w_k}(a, w) = -2 \left[\sum_{i=1}^{m} \frac{a_i}{(w_k + w_i)^2} - \sum_{j'=1}^{m^*} \frac{a^*_{j'}}{(w_k + w^*_{j'})^2} \right] - \frac{a_k}{2w_k^2} + 2\Delta_{1,\omega}(w_k), \tag{138}$$

$$\frac{\partial^2 J_\omega}{\partial w_k \partial w_j}(a, w) = \frac{4a_k a_j}{(w_k + w_j)^3}, \quad k \neq j, \tag{139}$$

$$\frac{\partial^2 J_\omega}{\partial w_k \partial w_k}(a, w) = 4a_k \left[\sum_{i=1}^{m} \frac{a_i}{(w_k + w_i)^3} - \sum_{j'=1}^{m^*} \frac{a^*_{j'}}{(w_k + w^*_{j'})^3} \right] + \frac{a_k^2}{2w_k^3} - 2a_k \Delta_{2,\omega}(w_k). \tag{140}$$

Fix any $(\hat{a}, \hat{w}) \in \mathbb{R}^m \times \mathbb{R}^m_+$ satisfying $\hat{\rho}(t; \hat{a}, \hat{w}) \equiv \bar{\rho}(t)$. By (131), we have

$$\begin{aligned}
&\frac{\partial^2 J_\omega}{\partial a_k \partial a_j}(\hat{a}, \hat{w}) = \frac{2}{\hat{w}_k + \hat{w}_j}, \\
&\frac{\partial^2 J_\omega}{\partial a_k \partial w_j}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_j}{(\hat{w}_k + \hat{w}_j)^2} \quad (k \neq j), \qquad \frac{\partial^2 J_\omega}{\partial a_k \partial w_k}(\hat{a}, \hat{w}) = -\frac{\hat{a}_k}{2\hat{w}_k^2} + 2\Delta_{1,\omega}(\hat{w}_k), \\
&\frac{\partial^2 J_\omega}{\partial w_k \partial w_j}(\hat{a}, \hat{w}) = \frac{4\hat{a}_k \hat{a}_j}{(\hat{w}_k + \hat{w}_j)^3} \quad (k \neq j), \qquad \frac{\partial^2 J_\omega}{\partial w_k \partial w_k}(\hat{a}, \hat{w}) = \frac{\hat{a}_k^2}{2\hat{w}_k^3} - 2\hat{a}_k \Delta_{2,\omega}(\hat{w}_k).
\end{aligned}$$

Let

$$\begin{aligned}
\bar{J}(a, w) &:= \|\hat{\rho}(t; a, w) - \bar{\rho}(t)\|^2_{L^2[0,\infty)} \\
&= \left\| \sum_{i=1}^{m} a_i e^{-w_i t} - \sum_{j=1}^{m^*} a^*_j e^{-w^*_j t} \right\|^2_{L^2[0,\infty)},
\end{aligned} \tag{141}$$

and

$$\mathcal{E}_\omega(a, w) := \begin{bmatrix} \boldsymbol{O}_{m\times m} & \mathrm{Diag}(\Delta_{1,\omega}(w)) \\ \mathrm{Diag}(\Delta_{1,\omega}(w)) & -\mathrm{Diag}(a \circ \Delta_{2,\omega}(w)) \end{bmatrix},$$

where $\Delta_{n,\omega}(\cdot)$ $(n=1,2)$ is performed element-wisely. One can verify that

$$\nabla^2 J_\omega(a,w) = \nabla^2 \bar{J}(a,w) + 2\mathcal{E}_\omega(a,w). \tag{142}$$

Then we analyze $\nabla^2\bar{J}(\hat{a},\hat{w})$ and $\mathcal{E}_\omega(\hat{a},\hat{w})$ respectively.

(i) $\nabla^2\bar{J}(\hat{a},\hat{w})$. Obviously $(\hat{a},\hat{w})$ is a global minimizer of $\bar{J}(a,w)$ due to $\bar{J}(\hat{a},\hat{w}) = 0$. Hence $\nabla\bar{J}(\hat{a},\hat{w}) = 0$ and $\nabla^2\bar{J}(\hat{a},\hat{w})$ is positive semi-definite. We further show that $\nabla^2\bar{J}(\hat{a},\hat{w})$ has multiple zero eigenvalues when $m > m^*$. In fact, since

$$\frac{\partial^2\bar{J}}{\partial a_k\partial a_j}(\hat{a},\hat{w}) = \frac{2}{\hat{w}_k+\hat{w}_j}, \quad \frac{\partial^2\bar{J}}{\partial a_k\partial w_j}(\hat{a},\hat{w}) = \frac{-2\hat{a}_j}{(\hat{w}_k+\hat{w}_j)^2}, \quad \frac{\partial^2\bar{J}}{\partial w_k\partial w_j}(\hat{a},\hat{w}) = \frac{4\hat{a}_k\hat{a}_j}{(\hat{w}_k+\hat{w}_j)^3},$$

it is straightforward to verify that for any $i,j\in\mathcal{I}_p$, $p\in[m^*]$ and any $i,j\in\mathcal{I}_{0,r}$, $r\in[i_0]$,

$$\nabla^2\bar{J}(\hat{a},\hat{w})^{i,:} = \nabla^2\bar{J}(\hat{a},\hat{w})^{j,:},\ \hat{a}_j\cdot\nabla^2\bar{J}(\hat{a},\hat{w})^{m+i,:} = \hat{a}_i\cdot\nabla^2\bar{J}(\hat{a},\hat{w})^{m+j,:},$$

where $A^{i,:}$ denotes the i-th row of matrix A. Notice that $\sum_{i\in\mathcal{I}_{0,r}}\nabla^2\bar{J}(\hat{a},\hat{w})^{m+i,:} = 0$ for any $r\in[i_0]$, we conclude that the Hessian $\nabla^2\bar{J}(\hat{a},\hat{w})$ has at most $2m^*+i_0+i_2 \le 2m^*+|\mathcal{I}_0| \le m+m^*$ different rows,[10] which yields $\operatorname{rank}(\nabla^2\bar{J}(\hat{a},\hat{w})) \le 2m^*+|\mathcal{I}_0| \le m+m^*$. Therefore, the number of zero eigenvalues of $\nabla^2\bar{J}(\hat{a},\hat{w}) \ge \dim\left\{x\in\mathbb{R}^{2m} : \nabla^2\bar{J}(\hat{a},\hat{w})\cdot x = 0\right\} = 2m - \operatorname{rank}(\nabla^2\bar{J}(\hat{a},\hat{w})) \ge 2(m-m^*) - |\mathcal{I}_0| \ge m-m^*$. Since $\nabla^2\bar{J}(\hat{a},\hat{w})$ is positive semi-definite, all the non-zero eigenvalues must be positive.

(ii) $\mathcal{E}_\omega(\hat{a},\hat{w})$. Let

$$\begin{aligned} G_k^{(1)} &:= \{y\in\mathbb{R} : |y| \le |\Delta_{1,\omega}(\hat{w}_k)|\}, \\ G_k^{(2)} &:= \{y\in\mathbb{R} : |y+\hat{a}_k\Delta_{2,\omega}(\hat{w}_k)| \le |\Delta_{1,\omega}(\hat{w}_k)|\}. \end{aligned}$$

By Gershgorin's circle theorem, for any eigenvalue of $\mathcal{E}_\omega(\hat{a},\hat{w})$, say $\lambda(\omega)$, we have $\lambda(\omega)\in\bigcup_{k=1}^m(G_k^{(1)}\cup G_k^{(2)})$. Combining with Lemma 27, we get

$$|\lambda(\omega)| \le \max_{k\in[m]}\left(|\hat{a}_k||\Delta_{2,\omega}(\hat{w}_k)| + |\Delta_{1,\omega}(\hat{w}_k)|\right) \to 0, \quad \omega\to 0^+. \tag{143}$$

If ρ_0 has the sub-Gaussian tails (103), again by Lemma 27 (similar to (132)), we further have

$$\begin{aligned} |\lambda(\omega)| &\le \max_{k\in[m]}\left(|\hat{a}_k|\Delta_{2,\omega}^+(\hat{w}_k) + \Delta_{1,\omega}^+(\hat{w}_k)\right) \\ &\le \max_{k\in[m]}\left(|\hat{a}_k|\Delta_{2,\omega}^+(\hat{w}_{\min}) + \Delta_{1,\omega}^+(\hat{w}_{\min})\right) \\ &\lesssim \omega^{-2}e^{-\hat{w}_{\min}/\omega}\left(c_2^{\hat{w}_{\min}^2} + c_3^{\hat{w}_{\min}}\right)(1+\|\hat{a}\|_\infty), \quad \omega\in(0,1], \end{aligned} \tag{144}$$

where $\omega\in(0,\min\{1/2,1/t_0,2c_1/\hat{w}_{\min}\})$, $\hat{w}_{\min} := \min_{i\in[m]}\hat{w}_i > 0$, $c_2,c_3>1$ are constants only related to c_1,t_0, and $\lesssim$ hides universal positive constants only depending on ρ_0, t_0, c_0, c_1.

10. Here $i_2 := |\{r\in[i_0] : |\mathcal{I}_{0,r}|\ge 2\}|$. When $\mathcal{I}_0=\varnothing$, the upper bound is $2m^*$; when $\mathcal{I}_0\neq\varnothing$, since $\mathcal{I}_{0,r}\neq\varnothing$ for any $r\in[i_0]$, let $i_1 := |\{r\in[i_0] : |\mathcal{I}_{0,r}|=1\}|$ and i_2 defined as before. Then $i_0 = i_1+i_2$, $|\mathcal{I}_0| = \sum_{r=1}^{i_0}|\mathcal{I}_{0,r}| \ge i_1+2i_2 = i_0+i_2$. The last inequality follows from $|\mathcal{I}_0| = m - \sum_{j=1}^{m^*}|\mathcal{I}_j| \le m-m^*$.

Combining (i), (ii) and applying Weyl's theorem gives the desired result. ■

(5) Local Linearization Analysis

The previous analysis can now be tied directly to a quantitative dynamics via linearization arguments. It is shown that under mild assumptions, the gradient flow (104) can become trapped in plateaus with an *exponentially* large timescale. That is, *the curse of memory* occurs, this time in optimization dynamics instead of approximation rates.

Proof [Proof of Theorem 20] Consider the asymptotic expansion with the form

$$\theta_\omega(\tau) = \theta_\omega^0(\tau) + \sum_{i=1}^{\infty} \delta^i \theta_\omega^i(\tau) = \theta_\omega^0(\tau) + \delta\theta_\omega^1(\tau) + \delta^2\theta_\omega^2(\tau) + o(\delta^2), \tag{145}$$

for some $\delta \in (0,1)$ (with $\delta \ll 1$) and $\theta_\omega^i(\tau) = \mathcal{O}(1)$ ($\tau \geq 0$, $i = 0, 1, \cdots$).[11] For consistency, we have $\theta_\omega^0(0) = \theta_0$ and $\theta_\omega^i(0) = 0$ for $i = 1, 2, \cdots$. By continuity, $\tau_0 > 0$ and $\|\theta_\omega(\tau) - \theta_0\|_2 \leq \delta$ for any $\tau \in [0, \tau_0]$. The aim is to quantify the scale of τ_0.

Let $g_0 := \nabla J_\omega(\theta_0)$ and $H_0 := \nabla^2 J_\omega(\theta_0)$. The local linearization on (104) shows

$$\frac{d}{d\tau}\theta_\omega(\tau) = -g_0 - H_0(\theta_\omega(\tau) - \theta_0) + \mathcal{O}(\delta^2), \quad \tau \in [0, \tau_0].$$

Combining with (145), we have

$$\begin{aligned}
&\frac{d}{d\tau}\theta_\omega^0(\tau) = -g_0 - H_0(\theta_\omega^0(\tau) - \theta_0), \quad \theta_\omega^0(0) = \theta_0, && \text{at } \mathcal{O}(1) \text{ scale},\\
&\frac{d}{d\tau}\theta_\omega^1(\tau) = -H_0\theta_\omega^1(\tau), \quad \theta_\omega^1(0) = 0, && \text{at } \mathcal{O}(\delta) \text{ scale},\\
&\frac{d}{d\tau}\theta_\omega^2(\tau) = -H_0\theta_\omega^2(\tau) + \mathcal{O}(1), \quad \theta_\omega^2(0) = 0, && \text{at } \mathcal{O}(\delta^2) \text{ scale}.
\end{aligned}$$

Therefore

$$\begin{aligned}
\theta_\omega^0(\tau) &= \theta_0 - \left(\int_0^\tau e^{-H_0 s} ds\right) g_0,\\
\theta_\omega^1(\tau) &= e^{-H_0\tau}\theta_\omega^1(0) = 0,
\end{aligned}$$

which gives

$$\theta_\omega(\tau) = \theta_0 - \left(\int_0^\tau e^{-H_0 s} ds\right) g_0 + \mathcal{O}(\delta^2), \quad \tau \in [0, \tau_0]. \tag{146}$$

To achieve a parameter separation gap δ_0, i.e. $\|\theta_\omega(\tau) - \theta_0\|_2 = \delta_0$ with $\delta_0 = c\delta$, $c \in (0, 1]$, we need to take τ such that

$$\left\| \left(\int_0^\tau e^{-H_0 s} ds\right) g_0 \right\|_2 \geq \frac{\delta_0}{2}. \tag{147}$$

11. Here $\theta_\omega^i(\tau)$ denotes the i-th term in the asymptotic expansion of $\theta_\omega(\tau)$, not the i-th power.

Let $H_0 = P^\top \Lambda P$ be the eigenvalue decomposition with P orthogonal and Λ diagonal consisting of the eigenvalues of H_0 (i.e. $\Lambda = \mathrm{diag}(\lambda_1, \cdots \lambda_{2m})$ with $\lambda_1 \geq \cdots \geq \lambda_{2m}$). Then

$$\begin{aligned}\left\| \left(\int_0^\tau e^{-H_0 s} ds \right) g_0 \right\|_2 = \left\| P^\top \left(\int_0^\tau e^{-\Lambda s} ds \right) P g_0 \right\|_2 &\leq \left\| \int_0^\tau e^{-\Lambda s} ds \right\|_2 \|g_0\|_2 \\ &\leq \|g_0\|_2 \cdot \max\left\{ \max_{i\in[2m], \lambda_i \neq 0} \frac{1}{|\lambda_i|} |e^{-\lambda_i \tau} - 1|,\ \tau \right\}.\end{aligned}$$

It is straightforward to verify that $h(\tau; \lambda) := \frac{1}{|\lambda|}|e^{-\lambda\tau} - 1|$, $\tau \geq 0$ monotonically decreases on $\lambda \in \mathbb{R}$ for any $\tau \geq 0$.[12] Hence

$$\left\| \left(\int_0^\tau e^{-H_0 s} ds \right) g_0 \right\|_2 \leq \|g_0\|_2 \cdot \begin{cases} \frac{1}{-\lambda_{2m}}(e^{-\lambda_{2m}\tau} - 1), & \lambda_{2m} < 0, \\ \tau, & \lambda_{2m} \geq 0, \end{cases} \tag{148}$$

and the right-hand side monotonically increases on $\tau \geq 0$. Combining (147), (148) gives

$$\frac{\delta_0}{2} \leq \|g_0\|_2 \cdot \begin{cases} \frac{1}{-\lambda_{2m}}(e^{-\lambda_{2m}\tau} - 1), & \lambda_{2m} < 0, \\ \tau, & \lambda_{2m} \geq 0. \end{cases} \tag{149}$$

We discuss for different cases:

(i) $\|g_0\|_2 = 0$. Obviously the inequality (149) fails since (147) fails for any $\tau \geq 0$, which gives $\tau_0 = +\infty$;

(ii) $\|g_0\|_2 \neq 0$ and $\lambda_{2m} \geq 0$. By (149), we get $\tau \geq \frac{\delta_0}{2\|g_0\|_2}$;

(iii) $\|g_0\|_2 \neq 0$ and $\lambda_{2m} < 0$. By (149), we get

$$\tau \geq \frac{1}{-\lambda_{2m}} \ln\left(1 + \delta_0 \frac{-\lambda_{2m}}{2\|g_0\|_2}\right).$$

If $-\lambda_{2m} \leq 2\|g_0\|_2$, we have

$$\begin{aligned}\tau &\geq \frac{1}{-\lambda_{2m}} \cdot \delta_0 \frac{-\lambda_{2m}}{2\|g_0\|_2} \cdot \frac{\ln\left(1 + \delta_0 \frac{-\lambda_{2m}}{2\|g_0\|_2}\right)}{\delta_0 \frac{-\lambda_{2m}}{2\|g_0\|_2}} \\ &= \frac{\delta_0}{2\|g_0\|_2}\left(1 + \mathcal{O}\left(\delta_0 \frac{-\lambda_{2m}}{2\|g_0\|_2}\right)\right) = \frac{\delta_0}{2\|g_0\|_2}(1 + \mathcal{O}(\delta_0));\end{aligned}$$

if $-\lambda_{2m} > 2\|g_0\|_2$, we have $\tau \geq \frac{\ln(1+\delta_0)}{-\lambda_{2m}}$.

Combining (i), (ii), (iii) gives

$$\tau_0 = \tau_0(\delta; \omega, m, \theta_0) \gtrsim \min\left\{ \frac{\delta}{\|g_0\|_2}, \frac{\ln(1+\delta)}{|\lambda_{2m}|} \right\}. \tag{150}$$

Let the initialization satisfy $\hat{\rho}(t; \theta_0) \equiv \bar{\rho}(t)$, and assume $m > m^*$. According to Proposition 33 and Proposition 34, we have

$$\lim_{\omega \to 0^+} \|g_0\|_2 = 0, \quad \lim_{\omega \to 0^+} \lambda_{2m} = 0 \Rightarrow \lim_{\omega \to 0^+} \tau_0(\delta; \omega, m, \theta_0) = +\infty. \tag{151}$$

12. With the convention that $h(\tau; 0) = \tau$ for any $\tau > 0$, and $h(0; \lambda) \equiv 0$ for any $\lambda \in \mathbb{R}$.

If ρ_0 has the sub-Gaussian tails (103), again by Proposition 33 and Proposition 34, we further have

$$\tau_0(\delta;\omega,m,\theta_0) \gtrsim \omega^2 e^{w_{0,\min}/\omega} \min\left\{\frac{\delta}{\sqrt{m}}, \ln(1+\delta)\right\} \frac{1}{\left(c_2^{w_{0,\min}^2} + c_3^{w_{0,\min}}\right)(1+\|a_0\|_\infty)} \tag{152}$$

for any $\omega \in (0, \min\{1/2, 1/t_0, 2c_1/w_{0,\min}\})$, where $w_{0,\min} := \min_{i\in[m]} w_{0,i} > 0$, $c_2, c_3 > 1$ are constants only related to c_1, t_0, and $\gtrsim$ hides universal positive constants only depending on ρ_0, t_0, c_0, c_1. Since the initialization is bounded as $(a_0, w_0) \in [a_l^0, a_r^0]^m \times [w_l^0, w_r^0]^m$ with $a_l^0 < a_r^0$, $0 < w_l^0 < w_r^0$, let $c_a^0 = \max\{|a_l^0|, |a_r^0|\}$, we get

$$\begin{aligned}\tau_0(\delta;\omega,m,\theta_0) &\gtrsim \omega^2 e^{w_l^0/\omega} \min\left\{\frac{\delta}{\sqrt{m}}, \ln(1+\delta)\right\} \frac{1}{\left(c_2^{(w_r^0)^2} + c_3^{w_r^0}\right)(1+c_a^0)} \\ &\gtrsim \omega^2 e^{w_l^0/\omega} \min\left\{\frac{\delta}{\sqrt{m}}, \ln(1+\delta)\right\}, \end{aligned} \tag{153}$$

where $\gtrsim$ hides universal positive constants only related to w_r^0, a_l^0 and a_r^0.

The last task is to show the dynamics of loss is much slower than the parameter separation when there are slow downs. The argument is trivial since for any $\tau \in [0, \tau_0]$,

$$\begin{aligned} J_\omega(\theta_\omega(\tau)) - J_\omega(\theta_0) &= g_0^\top(\theta_\omega(\tau) - \theta_0) + (\theta_\omega(\tau) - \theta_0)^\top H_0(\theta_\omega(\tau) - \theta_0) + o(\delta^2) \\ &\geq -\|g_0\|_2 \|\theta_\omega(\tau) - \theta_0\|_2 + \lambda_{2m}\|\theta_\omega(\tau) - \theta_0\|_2^2 + o(\delta^2) \\ &= o(1)\mathcal{O}(\delta) + o(1)\mathcal{O}(\delta^2) + o(\delta^2) \\ &= o(\delta^2), \quad \omega \to 0^+. \end{aligned}$$

By continuity, the proof is completed. ■

Remark 35 *The estimate in Theorem 20 shows a lower bound on the escape time, hence it does not appear to preclude the situation that the plateauing lasts forever. However, in the proof above, if one supposes $\tau_0 = +\infty$ in (105), i.e. the hypothetical situation where the parameters are trapped forever, and write $\tilde{g}_0 := Pg_0 = (\tilde{g}_{0,1}, \cdots, \tilde{g}_{0,2m})$, we have*

$$\left\|\left(\int_0^\tau e^{-H_0 s} ds\right) g_0\right\|_2^2 = \tilde{g}_0^\top \left(\int_0^\tau e^{-\Lambda s} ds\right)^2 \tilde{g}_0 = \sum_{i=1}^{2m} (\tilde{g}_{0,i})^2 (h(\tau;\lambda_i))^2 \geq (\tilde{g}_{0,j})^2 (h(\tau;\lambda_j))^2$$

for any j such that $\lambda_j < 0$. If $\tilde{g}_{0,j} \neq 0$, (146) gives

$$\begin{aligned} \|\theta_\omega(\tau) - \theta_0\|_2 &\geq \left\|\left(\int_0^\tau e^{-H_0 s} ds\right) g_0\right\|_2 + \mathcal{O}(\delta^2) \\ &\geq \frac{|\tilde{g}_{0,j}|}{-\lambda_j}(e^{-\lambda_j \tau} - 1) + \mathcal{O}(\delta^2) \to +\infty, \quad \tau \to \infty, \end{aligned}$$

which is a contradiction. That is to say, the parameter separation has to achieve the gap δ within a finite time, even if it is exponentially large.

Remark 36 *Recall Lemma 30, $\hat{\rho}(t; a_0, w_0) \equiv \bar{\rho}(t)$ if and only if $(a_0, w_0) \in \mathcal{M}^* = \bigcup_{\mathcal{P}} \mathcal{M}^*_{\mathcal{P}}$, where $\mathcal{P}$ is a partition over $[m]$ as defined in Definition 29. That is, as a union of affine spaces, $\mathcal{M}^*$ is in fact an equivalent set for qualified initializations. As discussed in Remark 31, when there is no degeneracy, the cardinality of $\mathcal{M}^*$ is $m^*! \left\{ {m \atop m^*} \right\}$ (i.e. the number of $\mathcal{P}$), with each $\mathcal{M}^*_{\mathcal{P}}$ an $(m-m^*)$-dimensional affine space; when there is degeneracy in some $\mathcal{M}^*_{\mathcal{P}}$, it then becoms an uncountable set. Certainly, initializations sufficiently near $\mathcal{M}^*$ are also qualified by continuity.*

Remark 37 *Motivated by the idea of weights degeneracy (Definition 29), we can further apply similar methods to a global landscape analysis on the loss function J_ω. The results there show that the plateaus are all over the landscape, even provided general targets without memory structures. See details in Appendix A.*

7.2 Extensions on Plateauing Time

Theorem 23 is proved in this section, which can be viewed as an extension to Theorem 20 in the following aspects:

- Recall $\bar{\rho}(t) = \int_{w_l}^{w_r} a^*(w) e^{-wt} d\pi_0(w)$. Theorem 23 considers general π_0, while Theorem 20 only takes the discrete one;
- More importantly, the condition that $\hat{\rho} \approx \bar{\rho}$ is quantified in the estimates of Theorem 23, while it is only qualified in Theorem 20.

The basic insight behind the proof of Theorem 23 is the same as that of Theorem 20 (see Section 7.1). That is, by adding long-term memories in targets, one can make little difference on the gradient and Hessian of the loss function, but affects the loss value to a large extent. This leads to a significant slow down of the gradient flow training dynamics near the short-term memory part of target.

However, when it comes to techniques of analysis, although both Theorem 20 and Theorem 23 are subsequently proved following the same procedure of "landscape analysis": i) show the large loss; ii) show small gradients; iii) show small eigenvalues of Hessian; iv) apply the local linearization argument to give a quantitative timescale on slow downs of the training dynamics, in fact i), ii), iii) are respectively performed under different regimes. That is, for Theorem 20, we firstly assume the condition $\hat{\rho} \equiv \bar{\rho}$ (at initialization) and then derive a fundamental relation on parameters (Lemma 30), which is used through the following analysis, and the conclusion is finally extended to $\hat{\rho} \approx \bar{\rho}$ by continuity, which is related to $\bar{\epsilon}$ in a *qualified* sense (where $\|\hat{\rho} - \bar{\rho}\|^2 \leq \bar{\epsilon}$). While in Theorem 23, to show a *quantitative* dependence on $\bar{\epsilon}$, obviously the parameter relation (Lemma 30) does not hold,[13] hence we are forced to directly work with the partial loss $\bar{\epsilon}$.

The analysis in this section follows the same organization as Section 7.1.

(1) Preliminary Results

We begin by a uniform upper bound on the initial loss.

13. In fact, it can be numerically observed that when $\bar{\rho}(t) = {a^*}^\top e^{-w^* t}$ (i.e. take a discrete distribution π_0), the model parameters (a, w) can be far away from the ground truth (a^*, w^*) even when $\hat{\rho}(t) = a^\top e^{-wt} \approx \bar{\rho}(t)$. That is to say, it is reasonable (and also direct) to investigate the partial loss (i.e. $\bar{\epsilon}$) instead of relations on parameters.

Lemma 38 (Bounded Initial Loss) *For any $\omega > 0$, $m \in \mathbb{N}_+$ and $(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m$, we have*

$$\sqrt{J_\omega(a,w)} \le \frac{\|a\|_1}{\sqrt{2w_{\min}}} + \frac{c^*}{\sqrt{2w_l}} + \|\rho_0\|_{L^2(\mathbb{R})}, \tag{154}$$

where $w_{\min} := \min_{i\in[m]} w_i > 0$, $c^* := \|a^*\|_{L^\infty[w_l,w_r]}$. Recall the bounded initialization (110), we have

$$\sqrt{J_\omega(\theta_0)} \le \frac{c_a^0}{\sqrt{2w_l^0}} + \frac{c^*}{\sqrt{2w_l}} + \|\rho_0\|_{L^2(\mathbb{R})}, \quad \theta_0 \in \Theta_0, \tag{155}$$

where $c_a^0 := \max\{|a_l^0|, |a_r^0|\}$.

Proof We have $|\bar\rho(t)| = \left|\int_{w_l}^{w_r} a^*(w)e^{-wt}d\pi_0(w)\right| \le c^* e^{-w_l t}$, then $\|\bar\rho\|_{L^2[0,\infty)} \le c^*/\sqrt{2w_l}$. Notice that $\|\rho_{0,\omega}\|_{L^2[0,\infty)} = \|\rho_0(t-1/\omega)\|_{L^2[0,\infty)} \le \|\rho_0\|_{L^2(\mathbb{R})}$, we have

$$\begin{aligned}
\sqrt{J_\omega(a,w)} &= \left\| \sum_{i=1}^m a_i e^{-w_i t} - \bar\rho(t) - \rho_{0,\omega}(t) \right\|_{L^2[0,\infty)} \\
&\le \sum_{i=1}^m |a_i| \left\| e^{-w_i t} \right\|_{L^2[0,\infty)} + \|\bar\rho(t)\|_{L^2[0,\infty)} + \|\rho_{0,\omega}(t)\|_{L^2[0,\infty)} \\
&= \sum_{i=1}^m |a_i| \frac{1}{\sqrt{2w_i}} + \|\bar\rho(t)\|_{L^2[0,\infty)} + \|\rho_0(t-1/\omega)\|_{L^2[0,\infty)} \\
&\le \frac{\|a\|_1}{\sqrt{2w_{\min}}} + \frac{c^*}{\sqrt{2w_l}} + \|\rho_0\|_{L^2(\mathbb{R})}.
\end{aligned}$$

For any $\theta_0 \in \Theta_0$, $w_0 \succeq w_l^0 \mathbf{1}_m$ and $\|a_0\|_1 = m^{-\beta}\|a_0'\|_1 \le m^{1-\beta} c_a^0 \le c_a^0$, which gives the desired conclusion. The proof is completed. ■

Remark 39 *Lemma 38 shows that to ensure a normal loss at initialization with respect to the model capacity (i.e. the width m), one has to take a particular scaling on the outer parameters. That is, it is necessary to take $\|a_0\|_1 = \mathcal{O}(1)$ for any $a_0 \in \mathbb{R}^m$, $m \in \mathbb{N}_+$. In addition, the initial loss is bounded uniformly in any $\omega > 0$ since $\|\rho_{0,\omega}\|_{L^2[0,\infty)} \le \|\rho_0\|_{L^2(\mathbb{R})}$.*

Then we prove the non-degeneracy property for the dynamics of parameters under the gradient flow (104). It is used throughout the subsequent analysis.

Lemma 40 (Training Stability) *Let $J^0 := \frac{c_a^0}{\sqrt{2w_l^0}} + \frac{c^*}{\sqrt{2w_l}} + \|\rho_0\|_{L^2(\mathbb{R})}$. Define the Cauchy problem*

$$v'(\tau) = \frac{1}{2} J^0 \left(c_a^0 + \sqrt{2} J^0 \int_0^\tau v(s) ds \right) v^6(\tau), \quad v(0) = \frac{1.1}{\sqrt{w_l^0}}. \tag{156}$$

Let $\tau_1 := \inf\left\{\tau \geq 0 : v(\tau) > \sqrt{\frac{2}{w_l^0}}\right\}$. *Then for any* $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \Theta_0$ *and* $\tau \in [0, \tau_1]$, *we have*

$$w_{\omega,k}(\tau) \geq \frac{w_l^0}{2} > 0, \quad k = 1, 2, \cdots, m. \tag{157}$$

Proof For any $\omega > 0$, $m \in \mathbb{N}_+$ and $\theta_0 \in \Theta_0$, let

$$\tau_k^+ = \tau_k^+(\omega, m, \theta_0) := \inf\left\{\tau \geq 0 : w_{\omega,k}(\tau) \leq 0\right\}, \quad k = 1, 2, \cdots, m,$$

and $\tau_k^+ := +\infty \Leftrightarrow w_{\omega,k}(\tau) > 0$ for any $\tau \geq 0$. Then by continuity, for any $\omega > 0$, $m \in \mathbb{N}_+$ and $\theta_0 \in \Theta_0$, $\tau_k^+ \in (0, +\infty]$, $w_{\omega,k}(\tau_k^+) = 0$,[14] and $w_{\omega,k}(\tau) > 0$ for any $\tau \in [0, \tau_k^+)$, $k = 1, 2, \cdots, m$. Assume that the conclusion does not hold, i.e. there exists $\omega' > 0$, $m' \in \mathbb{N}_+$, $\theta_0' \in \Theta_0$ and $\tau' \in [0, \tau_1]$, such that $w_{\omega',k'}(\tau') < w_l^0/2$ for some $k' \in [m']$. Write the corresponding solution to (104) by $\theta(\tau) = (a(\tau), w(\tau))$.[15] A straightforward computation shows, for $k = 1, 2, \cdots, m$,

$$\frac{\partial J_\omega}{\partial a_k}(a, w) = 2\int_0^\infty e^{-w_k t}\left(\sum_{i=1}^m a_i e^{-w_i t} - \rho_\omega(t)\right) dt, \tag{158}$$

$$\frac{\partial J_\omega}{\partial w_k}(a, w) = -2a_k\int_0^\infty t e^{-w_k t}\left(\sum_{i=1}^m a_i e^{-w_i t} - \rho_\omega(t)\right) dt. \tag{159}$$

By the gradient flow (104), for any $\tau \in [0, \tau_{k'}^+)$, $w_{k'}(\tau) > 0$, then

$$\begin{aligned}
|a_{k'}'(\tau)| &= 2\left|\int_0^\infty e^{-w_{k'}(\tau)t}\left(\sum_{i=1}^{m'} a_i(\tau)e^{-w_i(\tau)t} - \rho_\omega(t)\right) dt\right| \\
&\leq 2\left\{\int_0^\infty e^{-2w_{k'}(\tau)t} dt\right\}^{\frac{1}{2}}\left\{\int_0^\infty\left(\sum_{i=1}^{m'} a_i(\tau)e^{-w_i(\tau)t} - \rho_\omega(t)\right)^2 dt\right\}^{\frac{1}{2}} \\
&= 2\frac{1}{\sqrt{2w_{k'}(\tau)}}\sqrt{J_{\omega'}(\theta(\tau))} \leq \sqrt{2J_{\omega'}(\theta_0')}\frac{1}{\sqrt{w_{k'}(\tau)}} \leq \sqrt{2}J^0\frac{1}{\sqrt{w_{k'}(\tau)}},
\end{aligned} \tag{160}$$

where the Cauchy–Schwartz inequality and the monotonicity of gradient flow is used, and the last inequality is due to Lemma 38. Similarly, for any $\tau \in [0, \tau_{k'}^+)$, we have

$$|w_{k'}'(\tau)| \leq J^0\frac{|a_{k'}(\tau)|}{\sqrt{w_{k'}^3(\tau)}}. \tag{161}$$

Combining (160) and (161) gives

$$|w_{k'}'(\tau)| \leq J^0\left(c_a^0 + \sqrt{2}J^0\int_0^\tau \frac{ds}{\sqrt{w_{k'}(s)}}\right)\frac{1}{\sqrt{w_{k'}^3(\tau)}}. \tag{162}$$

14. Certainly, $w_{\omega,k}(\tau_k^+) = 0$ implies $a_{\omega,k}(\tau_k^+) = 0$, otherwise the loss will blow up.
15. Generally, the solution to (104) is related to ω (target memory), m (model capacity) and θ_0 (initialization). Here they are all omitted since fixed.

Let $\hat{v}(\tau) := 1/\sqrt{w_{k'}(\tau)}$ for any $\tau \in [0, \tau_{k'}^+)$, we get $w_{k'}(\tau) = 1/\hat{v}^2(\tau)$, $w'_{k'}(\tau) = -2/\hat{v}^3(\tau) \cdot \hat{v}'(\tau)$, hence

$$\begin{aligned} |\hat{v}'(\tau)| &= \frac{1}{2}\hat{v}^3(\tau)|w'_{k'}(\tau)| \\ &\le \frac{1}{2}J^0\left(c_a^0 + \sqrt{2}J^0\int_0^\tau \hat{v}(s)ds\right)\hat{v}^6(\tau). \end{aligned} \tag{163}$$

Now we make the comparison between (156) and (163). At the initialization, we have $\hat{v}(0) = \frac{1}{\sqrt{w_{k'}(0)}} \le \frac{1}{\sqrt{w_l^0}} < v(0)$, and

$$|\hat{v}'(0)| \le \frac{1}{2}J^0 c_a^0 \frac{1}{w_{k'}^3(0)} \le \frac{1}{2}J^0 c_a^0 \frac{1}{(w_l^0)^3} < \frac{1}{2}J^0 c_a^0 \frac{1.1^6}{(w_l^0)^3} = v'(0).$$

If we assume that $\hat{v}(s) < v(s)$ for any $s \le \tau \in [0, \tau_{k'}^+)$, then

$$|\hat{v}'(\tau)| < \frac{1}{2}J^0\left(c_a^0 + \sqrt{2}J^0\int_0^\tau v(s)ds\right)v^6(\tau) = v'(\tau).$$

As a result, for any $\tau \in [0, \tau_{k'}^+)$, $\hat{v}(\tau) \le v(\tau)$, which gives $w_{k'}(\tau) \ge 1/v^2(\tau)$. By the definition of τ_1, $v(\tau) \le \sqrt{\frac{2}{w_l^0}}$ for any $\tau \in [0, \tau_1]$. If $\tau_{k'}^+ \le \tau_1$, we get $w_{k'}(\tau) \ge 1/(2/w_l^0) = w_l^0/2$ for any $\tau \in [0, \tau_{k'}^+) \subset [0, \tau_1]$, which is contradictory with the fact that $w_{k'}(\tau_{k'}^+) = 0$ by continuity; if $\tau_1 < \tau_{k'}^+$, we get $w_{k'}(\tau) \ge 1/(2/w_l^0) = w_l^0/2$ for any $\tau \in [0, \tau_1] \subset [0, \tau_{k'}^+)$, which is also contradictory with the hypothesis. The proof is completed. ■

Remark 41 *Lemma 40 gives a "control" dynamical system (156) on the dynamics of the "feature" parameter w of the gradient flow (104). It is shown that the training dynamics is stable (i.e. $w_\omega(\tau) \succeq c\mathbf{1}_m \succ 0$) regardless of the target memory $1/\omega$, the model capacity (width m) and the initialization (specific θ_0), at least within a bounded time.*

The following lemma gives estimates for the change of parameters.

Lemma 42 (Dynamics of Parameters) *For any $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \Theta_0$ and $\tau \in [0, \tau_1]$, we have*

$$\|a_\omega(\tau) - a_\omega(0)\|_2 \lesssim \tau\sqrt{m}, \tag{164}$$

$$\|w_\omega(\tau) - w_\omega(0)\|_2 \lesssim \tau\left(\frac{1}{\sqrt{m}} + \tau\sqrt{m}\right), \tag{165}$$

where $\lesssim$ hides universal positive constants only depending on a_l^0, a_r^0, w_l^0, w_l, w_r and a^*, ρ_0.

Proof According to Lemma 40 (similar to (160) and (161)), for any $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \Theta_0$, $\tau \in [0, \tau_1]$ and $k = 1, \cdots, m$,

$$|a'_{\omega,k}(\tau)| \le \sqrt{2}J^0\frac{1}{\sqrt{w_{\omega,k}(\tau)}} \le \frac{2J^0}{\sqrt{w_l^0}}, \tag{166}$$

$$|w'_{\omega,k}(\tau)| \le J^0\frac{|a_{\omega,k}(\tau)|}{\sqrt{w_{\omega,k}^3(\tau)}} \le 2\sqrt{2}J^0\frac{|a_{\omega,k}(\tau)|}{\sqrt{(w_l^0)^3}}, \tag{167}$$

which gives

$$|a_{\omega,k}(\tau) - a_{\omega,k}(0)| \leq \frac{2J^0}{\sqrt{w_l^0}}\tau, \tag{168}$$

$$|w_{\omega,k}(\tau) - w_{\omega,k}(0)| \leq 2\sqrt{2}\left(\left(\frac{J^0}{w_l^0}\right)^2\tau^2 + \frac{J^0}{\sqrt{(w_l^0)^3}}|a_{\omega,k}(0)|\tau\right), \tag{169}$$

hence

$$\|a_\omega(\tau) - a_\omega(0)\|_2 \leq C\tau\sqrt{m}, \tag{170}$$

$$\|w_\omega(\tau) - w_\omega(0)\|_2 \leq C\tau(\|a_\omega(0)\|_2 + \tau\sqrt{m}), \tag{171}$$

where $C := C(J^0, w_l^0) > 0$ is some universal constant only related to J^0 and w_l^0. The proof is completed. ■

Remark 43 *Given that the network width $m \gg 1$. Comparing (164) with (165) implies that the dynamics of the outer weights a is much faster than the inner weights (or features) w within the timescale $\tau = o(1)$. In other words, a is the fast variable, while w is the slow variable. It appears in the typical training dynamics of the over-parameterized one-hidden-layer feed-forward neural networks.*[16]

Now we get down to perform the "landscape analysis". Motivated by the proof of Proposition 34, we define the corresponding optimization problem to the short-term memory part of target

$$\min_{(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m} \bar{J}(a,w) := \|\hat{\rho}(t;a,w) - \bar{\rho}(t)\|_{L^2[0,\infty)}^2, \tag{172}$$

the same as (141). We will subsequently show that (172) can be viewed as a good reference to the original optimization problem (101). Recall the assumption on dynamics of the loss in Theorem 23.

Conditions on Loss Dynamics. For any $\omega > 0$, $m \in \mathbb{N}_+$ and $\theta_0 \in \Theta_0$, assume

$$\bar{J}(\theta_\omega(\tau)) = \left\|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right\|_{L^2[0,\infty)}^2 \leq \bar{c}\bar{\epsilon}, \quad \forall\tau \in [\bar{\tau}/2, \bar{\tau}] \tag{173}$$

with $\bar{\epsilon} = \bar{\epsilon}(\omega, m)$, $\bar{\tau} = \bar{\tau}(\omega, m)$, and $\bar{c} > 0$ is a universal constant independent of ω, m and θ_0. That is, the gradient flow training dynamics (104) achieves an error tolerance $\bar{\epsilon}$ to the short-term memory part $\bar{\rho}(\cdot)$ within a timescale $\mathcal{O}(\bar{\tau})$.

(2) Loss

16. This is reasonable since the gradient flow training dynamics (104) aims to optimize the problem (101), which is in fact a population risk defined in the classical supervised learning, with the target possessing memories and the model to be one-hidden-layer feed-forward neural networks with the negative exponential activations.

Proposition 44 *There exist universal constants $C(\rho_0), C'(\rho_0) > 0$ only depending on ρ_0, such that for any $\omega \in (0, C'(\rho_0))$, $m \in \mathbb{N}_+$ with $\bar{\epsilon} = \bar{\epsilon}(\omega, m) \le \frac{C(\rho_0)}{4\bar{c}}$ and any $\theta_0 \in \Theta_0$, we have*

$$J_\omega(\theta_\omega(\tau)) \gtrsim C(\rho_0) > 0, \quad \forall \tau \in [\bar{\tau}/2, \bar{\tau}]. \tag{174}$$

That is, the loss is lower bounded away from zero uniformly in sufficiently small $\omega > 0$.

Proof Recall the proof of Proposition 32 up to the estimate (126), we have

$$\|\rho_{0,\omega}\|^2_{L^2[0,\infty)} \ge C(\rho_0) > 0, \quad \forall \omega \in (0, C'(\rho_0)),$$

where $C(\rho_0) = \delta_0 |\rho_0(t_1)|^2/2$ and $C'(\rho_0) = 1/|t_1 - \delta_0|$ the same as Proposition 32. By (173) and the assumption, for any $\omega \in (0, C'(\rho_0))$, $m \in \mathbb{N}_+$ such that $\bar{\epsilon} = \bar{\epsilon}(\omega, m) \le \frac{C(\rho_0)}{4\bar{c}}$ and any $\theta_0 \in \Theta_0$, $\bar{J}(\theta_\omega(\tau)) \le C(\rho_0)/4$ for any $\tau \in [\bar{\tau}/2, \bar{\tau}]$, which gives

$$\begin{aligned}
\sqrt{J_\omega(\theta_\omega(\tau))} &= \left\| [a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t) - \rho_{0,\omega}(t) \right\|_{L^2[0,\infty)} \\
&\ge \left| \sqrt{\bar{J}(\theta_\omega(\tau))} - \|\rho_{0,\omega}\|_{L^2[0,\infty)} \right| \\
&\ge \sqrt{C(\rho_0)} - \frac{1}{2}\sqrt{C(\rho_0)} = \frac{1}{2}\sqrt{C(\rho_0)} > 0,
\end{aligned}$$

which completes the proof. ∎

(3) Gradient

Proposition 45 *For any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l^0\})$, $m \in \mathbb{N}_+$ such that $\bar{\tau} = \bar{\tau}(\omega, m) \le \tau_1$ and any $\theta_0 \in \Theta_0$, there exists $\tau' = \tau'(\omega, m, \theta_0) \in [\bar{\tau}/2, \bar{\tau}]$, such that*

$$\|\nabla J_\omega(\theta_\omega(\tau'))\|_2^2 \lesssim \begin{cases} m\bar{\epsilon} + m\omega^{-2} e^{-\frac{w_l^0}{\omega}}, & \bar{\tau} = 0, \\ \bar{\epsilon}/\bar{\tau} + m\omega^{-2} e^{-\frac{w_l^0}{2\omega}}, & \bar{\tau} \ne 0, \end{cases} \tag{175}$$

where $\lesssim$ hides universal positive constants only depending on a_l^0, a_r^0, w_l^0, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1.

Proof By (158) and (159), for any $k = 1, 2, \cdots, m$, we have

$$\frac{\partial J_\omega}{\partial a_k}(a, w) = 2\int_0^\infty e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \bar{\rho}(t) \right) dt - 2\Delta_{0,\omega}(w_k), \tag{176}$$

$$\frac{\partial J_\omega}{\partial w_k}(a, w) = -2a_k \int_0^\infty t e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \bar{\rho}(t) \right) dt + 2a_k \Delta_{1,\omega}(w_k). \tag{177}$$

That is,

$$\nabla J_\omega(a, w) = \nabla \bar{J}(a, w) + \mathcal{E}_\omega(a, w), \tag{178}$$

where $\mathcal{E}_\omega(a,w) := 2[-\Delta_{0,\omega}(w), a\circ\Delta_{1,\omega}(w)]^\top$ in an element-wise sense. Then (104) can be rewritten as

$$\frac{d}{d\tau}\theta_\omega(\tau) = -\nabla\bar{J}(\theta_\omega(\tau)) - \mathcal{E}_\omega(\theta_\omega(\tau)), \quad \theta_\omega(0)=\theta_0, \tag{179}$$

hence

$$\frac{d}{d\tau}\bar{J}(\theta_\omega(\tau)) = \nabla\bar{J}(\theta_\omega(\tau))\frac{d}{d\tau}\theta_\omega(\tau) = -\|\nabla\bar{J}(\theta_\omega(\tau))\|_2^2 - \left[\nabla\bar{J}(\theta_\omega(\tau))\right]^\top\mathcal{E}_\omega(\theta_\omega(\tau)). \tag{180}$$

According to Lemma 40, for any $\omega>0$, $m\in\mathbb{N}_+$, $\theta_0\in\Theta_0$ and $\tau\in[0,\tau_1]$, we have $w_{\omega,k}(\tau)\geq w_l^0/2>0$, $k=1,2,\cdots,m$. Combining with Lemma 27 (similar to (132)), we have the estimste

$$|\Delta_{n,\omega}(w_{\omega,k}(\tau))| \leq \Delta_{n,\omega}^+(w_{\omega,k}(\tau)) \leq \Delta_{n,\omega}^+(w_l^0/2) \lesssim \omega^{-n}e^{-\frac{w_l^0}{2\omega}}\left(c_2^{(w_l^0)^2/4}+c_3^{w_l^0/2}\right) \tag{181}$$

holds for any $\omega\in(0,\min\{1/2,1/t_0,4c_1/w_l^0\})$, where $c_2,c_3>1$ are constants only depending on c_1,t_0, and $\lesssim$ hides universal positive constants only related to n and ρ_0, t_0, c_0, c_1. Therefore

$$\begin{aligned}
\|\mathcal{E}_\omega(\theta_\omega(\tau))\|_2^2 &= 4\sum_{k=1}^m|\Delta_{0,\omega}(w_{\omega,k}(\tau))|^2 + 4\sum_{k=1}^m|a_{\omega,k}(\tau)|^2|\Delta_{1,\omega}(w_{\omega,k}(\tau))|^2\\
&\lesssim \sum_{k=1}^m\left[1+|a_{\omega,k}(\tau)|^2\omega^{-2}\right]e^{-\frac{w_l^0}{\omega}}\left(c_2^{(w_l^0)^2/4}+c_3^{w_l^0/2}\right)^2\\
&\lesssim \left(m+\|a_\omega(\tau)\|_2^2\omega^{-2}\right)e^{-\frac{w_l^0}{\omega}},
\end{aligned}$$

where $c_2,c_3>1$ and $w_l^0>0$ are also hided. According to Lemma 42, for any $\omega>0$, $m\in\mathbb{N}_+$, $\theta_0\in\Theta_0$ and $\tau\in[0,\tau_1]$, we have $\|a_\omega(\tau)-a_0\|_2\lesssim\tau\sqrt{m}$, hence

$$\|a_\omega(\tau)\|_2 \lesssim \|a_0\|_2+\tau_1\sqrt{m}\lesssim\sqrt{m}, \tag{182}$$

where $\lesssim$ hides universal positive constants only related to c_a^0 and τ_1. Therefore

$$\|\mathcal{E}_\omega(\theta_\omega(\tau))\|_2 \lesssim \sqrt{m}\omega^{-1}e^{-\frac{w_l^0}{2\omega}}, \quad \forall\omega\in(0,\min\{1/2,1/t_0,4c_1/w_l^0\}). \tag{183}$$

According to Lemma 40, for any $\omega > 0$, $m \in \mathbb{N}_+$ such that $\bar{\tau} = \bar{\tau}(\omega, m) \leq \tau_1$ and any $\theta_0 \in \Theta_0$, $\tau \in [\bar{\tau}/2, \bar{\tau}]$, by (173) we have

$$\begin{aligned}
\left|\frac{\partial \bar{J}}{\partial a_k}(a_\omega(\tau), w_\omega(\tau))\right| &\leq 2\int_0^\infty e^{-w_{\omega,k}(\tau)t}\left|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right| dt \\
&\lesssim \frac{1}{\sqrt{w_{\omega,k}(\tau)}}\left\|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right\|_{L^2[0,\infty)} \\
&\lesssim \frac{\sqrt{\bar{\epsilon}}}{\sqrt{w_l^0}} \lesssim \sqrt{\bar{\epsilon}}, \\
\left|\frac{\partial \bar{J}}{\partial w_k}(a_\omega(\tau), w_\omega(\tau))\right| &\leq 2|a_{\omega,k}(\tau)|\int_0^\infty te^{-w_{\omega,k}(\tau)t}\left|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right| dt \\
&\lesssim |a_{\omega,k}(\tau)|\frac{1}{\sqrt{w_{\omega,k}^3(\tau)}}\left\|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right\|_{L^2[0,\infty)} \\
&\lesssim |a_{\omega,k}(\tau)|\frac{\sqrt{\bar{\epsilon}}}{\sqrt{(w_l^0)^3}} \lesssim |a_{\omega,k}(\tau)|\sqrt{\bar{\epsilon}}.
\end{aligned}$$

By (182), we get

$$\begin{aligned}
\|\nabla \bar{J}(\theta_\omega(\tau))\|_2^2 &= \sum_{k=1}^m \left|\frac{\partial \bar{J}}{\partial a_k}(a_\omega(\tau), w_\omega(\tau))\right|^2 + \sum_{k=1}^m \left|\frac{\partial \bar{J}}{\partial w_k}(a_\omega(\tau), w_\omega(\tau))\right|^2 \\
&\lesssim m\bar{\epsilon} + \bar{\epsilon}\|a_\omega(\tau)\|_2^2 \lesssim m\bar{\epsilon}.
\end{aligned} \tag{184}$$

By (183) and (184), we obtain that for any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l^0\})$, $m \in \mathbb{N}_+$ such that $\bar{\tau} = \bar{\tau}(\omega, m) \leq \tau_1$, and any $\theta_0 \in \Theta_0$, $\tau \in [\bar{\tau}/2, \bar{\tau}]$,

$$\left|\left[\nabla \bar{J}(\theta_\omega(\tau))\right]^\top \mathcal{E}_\omega(\theta_\omega(\tau))\right| \leq \left\|\nabla \bar{J}(\theta_\omega(\tau))\right\|_2 \|\mathcal{E}_\omega(\theta_\omega(\tau))\|_2 \lesssim \sqrt{\bar{\epsilon}} m\omega^{-1} e^{-\frac{w_l^0}{2\omega}}. \tag{185}$$

We discuss for different cases.

(i) $\bar{\tau} = 0$. Then $\tau' = 0$. By (178), (183) and (184), we get

$$\|\nabla J_\omega(\theta_\omega(\tau'))\|_2^2 \lesssim \|\nabla \bar{J}(\theta_\omega(\tau'))\|_2^2 + \|\mathcal{E}_\omega(\theta_\omega(\tau'))\|_2^2 \lesssim m\bar{\epsilon} + m\omega^{-2}e^{-\frac{w_l^0}{\omega}}.$$

(ii) $\bar{\tau} \neq 0$. If $\|\nabla \bar{J}(\theta_\omega(\tau))\|_2^2 \gtrsim 4\bar{\epsilon}/\bar{\tau} + \sqrt{\bar{\epsilon}} m\omega^{-1}e^{-\frac{w_l^0}{2\omega}}$ for any $\tau \in [\bar{\tau}/2, \bar{\tau}]$, then by (173), (180) and (185), we have

$$\begin{aligned}
\bar{J}(\theta_\omega(\bar{\tau})) &= \bar{J}(\theta_\omega(\bar{\tau}/2)) + \int_{\bar{\tau}/2}^{\bar{\tau}} \frac{d}{d\tau}\bar{J}(\theta_\omega(\tau))d\tau \\
&\lesssim \bar{\epsilon} - \int_{\bar{\tau}/2}^{\bar{\tau}} \|\nabla \bar{J}(\theta_\omega(\tau))\|_2^2 d\tau + \int_{\bar{\tau}/2}^{\bar{\tau}} \left|\left[\nabla \bar{J}(\theta_\omega(\tau))\right]^\top \mathcal{E}_\omega(\theta_\omega(\tau))\right| d\tau \\
&\lesssim \bar{\epsilon} - \left(2\bar{\epsilon} + \sqrt{\bar{\epsilon}} m\omega^{-1}e^{-\frac{w_l^0}{2\omega}}\frac{\bar{\tau}}{2}\right) + \sqrt{\bar{\epsilon}} m\omega^{-1}e^{-\frac{w_l^0}{2\omega}}\frac{\bar{\tau}}{2} = -\bar{\epsilon} < 0,
\end{aligned}$$

which is a contradictory. Hence, there exists $\tau' = \tau'(\omega, m, \theta_0) \in [\bar{\tau}/2, \bar{\tau}]$, such that $\|\nabla \bar{J}(\theta_\omega(\tau'))\|_2^2 \lesssim 4\bar{\epsilon}/\bar{\tau} + \sqrt{\bar{\epsilon}} m\omega^{-1} e^{-\frac{w_l^0}{2\omega}}$. Therefore, by (178) and (183), we get

$$\begin{aligned}\|\nabla J_\omega(\theta_\omega(\tau'))\|_2^2 &\lesssim \|\nabla \bar{J}(\theta_\omega(\tau'))\|_2^2 + \|\mathcal{E}_\omega(\theta_\omega(\tau'))\|_2^2 \\ &\lesssim \frac{4}{\bar{\tau}}\bar{\epsilon} + \sqrt{\bar{\epsilon}} m\omega^{-1} e^{-\frac{w_l^0}{2\omega}} + m\omega^{-2} e^{-\frac{w_l^0}{\omega}} \\ &\lesssim \frac{\bar{\epsilon}}{\bar{\tau}} + m\omega^{-2} e^{-\frac{w_l^0}{2\omega}}.\end{aligned}$$

Combining (i) (ii) completes the proof. ■

(4) Eigenvalues of Hessian

Proposition 46 *Under the conditions of Proposition 45, denote by $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{2m}$ the eigenvalues of $\nabla^2 J_\omega(\theta_\omega(\tau'))$. Then we have*

$$|\lambda_{m+i}| \lesssim mC^{-i} + \sqrt{\bar{\epsilon}} + \omega^{-2} e^{-\frac{w_l^0}{2\omega}}, \quad i = 1, 2, \cdots, m. \tag{186}$$

Here $C > 1$ and $\lesssim$ hide universal positive constants only depending on a_l^0, a_r^0, w_l^0, w_r^0, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1.

Proof A straightforward computation shows, for $k, j = 1, 2, \cdots, m$,

$$\begin{aligned}
&\frac{\partial^2 J_\omega}{\partial a_k \partial a_j}(a, w) = \frac{2}{w_k + w_j}, \\
&\frac{\partial^2 J_\omega}{\partial a_k \partial w_j}(a, w) = \frac{-2a_j}{(w_k + w_j)^2}, \quad k \neq j, \\
&\frac{\partial^2 J_\omega}{\partial a_k \partial w_k}(a, w) = -\frac{a_k}{2w_k^2} - 2\int_0^\infty t e^{-w_k t}\left(\sum_{i=1}^m a_i e^{-w_i t} - \bar{\rho}(t)\right) dt + 2\Delta_{1,\omega}(w_k), \\
&\frac{\partial^2 J_\omega}{\partial w_k \partial w_j}(a, w) = \frac{4a_k a_j}{(w_k + w_j)^3}, \quad k \neq j, \\
&\frac{\partial^2 J_\omega}{\partial w_k \partial w_k}(a, w) = \frac{a_k^2}{2w_k^3} + 2a_k \int_0^\infty t^2 e^{-w_k t}\left(\sum_{i=1}^m a_i e^{-w_i t} - \bar{\rho}(t)\right) dt - 2a_k \Delta_{2,\omega}(w_k).
\end{aligned}$$

Consider the decomposition $\nabla^2 J_\omega(a, w) = H_1(a, w) + 2H_2(a, w) + 2H_3(a, w)$, where

$$\begin{aligned}
H_1(a, w) &:= \begin{bmatrix} \left[\frac{2}{w_k + w_j}\right] & \left[\frac{-2a_j}{(w_k + w_j)^2}\right] \\ \left[\frac{-2a_k}{(w_k + w_j)^2}\right] & \left[\frac{4a_k a_j}{(w_k + w_j)^3}\right] \end{bmatrix}, \\
H_2(a, w) &:= \begin{bmatrix} \boldsymbol{O}_{m\times m} & \mathrm{Diag}\left(-\int_0^\infty \left(a^\top e^{-wt} - \bar{\rho}(t)\right) t e^{-wt} dt\right) \\ \mathrm{Diag}\left(-\int_0^\infty \left(a^\top e^{-wt} - \bar{\rho}(t)\right) t e^{-wt} dt\right) & \mathrm{Diag}\left(a \circ \int_0^\infty \left(a^\top e^{-wt} - \bar{\rho}(t)\right) t^2 e^{-wt} dt\right) \end{bmatrix}, \\
H_3(a, w) &:= \begin{bmatrix} \boldsymbol{O}_{m\times m} & \mathrm{Diag}(\Delta_{1,\omega}(w)) \\ \mathrm{Diag}(\Delta_{1,\omega}(w)) & \mathrm{Diag}(-a \circ \Delta_{2,\omega}(w)) \end{bmatrix}
\end{aligned}$$

are blocked matrices, where $[M_{kj}]$ denotes the matrix M with the (k,j)-element M_{kj}, and the integral is performed element-wisely. We analyze the eigenvalues of H_1, H_2 and H_3 respectively.

(i) Eigenvalues of H_1.

a) We first show that $H_1(a,w)$ is positive semi-definite (PSD) for any $(a,w) \in \mathbb{R}^m \times \mathbb{R}_+^m$. It is straightforward to verify that $H_1(a,w) = H_{1,1}(w) \circ 2H_{1,2}(a)$, where

$$H_{1,1}(w) := \begin{bmatrix} \left[\frac{1}{w_k+w_j}\right] & \left[\frac{1}{(w_k+w_j)^2}\right] \\ \left[\frac{1}{(w_k+w_j)^2}\right] & \left[\frac{2}{(w_k+w_j)^3}\right] \end{bmatrix}, \quad H_{1,2}(a) := \begin{bmatrix} \mathbf{1}_{m\times m} & -\mathbf{1}_m a^\top \\ -a\mathbf{1}_m^\top & aa^\top \end{bmatrix}.$$

Since

$$H_{1,1}(w) = \int_0^\infty \begin{bmatrix} \left[e^{-(w_k+w_j)t}\right] & \left[te^{-(w_k+w_j)t}\right] \\ \left[te^{-(w_k+w_j)t}\right] & \left[t^2e^{-(w_k+w_j)t}\right] \end{bmatrix} dt = \int_0^\infty \begin{bmatrix} e^{-wt} \\ te^{-wt} \end{bmatrix} \begin{bmatrix} e^{-w^\top t} & te^{-w^\top t} \end{bmatrix} dt,$$

and for any $w \in \mathbb{R}_+^m$ and $t \geq 0$, the matrix $\begin{bmatrix} e^{-wt} \\ te^{-wt} \end{bmatrix} \begin{bmatrix} e^{-w^\top t} & te^{-w^\top t} \end{bmatrix}$ is PSD, we obtain that $H_{1,1}(w)$ is PSD for any $w \in \mathbb{R}_+^m$.[17] The fact that $H_{1,2}(a)$ is also PSD for any $a \in \mathbb{R}^m$ is trivial, since $H_{1,2}(a) = \begin{bmatrix} \mathbf{1}_m \\ -a \end{bmatrix} \begin{bmatrix} \mathbf{1}_m^\top & -a^\top \end{bmatrix}$. Therefore, by Schur's product theorem, as an Hadamard product of two PSD matrices, $H_1(a,w)$ is PSD for any $(a,w) \in \mathbb{R}^m \times \mathbb{R}_+^m$.

b) Then we show that $H_1(a,w)$ has multiple near-zero eigenvalues for any $(a,w) \in \mathbb{R}^m \times \mathbb{R}_+^m$, provided $m \in \mathbb{N}_+$ appropriately large. Let $\mu_1 \geq \mu_2 \geq \cdots \geq \mu_{2m} \geq 0$ and $\nu_1 \geq \nu_2 \geq \cdots \geq \nu_m$ be the eigenvalues of $H_1(a,w)$ and $\left[\frac{2}{w_k+w_j}\right]$, respectively. By Cauchy's interlacing theorem, $\mu_{m+i} \leq \nu_i \leq \mu_i$ for $i = 1, 2, \cdots, m$. Then $\nu_m \geq \mu_{2m} \geq 0$, which gives $\left[\frac{2}{w_k+w_j}\right]$ is also PSD, hence all the eigenvalues $\{\nu_k\}_{k=1}^m$ are singular values. Since $\left[\frac{2}{w_k+w_j}\right]$ is in fact a Cauchy matrix, according to Corollary 7 in Beckermann and Townsend (2017), its singular values decay exponentially fast, i.e.

$$\nu_{j+k} \leq 4\left[\exp\left(\frac{\pi^2}{2\log(16\gamma)}\right)\right]^{-2k} \nu_j, \quad 1 \leq j+k \leq m,$$

where γ denotes the absolute value of the cross-ratio of $w_{\min}$, $w_{\max}$, $-w_{\max}$ and $-w_{\min}$ with $w_{\min} := \min_{i\in[m]} w_i$, $w_{\max} := \max_{i\in[m]} w_i$.[18] Since $\nu_1 = \left\|\left[\frac{2}{w_k+w_j}\right]\right\|_2 \leq \left\|\left[\frac{2}{w_k+w_j}\right]\right\|_F \leq \left\|\left[\frac{1}{w_{\min}}\right]\right\|_F = \frac{m}{w_{\min}}$, we get $\nu_{1+k} \lesssim \frac{m}{w_{\min}} C_w^{-k}$ for $k = 0, 1, \cdots, m-1$, where $C_w > 1$ is a universal constant only depending on $w_{\min}$, $w_{\max}$. This implies $0 \leq \mu_{m+i} \leq \nu_i \lesssim \frac{m}{w_{\min}} C_w^{-i}$.[19] That is, $H_1(a,w)$ has $\mathcal{O}(m)$ exponentially small eigenvalues for any $(a,w) \in \mathbb{R}^m \times \mathbb{R}_+^m$.

17. If $A(t)$ is PSD for any $t \in I \subset \mathbb{R}$, then the matrix $A := \int_I A(t)dt$ is also PSD. In fact, for any x, $x^\top Ax = \int_I x^\top A(t)x dt \geq \int_I \lambda_{\min}(A(t))\|x\|_2^2 dt \geq 0$.

18. Given four numbers a, b, c and d, the cross-ratio is given by $\gamma := \frac{(c-a)(d-b)}{(c-b)(d-a)}$. One can easily check $\gamma \geq 1$ here and the equality holds if and only if $w_{\min} = w_{\max}$, i.e. $w = c\mathbf{1}_m$ for some $c > 0$. This degenerate case is trivial since now $\operatorname{rank}\left(\left[\frac{2}{w_k+w_j}\right]\right) = 1$, which gives $\nu_2 = \cdots = \nu_m = 0$.

19. Since for any $w \in \mathbb{R}_+^m$, $\gamma = \gamma(w_{\min}, w_{\max}, -w_{\max}, -w_{\min}) \geq 1$, we have $C_w := \exp\left(\frac{\pi^2}{\log(16\gamma)}\right) \in \left(1, \exp\left(\frac{\pi^2}{\log 16}\right)\right]$.

c) Back to the dynamics of parameters. According to Lemma 40 and Lemma 42, for any $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \Theta_0$ and $\tau \in [0, \tau_1]$, we have $w_{\omega,k}(\tau) \geq w_l^0/2 > 0$, $k = 1, 2, \cdots, m$, and $|w_{\omega,k}(\tau) - w_{\omega,k}(0)| \lesssim (\tau^2 + |a_{\omega,k}(0)|\tau)$ by (169), where $\lesssim$ hides universal positive constants only related to J^0, w_l^0, which gives

$$|w_{\omega,k}(\tau)| \lesssim |w_{0,k}| + (\tau^2 + |a_{0,k}|\tau) \leq w_r^0 + \tau_1^2 + c_a^0\tau_1 \lesssim 1, \tag{187}$$

where $\lesssim$ hides universal positive constants only depending on w_r^0, c_a^0, τ_1. Take $\tau = \tau' = \tau'(\omega, m, \theta_0)$ with τ' derived from Proposition 45, and $w = w_\omega(\tau')$, we get $\tau' \in [\bar{\tau}/2, \bar{\tau}] \subset [0, \tau_1]$, hence $w_{\min} \geq w_l^0/2$ and $w_{\max} \lesssim 1$. A more refined estimate on above constants gives $1 \leq \gamma = \frac{1}{4}\left(\frac{w_{\max}}{w_{\min}} + \frac{w_{\min}}{w_{\max}}\right) + \frac{1}{2} \lesssim \frac{1}{4}\left(\frac{2}{w_l^0} + \frac{w_l^0}{2}\right) + \frac{1}{2} \lesssim 1$, where $w_l^0 > 0$ is also hided. Rewrite the upper bound as $C_0 \geq 1$, we get that C_0 only depends on $w_l^0, w_r^0, c_a^0, J^0, \tau_1$. This yields $C_w \geq \exp\left(\frac{\pi^2}{\log(16C_0)}\right) \triangleq C > 1$. Therefore, we obtain $0 \leq \mu_{m+i} \leq \nu_i \lesssim \frac{m}{w_{\min}} C_w^{-i} \lesssim \frac{m}{w_l^0} C_w^{-i} \lesssim mC^{-i}$ with $w_l^0 > 0$ hided.

(ii) Eigenvalues of H_2. By the Cauchy-Schwartz inequality, for any $w_k > 0$ we have

$$\left|\int_0^\infty te^{-w_k t}\left(a^\top e^{-wt} - \bar{\rho}(t)\right)dt\right| \lesssim \frac{1}{\sqrt{w_k^3}}\left\|a^\top e^{-wt} - \bar{\rho}(t)\right\|_{L^2[0,\infty)},$$

$$\left|a_k\int_0^\infty t^2e^{-w_k t}\left(a^\top e^{-wt} - \bar{\rho}(t)\right)dt\right| \lesssim \frac{|a_k|}{\sqrt{w_k^5}}\left\|a^\top e^{-wt} - \bar{\rho}(t)\right\|_{L^2[0,\infty)}.$$

Take $(a, w) = (a_\omega(\tau'), w_\omega(\tau'))$ with τ' derived from Proposition 45. According to Lemma 40 and (173), (168), we have

$$\left|\int_0^\infty te^{-w_{\omega,k}(\tau')t}\left([a_\omega(\tau')]^\top e^{-w_\omega(\tau')t} - \bar{\rho}(t)\right)dt\right| \lesssim \frac{1}{\sqrt{(w_l^0)^3}}\sqrt{\bar{\epsilon}} \lesssim \sqrt{\bar{\epsilon}}, \tag{188}$$

$$\left|a_{\omega,k}(\tau')\int_0^\infty t^2e^{-w_{\omega,k}(\tau')t}\left([a_\omega(\tau')]^\top e^{-w_\omega(\tau')t} - \bar{\rho}(t)\right)dt\right| \lesssim \frac{c_a^0 + \tau_1}{\sqrt{(w_l^0)^5}}\sqrt{\bar{\epsilon}} \lesssim \sqrt{\bar{\epsilon}}, \tag{189}$$

where $\lesssim$ hides universal positive constants only depending on $w_l^0, c_a^0, J^0, \tau_1$. By Gerŝgorin's circle theorem, for any ξ as the eigenvalue of $H_2(a_\omega(\tau'), w_\omega(\tau'))$, we have

$$|\xi| \leq \left|\int_0^\infty te^{-w_{\omega,k}(\tau')t}\left([a_\omega(\tau')]^\top e^{-w_\omega(\tau')t} - \bar{\rho}(t)\right)dt\right|,$$

$$\text{or} \quad \left|\xi - a_{\omega,k}(\tau')\int_0^\infty t^2e^{-w_{\omega,k}(\tau')t}\left([a_\omega(\tau')]^\top e^{-w_\omega(\tau')t} - \bar{\rho}(t)\right)dt\right|$$
$$\leq \left|\int_0^\infty te^{-w_{\omega,k}(\tau')t}\left([a_\omega(\tau')]^\top e^{-w_\omega(\tau')t} - \bar{\rho}(t)\right)dt\right|,$$

and each of them gives $|\xi| \lesssim \sqrt{\bar{\epsilon}}$ by (188) and (189).

(iii) Eigenvalues of H_3. According to Lemma 27 and Lemma 40, we have the same bound as (181). That is,

$$|\Delta_{n,\omega}(w_{\omega,k}(\tau'))| \leq \Delta_{n,\omega}^+(w_{\omega,k}(\tau')) \leq \Delta_{n,\omega}^+(w_l^0/2) \lesssim \omega^{-n}e^{-\frac{w_l^0}{2\omega}}\left(c_2^{(w_l^0)^2/4} + c_3^{w_l^0/2}\right) \tag{190}$$

holds for any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l^0\})$, where $c_2, c_3 > 1$ are constants only depending on c_1, t_0, and $\lesssim$ hides universal positive constants only depending on n and ρ_0, t_0, c_0, c_1. By (168), we have $|a_{\omega,k}(\tau')| \lesssim c_a^0 + \tau_1 \lesssim 1$ with $\lesssim$ hiding universal positive constants only related to $w_l^0, c_a^0, J^0, \tau_1$. Again, by Gerŝgorin's circle theorem, for any η as the eigenvalue of $H_3(a_\omega(\tau'), w_\omega(\tau'))$, we have

$$\begin{aligned} & |\eta| \le |\Delta_{1,\omega}(w_{\omega,k}(\tau'))|, \\ \text{or} \quad & |\eta + a_{\omega,k}(\tau')\Delta_{2,\omega}(w_{\omega,k}(\tau'))| \le |\Delta_{1,\omega}(w_{\omega,k}(\tau'))|, \end{aligned}$$

which gives $|\eta| \lesssim \omega^{-2} e^{-\frac{w_l^0}{2\omega}}$, where $\lesssim$ also hides universal positive constants only related to w_l^0, c_2, c_3.

Combining the estimates of (i), (ii) and (iii) and applying Weyl's theorem yields, for $i = 1, 2, \cdots, m$,

$$\begin{aligned} \lambda_{m+i} &\le \mu_{m+i} + \xi_1 + \eta_1 \lesssim mC^{-i} + \sqrt{\epsilon} + \omega^{-2} e^{-\frac{w_l^0}{2\omega}}, \\ \lambda_{m+i} &\ge \mu_{m+i} + \xi_{2m} + \eta_{2m} \gtrsim -\sqrt{\epsilon} - \omega^{-2} e^{-\frac{w_l^0}{2\omega}}, \end{aligned}$$

where $C > 1$ is a universal constant only related to $w_l^0, w_r^0, c_a^0, J^0, \tau_1$. The proof is completed. ■

Proposition 46 naturally implies that learning for the teacher-student model of exponential sums is ill-conditioned, in the sense that all the global minimizers are singular. Specifically, we can show that the Hessian around these global minimizers has multiple exponentially small eigenvalues, provided the network width (model capacity) $m \in \mathbb{N}_+$ appropriately large.

Corollary 47 *For any $(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m$ satisfying $\bar{J}(a, w) = \|a^\top e^{-wt} - \bar{\rho}(t)\|_{L^2[0,\infty)}^2 \le \bar{\varepsilon} = \bar{\varepsilon}(m)$, then we have*

$$\|\nabla \bar{J}(a, w)\|_2^2 \lesssim m\bar{\varepsilon}, \tag{191}$$

$$|\bar{\lambda}_{m+i}| \lesssim mC_w^{-i} + \sqrt{\bar{\varepsilon}}, \quad i = 1, 2, \cdots, m, \tag{192}$$

where $\bar{\lambda}_1 \ge \bar{\lambda}_2 \ge \cdots \ge \bar{\lambda}_{2m}$ are the eigenvalues of $\nabla^2 \bar{J}(a, w)$, and $\lesssim$ hides universal positive constants only related to (a, w), and $C_w > 1$ is a universal constant only depending on w. Therefore, $\bar{\varepsilon} = 0$ implies a zero gradient and $|\bar{\lambda}_{m+i}| \lesssim mC_w^{-i}$ for any $i \in [m]$.[20]

Proof For any any $(a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m$, by the Cauchy-Schwartz inequality, we have

$$\begin{aligned} \left| \frac{\partial \bar{J}}{\partial a_k}(a, w) \right| &\le 2 \int_0^\infty e^{-w_k t} \left| a^\top e^{-wt} - \bar{\rho}(t) \right| dt \le \frac{\sqrt{2}}{\sqrt{w_k}} \|a^\top e^{-wt} - \bar{\rho}(t)\|_{L^2[0,\infty)}, \\ \left| \frac{\partial \bar{J}}{\partial w_k}(a, w) \right| &\le 2|a_k| \int_0^\infty t e^{-w_k t} \left| a^\top e^{-wt} - \bar{\rho}(t) \right| dt \le \frac{|a_k|}{\sqrt{w_k^3}} \|a^\top e^{-wt} - \bar{\rho}(t)\|_{L^2[0,\infty)}, \end{aligned}$$

20. Obviously, a sufficient condition to lead $\bar{\varepsilon} = 0$ is to take a discrete distribution π_0, e.g. $d\pi_0(w)/dw = \sum_{j=1}^{m^*} \delta(w - w_j^*)$ with $m^* \le m$ and $w_j^* > 0$ for any $j \in [m^*]$, where $\delta(\cdot)$ denotes the common Dirac function. Then there exists $(b^*, v^*) \in \mathbb{R}^m \times \mathbb{R}_+^m$ such that $b^{*\top} e^{-v^* t} \equiv \bar{\rho}(t)$.

which gives (191). Notice that $\nabla^2\bar{J}(a,w) = H_1(a,w) + H_2(a,w)$, according to the arguments (i), (ii) in the proof of Proposition 46, we get (192). The proof is completed. ■

(5) Local Linearization Analysis

Now we can derive a quantitative timescale of the trap in plateauing via linearization arguments.

Proof [Proof of Theorem 23] (i) The estimate (112) follows from Proposition 44.

(ii) The proof for hitting time (115) follows the same procedure as that of Theorem 20 (i.e. Section 7.1 **(5)**), where the asymptotic analysis and linearization techniques are applied. Based on estimates of local curvature of the loss landscape, we can obtain the same lower bound as (150) of the timescale of slow downs

$$\tau_0 - \tau' \gtrsim \min\left\{\frac{\delta}{\|g_0\|_2}, \frac{\ln(1+\delta)}{|\lambda_{2m}|}\right\}, \tag{193}$$

where $g_0 := \nabla J_\omega(\theta_\omega(\tau'))$, and λ_{2m} is the minimal eigenvalue of $H_0 := \nabla^2 J_\omega(\theta_\omega(\tau'))$. According to Proposition 45 and Proposition 46, we obtain

$$\tau_0 - \tau' \gtrsim \begin{cases} \min\left\{\frac{\delta}{\sqrt{m\bar{\epsilon}}+\sqrt{m}\omega^{-1}e^{-\frac{w_l^0}{2\omega}}}, \frac{\ln(1+\delta)}{mC^{-m}+\sqrt{\bar{\epsilon}}+\omega^{-2}e^{-\frac{w_l^0}{2\omega}}}\right\}, & \bar{\tau}=0, \\ \min\left\{\frac{\delta}{\sqrt{\bar{\epsilon}/\bar{\tau}}+\sqrt{m}\omega^{-1}e^{-\frac{w_l^0}{4\omega}}}, \frac{\ln(1+\delta)}{mC^{-m}+\sqrt{\bar{\epsilon}}+\omega^{-2}e^{-\frac{w_l^0}{2\omega}}}\right\}, & \bar{\tau}\neq 0, \end{cases} \tag{194}$$

where $C > 1$ and $\gtrsim$ hide universal positive constants only depending on a_l^0, a_r^0, w_l^0, w_r^0, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1. The last task is to show the loss dynamics is much slower than the parameter dynamics when there are slow downs. It is straightforward since for any $\tau \in [\tau', \tau_0]$,

$$\begin{aligned} J_\omega(\theta_\omega(\tau)) - J_\omega(\theta_\omega(\tau')) &= g_0^\top(\theta_\omega(\tau) - \theta_\omega(\tau')) + (\theta_\omega(\tau) - \theta_\omega(\tau'))^\top H_0(\theta_\omega(\tau) - \theta_\omega(\tau')) + o(\delta^2) \\ &\geq -\|g_0\|_2\|\theta_\omega(\tau) - \theta_\omega(\tau')\|_2 + \lambda_{2m}\|\theta_\omega(\tau) - \theta_\omega(\tau')\|_2^2 + o(\delta^2) \\ &= o(1)\mathcal{O}(\delta) + o(1)\mathcal{O}(\delta^2) + o(\delta^2) \\ &= o(\delta^2), \quad \omega \to 0^+, \end{aligned}$$

provided $m \in \mathbb{N}_+$ appropriately large and $\max\{\bar{\epsilon}, \bar{\epsilon}/\bar{\tau}\} \ll 1$. The proof is completed. ■

7.3 Sufficiently Wide RNNs: an Example

In this section, we prove Theorem 25 for sufficiently wide RNNs, which gives an example to illustrate the conditions in Theorem 23. Connecting Theorem 25 with Theorem 23, we obtain that the gradient flow training dynamics (104) for learning with linear RNNs (101) appears a typical *two-stage process with separation of timescales*. That is, the model $\hat{\rho}$ first learns the short-term memory part of target (i.e. $\bar{\rho}$) rapidly, then the training becomes stuck for a long time.

Motivated by E et al. (2020), the basic insight here is to compare the non-convex training dynamics (104) with the random feature model (i.e. sampling the feature parameters w at initialization and freezing them during training), which is convex and hence easy to analyze.

Random Feature Model. For reference, we consider a random feature version of (172):

$$\min_{a\in\mathbb{R}^m} \bar{J}(a, w_0) = \|\hat{\rho}(t; a, w_0) - \bar{\rho}(t)\|^2_{L^2[0,\infty)}, \tag{195}$$

with the corresponding gradient flow training dynamics

$$\frac{d}{d\tau}\tilde{a}(\tau) = -\nabla_a \bar{J}(\tilde{a}(\tau), w_0), \quad \tilde{a}(0) = a_0. \tag{196}$$

Here $(a_0, w_0) = \theta_0 \in \tilde{\Theta}_0$, i.e. under the random bounded initialization (recall (116)).

Theorem 25 is proved subsequently in the following procedure:

1. For the random feature model (195), we prove that the optimal solution exists, and it can be achieved by the gradient flow training dynamics (196);
2. We prove that the solution of the "full" gradient flow (104) and the "partial" one (196) starting from the same initialization can be close, provided we have sufficiently wide RNNs ($m \gg 1$) and long-term memories in targets ($0 < \omega \ll 1$);
3. Combining the former two steps gives the desired conclusion.

(1) Model for Reference

The analysis of the random feature model (195) is given in this part. We first show the existence of optimal solutions.

Lemma 48 *Fix any $m \in \mathbb{N}_+$. For any $\delta > 0$, with probability of at least $1-\delta$ over the choice of w_0, there exists $\hat{a} = \hat{a}(w_0)$ such that*

$$\bar{J}(\hat{a}, w_0) \lesssim \frac{1}{m}(1 + \ln(2/\delta)), \tag{197}$$

$$\|\hat{a}\|_2 \lesssim \frac{1}{\sqrt{m}}, \tag{198}$$

where $\lesssim$ hides universal positive constants only depending on w_l, w_r and a^*.

Proof For any w_0, let $\hat{a} = \hat{a}(w_0) = a^*(w_0)/m$ with $a^*(\cdot)$ performed element-wisely, then

$$\hat{\rho}(t; \hat{a}, w_0) = \frac{1}{m}\sum_{k=1}^{m} a^*(w_{0,k}) e^{-w_{0,k}t}, \tag{199}$$

hence $\mathbb{E}_{w_0}[\hat{\rho}(t; \hat{a}, w_0)] = \mathbb{E}_{w\sim\pi_0}[a^*(w)e^{-wt}] = \bar{\rho}(t)$. Let

$$Z(w_0) := \sqrt{\bar{J}(\hat{a}(w_0), w_0)} = \|\hat{\rho}(t; \hat{a}(w_0), w_0) - \bar{\rho}(t)\|_{L^2[0,\infty)}. \tag{200}$$

If $\tilde{w}_0$ is different from w_0 at only one component indexed by i, the triangle inequality gives

$$\begin{aligned}
|Z(w_0) - Z(\tilde{w}_0)| &= \left| \|\hat{\rho}(t; \hat{a}(w_0), w_0) - \bar{\rho}(t)\|_{L^2[0,\infty)} - \|\hat{\rho}(t; \hat{a}(\tilde{w}_0), \tilde{w}_0) - \bar{\rho}(t)\|_{L^2[0,\infty)} \right| \\
&\le \|\hat{\rho}(t; \hat{a}(w_0), w_0) - \hat{\rho}(t; \hat{a}(\tilde{w}_0), \tilde{w}_0)\|_{L^2[0,\infty)} \\
&= \frac{1}{m} \left\| a^*(w_{0,i}) e^{-w_{0,i}t} - a^*(\tilde{w}_{0,i}) e^{-\tilde{w}_{0,i}t} \right\|_{L^2[0,\infty)} \\
&\le \frac{1}{m} \left(|a^*(w_{0,i})| \sqrt{\frac{1}{2w_{0,i}}} + |a^*(\tilde{w}_{0,i})| \sqrt{\frac{1}{2\tilde{w}_{0,i}}} \right) \le \frac{c^*}{m} \sqrt{\frac{2}{w_l}},
\end{aligned}$$

where $c^* := \|a^*\|_{L^\infty[w_l, w_r]}$. By McDiarmid's inequality, for any $\delta > 0$, with probability of at least $1 - \delta$, we have

$$|Z(w_0) - \mathbb{E}_{w_0}[Z(w_0)]| \le \frac{c^*}{m} \sqrt{\frac{2}{w_l}} \sqrt{m \ln(2/\delta)/2} \lesssim \sqrt{\frac{\ln(2/\delta)}{m}}, \tag{201}$$

where $\lesssim$ hides universal positive constants only related to w_l, w_r and a^*. Since

$$\begin{aligned}
\mathbb{E}_{w_0}\left[Z^2(w_0)\right] &= \mathbb{E}_{w_0}\left[\bar{J}(\hat{a}(w_0), w_0)\right] \\
&= \int_0^\infty \mathbb{E}_{w_0} \left[\left(\frac{1}{m} \sum_{k=1}^m a^*(w_{0,k}) e^{-w_{0,k}t} - \bar{\rho}(t) \right)^2 \right] dt \\
&= \frac{1}{m^2} \int_0^\infty \mathbb{E}_{w_0} \left[\left(\sum_{k=1}^m (a^*(w_{0,k}) e^{-w_{0,k}t} - \bar{\rho}(t)) \right)^2 \right] dt \\
&= \frac{1}{m^2} \sum_{k=1}^m \int_0^\infty \mathbb{E}_{w \sim \pi_0}\left[(a^*(w) e^{-wt} - \bar{\rho}(t))^2 \right] dt \\
&\quad + \frac{1}{m^2} \sum_{i \ne j} \int_0^\infty \mathbb{E}^2_{w \sim \pi_0}[a^*(w) e^{-wt} - \bar{\rho}(t)] dt \\
&= \frac{1}{m} \int_0^\infty \mathbb{E}_{w \sim \pi_0}\left[(a^*(w) e^{-wt} - \bar{\rho}(t))^2 \right] dt \\
&\le \frac{1}{m} \int_0^\infty \mathbb{E}_{w \sim \pi_0}\left[a^*(w)^2 e^{-2wt} \right] dt \le \frac{1}{m} \frac{(c^*)^2}{2w_l},
\end{aligned}$$

by Jensen's inequality,

$$\begin{aligned}
\bar{J}(\hat{a}(w_0), w_0) = Z^2(w_0) &\lesssim \left(|\mathbb{E}_{w_0}[Z(w_0)]| + \sqrt{\frac{\ln(2/\delta)}{m}} \right)^2 \\
&\le 2 \left(\mathbb{E}_{w_0}\left[Z^2(w_0)\right] + \frac{\ln(2/\delta)}{m} \right) \lesssim \frac{1}{m}(1 + \ln(2/\delta)).
\end{aligned}$$

Obviously, $\|\hat{a}\|_2 = \|a^*(w_0)\|_2/m \le c^*/\sqrt{m}$. The proof is completed. ■

Then we show that the gradient flow training dynamics (196) can find the optimal solution for the random feature model (195).

Lemma 49 *Fix any $m \in \mathbb{N}_+$. For any $\delta > 0$, with probability of at least $1-\delta$ over the choice of w_0, we have*

$$\bar{J}(\tilde{a}(\tau), w_0) \lesssim \frac{1}{m\tau} + \frac{1}{m} C_\delta^2, \tag{202}$$

$$\|\tilde{a}(\tau)\|_2 \lesssim \frac{1}{\sqrt{m}}(1 + C_\delta\sqrt{\tau}). \tag{203}$$

Here $\tilde{a}(\tau)$ is the solution to (196), and $\lesssim$ hides universal positive constants only depending on a_l, a_r, w_l, w_r and a^*, and $C_\delta := \sqrt{1+\ln(2/\delta)}$.

Proof According to Lemma 48, for any $\delta > 0$, with probability of at least $1-\delta$ over the choice of w_0, there exists $\hat{a} = \hat{a}(w_0)$ such that $\bar{J}(\hat{a}, w_0) \lesssim (1+\ln(2/\delta))/m$ and $\|\hat{a}\|_2 \lesssim 1/\sqrt{m}$, where $\lesssim$ hides universal positive constants only related to w_l, w_r and a^*. Consider the Lyapunov function

$$E(\tau) := \tau\big(\bar{J}(\tilde{a}(\tau), w_0) - \bar{J}(\hat{a}, w_0)\big) + \frac{1}{2}\|\tilde{a}(\tau) - \hat{a}\|_2^2, \tag{204}$$

It is straightforward to verify that $\bar{J}(\tilde{a}(\tau), w_0)$ is quadratic, and hence convex to $\tilde{a}(\tau)$. Therefore

$$\begin{aligned}
E'(\tau) &= \bar{J}(\tilde{a}(\tau), w_0) - \bar{J}(\hat{a}, w_0) + \tau\left[\nabla_a \bar{J}(\tilde{a}(\tau), w_0)\right]^\top \frac{d}{d\tau}\tilde{a}(\tau) + (\tilde{a}(\tau) - \hat{a})^\top \frac{d}{d\tau}\tilde{a}(\tau) \\
&= \bar{J}(\tilde{a}(\tau), w_0) - \bar{J}(\hat{a}, w_0) - \tau\|\nabla_a \bar{J}(\tilde{a}(\tau), w_0)\|_2^2 - (\tilde{a}(\tau) - \hat{a})^\top \nabla_a \bar{J}(\tilde{a}(\tau), w_0) \\
&\le -\left[\bar{J}(\hat{a}, w_0) - \Big(\bar{J}(\tilde{a}(\tau), w_0) + (\hat{a} - \tilde{a}(\tau))^\top \nabla_a \bar{J}(\tilde{a}(\tau), w_0)\Big)\right] \le 0,
\end{aligned}$$

which gives

$$\begin{aligned}
E(\tau) \le E(0) &\Leftrightarrow \tau(\bar{J}(\tilde{a}(\tau), w_0) - \bar{J}(\hat{a}, w_0)) + \frac{1}{2}\|\tilde{a}(\tau) - \hat{a}\|_2^2 \le \frac{1}{2}\|a_0 - \hat{a}\|_2^2 \qquad (205) \\
&\Rightarrow \|\tilde{a}(\tau) - \hat{a}\|_2^2 \le \|a_0 - \hat{a}\|_2^2 + 2\tau\bar{J}(\hat{a}, w_0) \\
&\Rightarrow \|\tilde{a}(\tau)\|_2^2 \le 2\|\tilde{a}(\tau) - \hat{a}\|_2^2 + 2\|\hat{a}\|_2^2 \le 2\|a_0 - \hat{a}\|_2^2 + 4\tau\bar{J}(\hat{a}, w_0) + 2\|\hat{a}\|_2^2 \\
&\le 4\|a_0\|_2^2 + 6\|\hat{a}\|_2^2 + 4\tau\bar{J}(\hat{a}, w_0) \lesssim \frac{c_a^2}{m} + \frac{1}{m} + \frac{\tau}{m}(1 + \ln(2/\delta)),
\end{aligned}$$

where $c_a := \max\{|a_l|, |a_r|\}$. In addition, (205) also gives

$$\bar{J}(\tilde{a}(\tau), w_0) \le \frac{1}{2\tau}\|a_0 - \hat{a}\|_2^2 + \bar{J}(\hat{a}, w_0) \lesssim \frac{1}{m\tau}(1 + c_a^2) + \frac{1}{m}(1 + \ln(2/\delta)).$$

The proof is completed. ■

(2) Comparison

To ensure the stability of training dynamics (104) under the new (random bounded) initialization, we first restate Lemma 40 by taking $(a_l^0, a_r^0, w_l^0, w_r^0) = (a_l, a_r, w_l, w_r)$.

Lemma 50 (Training Stability, Restatement) *Let* $J^0 := (c_a + c^*)/\sqrt{2w_l} + \|\rho_0\|_{L^2(\mathbb{R})}$ *with* $c_a := \max\{|a_l|, |a_r|\}$, $c^* := \|a^*\|_{L^\infty[w_l,w_r]}$. *Define the Cauchy problem*

$$v'(\tau) = \frac{1}{2}J^0\left(c_a + \sqrt{2}J^0\int_0^\tau v(s)ds\right)v^6(\tau), \quad v(0) = \frac{1.1}{\sqrt{w_l}}. \tag{206}$$

Let $\tau_1 := \inf\left\{\tau \geq 0 : v(\tau) > \sqrt{\frac{2}{w_l}}\right\}$. *Then for any* $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \tilde{\Theta}_0$ *and* $\tau \in [0, \tau_1]$, *we have*

$$w_{\omega,k}(\tau) \geq \frac{w_l}{2} > 0, \quad k = 1, 2, \cdots, m. \tag{207}$$

Then Lemma 42 is restated as follows.

Lemma 51 (Dynamics of Parameters, Restatement) *For any* $\omega > 0$, $m \in \mathbb{N}_+$, $\theta_0 \in \tilde{\Theta}_0$ *and* $\tau \in [0, \tau_1]$, *we have*

$$\|a_\omega(\tau) - a_\omega(0)\|_2 \lesssim \tau\sqrt{m}, \tag{208}$$

$$\|w_\omega(\tau) - w_\omega(0)\|_2 \lesssim \tau\left(\frac{1}{\sqrt{m}} + \tau\sqrt{m}\right), \tag{209}$$

where $\lesssim$ *hides universal positive constants only depending on* a_l, a_r, w_l, w_r *and* a^*, ρ_0.

Now we bound the difference between solutions to the "full" gradient flow (104) and the "partial" one (196).

Lemma 52 *Fix any* $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l\})$ *and* $m \in \mathbb{N}_+$. *For any* $\delta > 0$, *with probability of at least* $1 - \delta$ *over the choice of* $\theta_0 \in \tilde{\Theta}_0$, *we have*

$$\|a_\omega(\tau) - \tilde{a}(\tau)\|_2 \lesssim \tau^2\left(\frac{1}{\sqrt{m}} + \tau\sqrt{m}\right)(1 + C_\delta\sqrt{\tau}) + \tau\sqrt{m}e^{-\frac{w_l}{4\omega}}, \quad \forall\tau \in [0, \tau_1]. \tag{210}$$

Here $a_\omega(\tau)$, $\tilde{a}(\tau)$ *are solutions to (104), (196) respectively with* $\theta_0 = (a_0, w_0) = (a_\omega(0), w_\omega(0))$ $= (\tilde{a}(0), w_0)$, *and* $\lesssim$ *hides universal positive constants only depending on* a_l, a_r, w_l, w_r *and* a^*, ρ_0, t_0, c_0, c_1, *and* $C_\delta = \sqrt{1 + \ln(2/\delta)}$.

Proof For any $\omega > 0$, $m \in \mathbb{N}_+$, we have

$$\begin{aligned}
&\frac{d}{d\tau}(a_\omega(\tau) - \tilde{a}(\tau)) \\
=& -2\int_0^\infty \left[\left([a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right)e^{-w_\omega(\tau)t} - \left([\tilde{a}(\tau)]^\top e^{-w_0 t} - \bar{\rho}(t)\right)e^{-w_0 t}\right]dt + 2\Delta_{0,\omega}(w_\omega(\tau)) \\
=& -2\int_0^\infty \left[(a_\omega(\tau) - \tilde{a}(\tau))^\top e^{-w_\omega(\tau)t}\right]e^{-w_\omega(\tau)t}dt + 2\int_0^\infty \bar{\rho}(t)\left(e^{-w_\omega(\tau)t} - e^{-w_0 t}\right)dt \\
&-2\int_0^\infty \left[\left([\tilde{a}(\tau)]^\top e^{-w_\omega(\tau)t}\right)e^{-w_\omega(\tau)t} - \left([\tilde{a}(\tau)]^\top e^{-w_0 t}\right)e^{-w_0 t}\right]dt + 2\Delta_{0,\omega}(w_\omega(\tau)) \\
=& -2\int_0^\infty e^{-w_\omega(\tau)t}e^{-[w_\omega(\tau)]^\top t}(a_\omega(\tau) - \tilde{a}(\tau))dt + 2\int_0^\infty \bar{\rho}(t)\left(e^{-w_\omega(\tau)t} - e^{-w_0 t}\right)dt \\
&-2\int_0^\infty \left(e^{-w_\omega(\tau)t}e^{-[w_\omega(\tau)]^\top t} - e^{-w_0 t}e^{-w_0^\top t}\right)\tilde{a}(\tau)dt + 2\Delta_{0,\omega}(w_\omega(\tau)),
\end{aligned}$$

which gives

$$
\begin{aligned}
&\frac{1}{4}\frac{d}{d\tau}\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2^2=\frac{1}{2}(a_\omega(\tau)-\tilde{a}(\tau))^\top\frac{d}{d\tau}(a_\omega(\tau)-\tilde{a}(\tau))\\
=&-\int_0^\infty\left[(a_\omega(\tau)-\tilde{a}(\tau))^\top e^{-w_\omega(\tau)t}\right]^2dt+\int_0^\infty\bar{\rho}(t)(a_\omega(\tau)-\tilde{a}(\tau))^\top\left(e^{-w_\omega(\tau)t}-e^{-w_0t}\right)dt\\
&-\int_0^\infty(a_\omega(\tau)-\tilde{a}(\tau))^\top\left(e^{-w_\omega(\tau)t}e^{-[w_\omega(\tau)]^\top t}-e^{-w_0t}e^{-w_0^\top t}\right)\tilde{a}(\tau)dt\\
&+(a_\omega(\tau)-\tilde{a}(\tau))^\top\Delta_{0,\omega}(w_\omega(\tau))\\
\le&\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\int_0^\infty|\bar{\rho}(t)|\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2dt+\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\Delta_{0,\omega}(w_\omega(\tau))\right\|_2\\
&+\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\tilde{a}(\tau)\right\|_2\int_0^\infty\left\|e^{-w_\omega(\tau)t}e^{-[w_\omega(\tau)]^\top t}-e^{-w_0t}e^{-w_0^\top t}\right\|_2dt\\
\le&\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\{\int_0^\infty|\bar{\rho}(t)|^2dt\right\}^{\frac{1}{2}}\left\{\int_0^\infty\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2^2dt\right\}^{\frac{1}{2}}\\
&+\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\tilde{a}(\tau)\right\|_2\int_0^\infty\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2\left(\|e^{-w_\omega(\tau)t}\|_2+\|e^{-w_0t}\|_2\right)dt\\
&+\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\Delta_{0,\omega}(w_\omega(\tau))\right\|_2\\
\le&\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\|\bar{\rho}\|_{L^2[0,\infty)}\left\{\int_0^\infty\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2^2dt\right\}^{\frac{1}{2}}\\
&+\sqrt{2}\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\tilde{a}(\tau)\right\|_2\left\{\int_0^\infty\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2^2dt\right\}^{\frac{1}{2}}\\
&\times\left\{\int_0^\infty\|e^{-w_\omega(\tau)t}\|_2^2dt+\int_0^\infty\|e^{-w_0t}\|_2^2dt\right\}^{\frac{1}{2}}+\left\|a_\omega(\tau)-\tilde{a}(\tau)\right\|_2\left\|\Delta_{0,\omega}(w_\omega(\tau))\right\|_2,
\end{aligned}
\tag{211}
$$

where the Cauchy-Schwartz inequality is repeatedly used.[21] By Lemma 50, for any $\omega>0$, $m\in\mathbb{N}_+$, $\theta_0\in\tilde{\Theta}_0$ and $\tau\in[0,\tau_1]$, we have $w_{\omega,k}(\tau)\ge w_l/2>0$, $k=1,2,\cdots,m$. This gives

$$
\begin{aligned}
&\int_0^\infty\left\|e^{-w_\omega(\tau)t}-e^{-w_0t}\right\|_2^2dt=\sum_{k=1}^m\int_0^\infty\left(e^{-w_{\omega,k}(\tau)t}-e^{-w_{0,k}t}\right)^2dt\\
=&\sum_{k=1}^m\left(\frac{1}{2w_{\omega,k}(\tau)}-\frac{2}{w_{\omega,k}(\tau)+w_{0,k}}+\frac{1}{2w_{0,k}}\right)=\sum_{k=1}^m\frac{(w_{\omega,k}(\tau)-w_{0,k})^2}{2w_{\omega,k}(\tau)w_{0,k}(w_{\omega,k}(\tau)+w_{0,k})}\\
\le&\frac{2}{3w_l^3}\|w_\omega(\tau)-w_0\|_2^2,
\end{aligned}
\tag{212}
$$

and similarly,

$$
\int_0^\infty\left\|e^{-w_\omega(\tau)t}\right\|_2^2dt\le\frac{m}{w_l},\qquad\int_0^\infty\|e^{-w_0t}\|_2^2dt\le\frac{m}{2w_l}.
\tag{213}
$$

By Lemma 27, we also have

$$
\left\|\Delta_{0,\omega}(w_\omega(\tau))\right\|_2^2\le\left\|\Delta_{0,\omega}^+(w_\omega(\tau))\right\|_2^2\le m\left[\Delta_{0,\omega}^+(w_l/2)\right]^2\lesssim me^{-\frac{w_l}{2\omega}}\left(c_2^{w_l^2/4}+c_3^{w_l/2}\right)
\tag{214}
$$

21. Here we also use the inequality $\|bb^\top-cc^\top\|_2\le\|b-c\|_2(\|b\|_2+\|c\|_2)$ for any vectors b,c.

holds for any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l\})$, where $c_2, c_3 > 1$ are constants only depending on c_1, t_0, and $\lesssim$ hides universal positive constants only related to ρ_0 and t_0, c_0, c_1. Notice that $\frac{d}{d\tau}\|a_\omega(\tau) - \tilde{a}(\tau)\|_2^2 = 2\|a_\omega(\tau) - \tilde{a}(\tau)\|_2 \frac{d}{d\tau}\|a_\omega(\tau) - \tilde{a}(\tau)\|_2$, combining with (211), (212), (213) and (214), we get

$$\begin{aligned}
&\frac{d}{d\tau}\|a_\omega(\tau) - \tilde{a}(\tau)\|_2 \\
\leq\ & 2\|\bar{\rho}\|_{L^2[0,\infty)} \left\{\int_0^\infty \left\|e^{-w_\omega(\tau)t} - e^{-w_0 t}\right\|_2^2 dt\right\}^{\frac{1}{2}} + 2\left\|\Delta_{0,\omega}(w_\omega(\tau))\right\|_2 \\
&+ 2\sqrt{2}\|\tilde{a}(\tau)\|_2 \left\{\int_0^\infty \left\|e^{-w_\omega(\tau)t} - e^{-w_0 t}\right\|_2^2 dt\right\}^{\frac{1}{2}} \left\{\int_0^\infty \|e^{-w_\omega(\tau)t}\|_2^2 dt + \int_0^\infty \|e^{-w_0 t}\|_2^2 dt\right\}^{\frac{1}{2}} \\
\lesssim\ & \|\bar{\rho}\|_{L^2[0,\infty)}\|w_\omega(\tau) - w_0\|_2 + \|\tilde{a}(\tau)\|_2\|w_\omega(\tau) - w_0\|_2\sqrt{m} + \sqrt{m}e^{-\frac{w_l}{4\omega}},
\end{aligned}$$

where $\lesssim$ hides universal positive constants only related to w_l, c_2, c_3. It is shown that $\|\bar{\rho}\|_{L^2[0,\infty)} \leq c^*/\sqrt{2w_l}$ (see the proof of Lemma 38). Combining with Lemma 49 and Lemma 51, we obtain that for any $\delta > 0$, with probability of at least $1 - \delta$ over the choice of w_0,

$$\begin{aligned}
\frac{d}{d\tau}\|a_\omega(\tau) - \tilde{a}(\tau)\|_2 &\lesssim \tau\left(\frac{1}{\sqrt{m}} + \tau\sqrt{m}\right)\left(\frac{c^*}{\sqrt{2w_l}} + (1 + C_\delta\sqrt{\tau})\right) + \sqrt{m}e^{-\frac{w_l}{4\omega}} \\
&\lesssim \tau(1 + C_\delta\sqrt{\tau})\left(\frac{1}{\sqrt{m}} + \tau\sqrt{m}\right) + \sqrt{m}e^{-\frac{w_l}{4\omega}},
\end{aligned}$$

where $\lesssim$ hides universal positive constants only related to a_l, a_r, w_l, w_r and a^*, ρ_0, and $C_\delta = \sqrt{1 + \ln(2/\delta)}$. Since $a_\omega(0) = \tilde{a}(0) = a_0$, we finally get

$$\begin{aligned}
\|a_\omega(\tau) - \tilde{a}(\tau)\|_2 &\lesssim \int_0^\tau \left[s(1 + C_\delta\sqrt{s})\left(\frac{1}{\sqrt{m}} + s\sqrt{m}\right) + \sqrt{m}e^{-\frac{w_l}{4\omega}}\right] ds \\
&= \tau^2\left[\frac{1}{\sqrt{m}}\left(\frac{1}{2} + \frac{2C_\delta\sqrt{\tau}}{5}\right) + \tau\sqrt{m}\left(\frac{1}{3} + \frac{2C_\delta\sqrt{\tau}}{7}\right)\right] + \tau\sqrt{m}e^{-\frac{w_l}{4\omega}},
\end{aligned}$$

which completes the proof. ■

(3) Result

Eventually, we can establish the optimization result of (172) under the gradient flow training dynamics (104).

Proof [Proof of Theorem 25] For any $\omega > 0$ and $m \in \mathbb{N}_+$, we have

$$\begin{aligned}
&\left\|[a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t)\right\|_{L^2[0,\infty)}^2 \\
\lesssim\ & \left\|\left(a_\omega(\tau) - \tilde{a}(\tau)\right)^\top e^{-w_\omega(\tau)t}\right\|_{L^2[0,\infty)}^2 + \left\|[\tilde{a}(\tau)]^\top\left(e^{-w_\omega(\tau)t} - e^{-w_0 t}\right)\right\|_{L^2[0,\infty)}^2 + \bar{J}(\tilde{a}(\tau), w_0).
\end{aligned}$$

By the Cauchy-Schwartz inequality, we get

$$\begin{aligned}
\left\|\left(a_\omega(\tau) - \tilde{a}(\tau)\right)^\top e^{-w_\omega(\tau)t}\right\|_{L^2[0,\infty)}^2 &\leq \|a_\omega(\tau) - \tilde{a}(\tau)\|_2^2 \int_0^\infty \|e^{-w_\omega(\tau)t}\|_2^2 dt, \\
\left\|[\tilde{a}(\tau)]^\top\left(e^{-w_\omega(\tau)t} - e^{-w_0 t}\right)\right\|_{L^2[0,\infty)}^2 &\leq \|\tilde{a}(\tau)\|_2^2 \int_0^\infty \|e^{-w_\omega(\tau)t} - e^{-w_0 t}\|_2^2 dt.
\end{aligned}$$

Combining with Lemma 49, Lemma 51, Lemma 52 and (212), (213), we obtain that for any $\omega \in (0, \min\{1/2, 1/t_0, 4c_1/w_l\})$, $m \in \mathbb{N}_+$, and for any $\delta > 0$, with probability of at least $1-\delta$ over the choice of $\theta_0 \in \widetilde{\Theta}_0$,

$$\begin{aligned}
\left\| [a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t) \right\|_{L^2[0,\infty)}^2 &\lesssim \tau^4 m \left(\frac{1}{\sqrt{m}} + \tau\sqrt{m} \right)^2 (1 + C_\delta\sqrt{\tau})^2 + \tau^2 m^2 e^{-\frac{w_l}{2\omega}} \\
&\quad + \frac{1}{m}(1 + C_\delta\sqrt{\tau})^2 \|w_\omega(\tau) - w_0\|_2^2 + \left(\frac{1}{m\tau} + \frac{1}{m} C_\delta^2 \right) \\
&\lesssim \tau^2 \left(\frac{1}{\sqrt{m}} + \tau\sqrt{m} \right)^2 (1 + C_\delta\sqrt{\tau})^2 \left(\tau^2 m + \frac{1}{m} \right) \\
&\quad + \tau^2 m^2 e^{-\frac{w_l}{2\omega}} + \left(\frac{1}{m\tau} + \frac{1}{m} C_\delta^2 \right), \quad \forall \tau \in [0, \tau_1],
\end{aligned}$$

where $\lesssim$ hides universal positive constants only related to a_l, a_r, w_l, w_r and a^*, ρ_0, t_0, c_0, c_1, and $C_\delta = \sqrt{1 + \ln(2/\delta)}$. Let $C'_\delta := (1 + C_\delta)^2$. For any $m \geq 1/\tau_1^{1/p}$, $p \in (1/3, 1)$, and any $\tau \in [m^{-p}/2, m^{-p}] \subset [0, \tau_1]$, we get

$$\begin{aligned}
&\left\| [a_\omega(\tau)]^\top e^{-w_\omega(\tau)t} - \bar{\rho}(t) \right\|_{L^2[0,\infty)}^2 \\
\lesssim\, & \frac{1}{m^{2p}} \left(\frac{1}{\sqrt{m}} + \frac{1}{m^{p-\frac{1}{2}}} \right)^2 C'_\delta \left(\frac{1}{m^{2p-1}} + \frac{1}{m} \right) + m^{2(1-p)} e^{-\frac{w_l}{2\omega}} + \left(\frac{2}{m^{1-p}} + \frac{1}{m} C'_\delta \right) \\
=\, & C'_\delta \left(\frac{2}{m^{4p}} + \frac{1}{m^{2p+2}} + \frac{1}{m^{6p-2}} + \frac{2}{m^{5p-1}} + \frac{2}{m^{3p+1}} + \frac{1}{m} \right) + m^{2(1-p)} e^{-\frac{w_l}{2\omega}} + \frac{2}{m^{1-p}} \\
\lesssim\, & C'_\delta \left(\frac{1}{m^{6p-2}} + \frac{1}{m} \right) + m^{2(1-p)} e^{-\frac{w_l}{2\omega}} + \frac{1}{m^{1-p}},
\end{aligned}$$

which completes the proof. ■

8. Conclusion

In this paper, we analyze the basic approximation and optimization aspects of using RNNs to learn input-output relationships involving temporal sequences in the linear, continuous-time setting. In both aspects, our analysis reveals that the dynamical nature of the problem connects the idea of memory and learning in a precise way. In particular, we theoretically and numerically uncover phenomena called the *curse of memory*, and reveals two of its facets: when the target relationship to be learned has long-term memory, both approximation and optimization become exceedingly difficult. The analysis makes concrete the heuristic observations of the adverse effect of memory on learning with RNNs. Moreover, it quantifies the interaction between the model architectures (RNN functionals) and the data structures (target functionals). The latter is a much less-studied topic. The current analyses focus on the linear case. A natural question is whether such interactions between approximation, optimization and memory structures persist in the nonlinear case. We showed using numerical experiments in Figure 4 that even with nonlinear activation functions, the behavior of RNNs with respect to memory structures remains similar. Nevertheless, it is

important to develop the mathematical theory for this case, and the key ingredient is to understand whether point-wise non-linearity can induce fundamental changes in memory patterns for functionals parameterized by otherwise linear dynamics. More broadly, the approach here may act as a basic starting point for understanding partially-observed time series data in general, including gated variants of RNNs (Hochreiter and Schmidhuber, 1997; Cho et al., 2014), and other models such as transformers and convolution-based approaches such as WaveNet (Vaswani et al., 2017; Oord et al., 2016). These are certainly worthy of future exploration.

Acknowledgments

Qianxiao Li is supported by the National Research Foundation, Singapore, under the NRF fellowship (NRF-NRFF13-2021-0005).

Appendix A. Landscape Analysis: Weights Degeneracy

As mentioned in Remark 37, we can perform a global landscape analysis on the loss function based on the idea of weights degeneracy, which arises from Definition 29. Recall that the loss function reads

$$\min_{(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m} J_m(a,w) := \int_0^\infty \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t)\right)^2 dt. \tag{215}$$

Here we use the subscript m to emphasize the effect of model capacity, since different choices of m are discussed in the following analysis. The main results of the appendix are summarized as follows.

- In Theorem 57, we prove that the loss function has infinitely many critical points, which form a factorial number of affine spaces (affine spaces);
- In Theorem 58, we prove that such (critical) affine spaces are much more than global minimizers provided the target as an exponential sum;[22]
- In Theorem 61, we prove that on such (critical) affine spaces, the Hessian is singular in the sense of processing multiple zero eigenvalues;
- In Proposition 71, we prove that the (critical) affine spaces contain both saddles and degenerate stable points which are not global optimal.

Instead of a local dynamical analysis in the main text, we generalize similar methods to a *global landscape analysis* here, and the results hold for the loss function associated with *general targets*. More specifically, these results complement our main results (see Section 6.2.1) in the following aspects.

22. The global minimizers are distinct when the target is an exponential sum. Here we compare the number of (critical) affine spaces with the number of global minimizers (both of them are finite). When the target is not an exponential sum, the same conclusion holds if there are still finite number of global minimizers. See Remark 60 in Section A.2 for details.

- It is shown that the weights degeneracy is quite common in the whole landscape of the loss function. Unfortunately, weights degeneracy often worsens the landscape to a large extent;
- It is shown that the weights degeneracy leads to a large number of stable areas (i.e. critical affine spaces), but most of them contribute to non-global minimizers;
- It is shown that these stable areas can also be quite flat, which often connect with local plateaus;
- For the structure of these stable areas, there are both saddles and degenerate critical points (not global optimal). In certain regimes, even saddles can be rather difficult to escape (Theorem 20).

As a consequence, the optimization problem of linear RNNs is globally and essentially difficult to solve.

This section consists of three parts: in Section A.1, we give main results provided the existence of weights degeneracy; in Section A.2, we give sufficient conditions to guarantee the existence. A low-dimensional example is investigated in Section A.3. Since the key observation to use weights degeneracy is to notice the permutation symmetry of coordinates of gradients, we also called it "symmetry analysis".

A.1 Generic Theories

We begin with the following definition, which describes the idea of weights degeneracy in a natural and rigorous way.

Definition 53 *(coincided critical solutions and affine spaces) Let $d \in \mathbb{N}_+$ and $1 \leq d \leq m$. We call that (a, w) is a d-coincided critical solution of J_m, if $\nabla J_m(a, w) = 0$, and $w = (w_i) \in \mathbb{R}_+^m$ has d different components. The coincided critical affine spaces are defined as coincided critical solutions that form affine spaces.*

To guarantee the existence of such solutions, it is necessary to have the following definition.

Definition 54 *$(\hat{a}, \hat{w}) \in \mathbb{R}^m \times \mathbb{R}_+^m$ is called the non-degenerate global minimizer of J_m, if and only if*

$$J_m(\hat{a}, \hat{w}) = \inf_{(a,w)\in\mathbb{R}^m\times\mathbb{R}_+^m} J_m(a, w), \tag{216}$$

and $(\hat{a}, \hat{w})$ takes a non-degenerate form

$$\hat{a}_i \neq 0,\ \hat{w}_i \neq \hat{w}_j \text{ for } i \neq j, \quad i, j = 1, 2, \cdots, m. \tag{217}$$

For convenience, we also define an index set

$$\mathcal{N} := \{n \in \mathbb{N}_+ :\ J_m \text{ has non-degenerate global minimizers for any } m \leq n\}, \tag{218}$$

which is used frequently in the following analysis. For any $f \in L^2[0, \infty)$, let $\mathcal{L}[f]$ be the Laplace transform of f, i.e. $\mathcal{L}[f](s) = \int_0^\infty e^{-st} f(t)dt,\ s > 0$.

We begin with the following lemma.

Lemma 55 *Assume that ρ is smooth and $\sqrt{w}\,|\mathcal{L}[\rho](w)| \to 0$ as $w \to 0^+$ and $w \to \infty$. Then we have $1 \in \mathcal{N}$ and thus $\mathcal{N} \neq \varnothing$.*

Proof We aim to show that there exists $\hat{a} \neq 0$ and $\hat{w} > 0$, such that

$$J_1(\hat{a}, \hat{w}) = \inf_{(a,w)\in\mathbb{R}\times\mathbb{R}_+} J_1(a, w). \tag{219}$$

The basic idea is to limit the unbounded domain $\mathbb{R} \times \mathbb{R}_+$ to a compact set without effecting the minimization of $J_1(a, w)$. We have

$$\begin{aligned}
\min_{a,w>0} J_1(a, w) &= \min_{w>0} \min_{a} \left\{ \frac{1}{2w} \cdot a^2 - 2\mathcal{L}[\rho](w) \cdot a + \|\rho\|^2_{L^2[0,\infty)} \right\} \\
&= \min_{w>0} \min_{a} \left\{ \frac{1}{2w} \big(a - 2w\mathcal{L}[\rho](w)\big)^2 + \Big[\|\rho\|^2_{L^2[0,\infty)} - 2w(\mathcal{L}[\rho](w))^2 \Big] \right\} \\
&= \min_{w>0} \left\{ \|\rho\|^2_{L^2[0,\infty)} - 2w(\mathcal{L}[\rho](w))^2 \right\} = J_1(a(w), w),
\end{aligned}$$

where $a(w) := 2w\mathcal{L}[\rho](w)$. Write $h(w) := J_1(a(w), w)$, then $h(0^+) = h(\infty) = \|\rho\|^2_{L^2[0,\infty)}$. Obviously $h(w) < \|\rho\|^2_{L^2[0,\infty)}$ for any $w > 0$, hence

$$\min_{w>0} h(w) = \min_{w\in[w_{lb},w_{ub}]} h(w), \quad 0 < w_{lb} < w_{ub} < \infty,$$

which implies

$$\min_{a,w>0} J_1(a, w) = \min_{w>0} J_1(a(w), w) = \min_{w\in[w_{lb},w_{ub}]} J_1(a(w), w).$$

That is to say, the minimization of $J_1(a, w)$ can be equivalently performed on a 2-dimensional smooth curve
$(w, a(w))_{w\in[w_{lb},w_{ub}]}$, which is certainly a compact set. By continuity, $J_1(a, w)$ has global minimizers, say $(\hat{a}, \hat{w})$. Obviously $\hat{w} > 0$ and $\hat{a} = a(\hat{w}) \neq 0$ (since $\hat{a} = 0$ implies $J_1(\hat{a}, w) = \|\rho\|^2_{L^2[0,\infty)}$, certainly not a minimum), which completes the proof. ■

Remark 56 *If the ground truth is an exponential sum, i.e. $\rho(t) = \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t}$, we know ρ is smooth and $\sqrt{w}\,|\mathcal{L}[\rho](w)| \to 0$ as $w \to 0^+$ and $w \to \infty$; hence $1 \in \mathcal{N}$ by Lemma 55, and thus $\mathcal{N} \neq \varnothing$. In fact, $\mathcal{L}[\rho](w) = \sum_{j=1}^{m^*} \frac{a_j^*}{w+w_j^*}$ implies that $\mathcal{L}[\rho](w) = \mathcal{O}(1)$ when $w \to 0^+$, and $\mathcal{L}[\rho](w) = \mathcal{O}(1/w)$ when $w \to \infty$.*

Theorem 57 *Assume that $\mathcal{N} \neq \varnothing$ with $\mathcal{N}$ defined as (218). Let $M := \sup \mathcal{N}$. Then for any $m \in \mathbb{N}_+$, $1 \leq d \leq \min\{m, M\}$, there exists at least $d! \left\{ {m \atop d} \right\}$ d-coincided critical affine spaces of J_m,[23] where $\left\{ {m \atop d} \right\} \in \mathbb{N}_+$ is called the Stirling number of the second kind.*

23. Certainly, the affine spaces degenerate to distinct points when $d = m$. For sufficient conditions to guarantee $M > 1$ (to give meaningful results), see Theorem 68 and Remark 69 in Section A.2.

Proof (i) Existence. The key observation is the permutation symmetry of ∇J_m: by (97) and (98), if $a_i = a_j$ and $w_i = w_j$ for some $i \neq j$, then $\frac{\partial J_m}{\partial a_i} = \frac{\partial J_m}{\partial a_j}$ and $\frac{\partial J_m}{\partial w_i} = \frac{\partial J_m}{\partial w_j}$.

For any $m, d \in \mathbb{N}_+$, $1 \leq d \leq m$, suppose that $w = (w_i) \in \mathbb{R}_+^m$ has d different components. Then for any partition $\mathcal{P}$: $\{1, \cdots, m\} = \cup_{j=1}^d \mathcal{I}_j$ with $\mathcal{I}_{j_1} \cap \mathcal{I}_{j_2} = \varnothing$ for any $j_1 \neq j_2$, $j_1, j_2 = 1, \cdots, d$, define the affine space

$$\mathcal{M}_{\mathcal{P},(b,v),(m,d)} := \left\{ (a, w) \in \mathbb{R}^m \times \mathbb{R}_+^m : w_i = v_j \text{ for any } i \in \mathcal{I}_j, \ \sum_{i \in \mathcal{I}_j} a_i = b_j, \quad j = 1, \cdots, d \right\}$$

for some $(b, v) \in \mathbb{R}^d \times \mathbb{R}_+^d$, where v has exactly d different components. Therefore, for any $(a, w) \in \mathcal{M}_{\mathcal{P},(b,v),(m,d)}$, we have

$$J_m(a, w) = \left\| \sum_{j=1}^d \sum_{i \in \mathcal{I}_j} a_i e^{-w_i t} - \rho(t) \right\|_{L^2[0,\infty)}^2 = \left\| \sum_{j=1}^d b_j e^{-v_j t} - \rho(t) \right\|_{L^2[0,\infty)}^2 = J_d(b, v),$$

and similarly

$$\frac{\partial J_m}{\partial a_k}(a, w) = 2 \int_0^\infty e^{-v_s t} \left(\sum_{j=1}^d b_j e^{-v_j t} - \rho(t) \right) dt, \quad k \in \mathcal{I}_s, \ s = 1, 2, \cdots, d,$$

$$\frac{\partial J_m}{\partial w_k}(a, w) = 2 a_k \int_0^\infty (-t) e^{-v_s t} \left(\sum_{j=1}^d b_j e^{-v_j t} - \rho(t) \right) dt, \quad k \in \mathcal{I}_s, \ s = 1, 2, \cdots, d.$$

Notice that

$$\frac{\partial J_d}{\partial b_s}(b, v) = 2 \int_0^\infty e^{-v_s t} \left(\sum_{j=1}^d b_j e^{-v_j t} - \rho(t) \right) dt, \quad s = 1, 2, \cdots, d,$$

$$\frac{\partial J_d}{\partial v_s}(b, v) = 2 b_s \int_0^\infty (-t) e^{-v_s t} \left(\sum_{j=1}^d b_j e^{-v_j t} - \rho(t) \right) dt, \quad s = 1, 2, \cdots, d,$$

we have

$$\frac{\partial J_m}{\partial a_k}(a, w) = \frac{\partial J_d}{\partial b_s}(b, v), \quad b_s \frac{\partial J_m}{\partial w_k}(a, w) = a_k \frac{\partial J_d}{\partial v_s}(b, v), \quad k \in \mathcal{I}_s, \ s = 1, 2, \cdots, d, \tag{220}$$

which is in fact a model reduction.[24] Since $d \leq \min\{m, M\}$, $d \in \mathcal{N}$. In fact, for any $k \in \mathbb{N}_+$, if $k \notin \mathcal{N}$, there exists $i \leq k$ such that $J_{(i)}$ has no non-degenerate global minimizers, we have $j \notin \mathcal{N}$ for any $j \geq i$, hence $M \leq i - 1 \leq k - 1$. Hence $M = \infty$ implies $\mathcal{N} = \mathbb{N}_+$ and $M < \infty$

24. By considering the gradient flow dynamic of J_d instead of J_m, a model reduction (from m-dimensional to d-dimensional) is almost completed on $\mathcal{M}_{\mathcal{P},(b,v),(m,d)}$, except for the trivial degenerate cases (e.g. $a_k = 0$ or $b_s = 0$).

implies $M \in \mathcal{N}$, and both of them lead to $d \in \mathcal{N}$. Therefore, J_d has non-degenerate global minimizers, i.e. there exists $(\hat{b}, \hat{v}) \in \mathbb{R}^d \times \mathbb{R}^d_+$ such that

$$J_d(\hat{b}, \hat{v}) = \inf_{(b,v)\in\mathbb{R}^d\times\mathbb{R}^d_+} J_d(b, v), \tag{221}$$

and $(\hat{b}, \hat{v})$ takes a non-degenerate form

$$\hat{b}_i \neq 0, \ \hat{v}_i \neq \hat{v}_j \text{ for any } i \neq j, \quad i, j = 1, 2, \cdots, d. \tag{222}$$

By (221), we get $\nabla J_d(\hat{b}, \hat{v}) = 0$. Combining with (220) and (222), we obtain $\nabla J_m(\hat{a}, \hat{w}) = 0$ for any $(\hat{a}, \hat{w}) \in \mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$, i.e. $(\hat{a}, \hat{w})$ belongs to a d-coincided critical affine space. Note that the affine space is with the dimension $\sum_{j=1}^{d}(|\mathcal{I}_j| - 1) = m - d$, since there are d linear equality constrains on the m-dimensional vector a.

(ii) Counting. By the structure of affine spaces discussed above, we can identify different affine spaces with respect to the partition $\mathcal{P}$. For counting the number of different partitions $\mathcal{P}$: $\{1, \cdots, m\} = \cup_{j=1}^{d}\mathcal{I}_j$, it can be decomposed into the following two steps. First, partitioning a set of m labelled objects into d non-empty unlabelled subsets. By definition, the answer is the Stirling number of the second kind $\left\{ {m \atop d} \right\}$. Second, assign each partition to $\mathcal{I}_1, \cdots, \mathcal{I}_d$ accordingly. There are $d!$ ways in total. Therefore, the number of d-coincided critical affine spaces is at least $d! \left\{ {m \atop d} \right\}$. The proof is completed. ■

Combining Lemma 55, Remark 56 and Theorem 57 gives the following theorem, which states that there are much more saddles and degenerate stable points which are not global optimal than global minimizers in the landscape (provided the target as an exponential sum).

Theorem 58 *Fix any $m \in \mathbb{N}_+$ relatively large. Consider the loss J_m with the ground truth as a non-degenerate exponential sum, i.e. $\rho(t) = \sum_{j=1}^{m} a_j^* e^{-w_j^* t}$, where $a_j^* \neq 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j = 1, \cdots, m$. Assume that $m \in \mathcal{N}$ with $\mathcal{N}$ defined in (218).*[25] *Then in the landscape of J_m, the number of coincided critical affine spaces is at least $Poly(m)$ times larger than the number of global minimizers.*

Proof (i) Global minimizers. Since the ground truth is an exponential sum, we have $J_m(a, w) \geq 0$ and $J_m(\bar{a}^*, \bar{w}^*) = 0$, where $\bar{a}^* = Pa^*$ and $\bar{w}^* = Pw^*$ with $P \in \mathbb{R}^{m\times m}$ to be some permutation matrix. Next we show J_m has no other global minimizers.

Suppose $J_m(a, w) = 0$, we have

$$\sum_{i=1}^{m} a_i e^{-w_i t} - \sum_{j=1}^{m} a_j^* e^{-w_j^* t} = 0, \quad t \geq 0. \tag{223}$$

It is easy to get that for any $j = 1, 2, \cdots, m$, there exists $i(j)$ such that $w_{i(j)} = w_j^*$. Otherwise, if $w_i \neq w_j^*$, $i = 1, \cdots, m$, by (123) or Lemma 28, we have $a_j^* = 0$, which is a

25. Although the assumption $m \in \mathcal{N}$ seems strong, we will provide sufficient conditions to guarantee its validity in Section A.2. See Theorem 70.

contradiction. Notice that $w_i^* \neq w_j^*$ for any $i \neq j$, different w_j^* will correspond to different w_i, hence the correspondence is one-to-one. Therefore, let $w_i = w_{j(i)}^*$, (223) can be rewritten as

$$0 = \sum_{i=1}^{m} a_i e^{-w_{j(i)}^* t} - \sum_{i=1}^{m} a_{j(i)}^* e^{-w_{j(i)}^* t} = \sum_{i=1}^{m} (a_i - a_{j(i)}^*) e^{-w_{j(i)}^* t}, \quad t \geq 0.$$

Again by Lemma 28, we have $a_i = a_{j(i)}^*$. That is to say, $J_m(a, w) = 0$ implies $a = Pa^*$ and $w = Pw^*$ with $P \in \mathbb{R}^{m \times m}$ to be some permutation matrix. This gives $m!$ global minimizers.

(ii) Coincided critical affine spaces. Obviously $\mathcal{N} \neq \varnothing$, and $M = \sup \mathcal{N} \geq m$. According to Theorem 57, for any d, $1 \leq d \leq \min\{m, M\} = m$, we have at least $d! \left\{ {m \atop d} \right\}$ d-coincided critical affine spaces of J_m. By (i), for any $d \leq m-1$, there are no global minimizers in these affine spaces. Counting the total number

$$\sum_{d=1}^{m-1} d! \left\{ {m \atop d} \right\}. \tag{224}$$

(iii) Comparison. To give a bound between (224) and $m!$, we need an elementary recurrence

$$\left\{ {m \atop d} \right\} = d \left\{ {m-1 \atop d} \right\} + \left\{ {m-1 \atop d-1} \right\}.$$

- For $d = m-1$, let $p_m := \left\{ {m \atop m-1} \right\}$, then

$$p_m = (m-1) \left\{ {m-1 \atop m-1} \right\} + \left\{ {m-1 \atop m-2} \right\} = (m-1) + p_{m-1} = \cdots = \frac{m(m-1)}{2}.$$

- For $d = m-2$, let $q_m := \left\{ {m \atop m-2} \right\}$, then

$$\begin{aligned} q_m &= (m-2) \left\{ {m-1 \atop m-2} \right\} + \left\{ {m-1 \atop m-3} \right\} = (m-2) p_{m-1} + q_{m-1} = \cdots \\ &= \frac{1}{24} [2(m-2)(m-1)(2m-3) + 3(m-2)^2 (m-1)^2]. \end{aligned}$$

Combining above gives

$$\frac{1}{m!} \sum_{d=1}^{m-1} d! \left\{ {m \atop d} \right\} > \frac{1}{m!} [(m-1)! p_m + (m-2)! q_m] = \frac{(m+1)(3m-2)}{24},$$

which is a quadratic polynominal on m. The proof is completed. ■

Remark 59 *We only take the last two terms of (224) for a lower bound, which is obviously rather loose. In principle, a* Poly(m) *bound with higher degrees can be similarly obtained. That is to say, on one hand, there are infinitely many critical points forming affine spaces in the landscape of J_m; on the other hand, we deduce that even only counting the affine spaces, there are still much less global minimizers (given the width m relatively large).*

Remark 60 *When the target ρ is not an exponential sum, it is straightforward to verify Theorem 58 still holds if there are finite number (with the scale of no more than factorial) of global minimizers.*

Now we get down to investigate $\nabla^2 J_m$ on the above coincided critical affine spaces. It is shown that $\nabla^2 J_m$ is singular and can have multiple zero eigenvalues.

Theorem 61 *Fix any $m, d \in \mathbb{N}_+$, $1 \le d \le m$. On the d-coincided critical affine spaces of J_m,*[26] *$\nabla^2 J_m$ is with rank at most $m+d$, and hence has at least $m-d$ zero eigenvalues.*

Proof A straightforward computation shows that, for $k, l = 1, 2, \cdots, m$,

$$\frac{\partial^2 J_m}{\partial a_k \partial a_l}(a, w) = \frac{2}{w_k + w_l}, \tag{225}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_l}(a, w) = \frac{-2a_l}{(w_k + w_l)^2}, \quad k \neq l, \tag{226}$$

$$\frac{\partial^2 J_m}{\partial a_k \partial w_k}(a, w) = \frac{-a_k}{2w_k^2} + 2\int_0^\infty (-t) e^{-w_k t} \left(\sum_{i=1}^m a_i e^{-w_i t} - \rho(t) \right) dt. \tag{227}$$

Let the induced d-coincided critical affine space be $\mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$, as is derived in the proof of Theorem 57. Since $(\hat{b}, \hat{v})$ is the non-degenerate global minimizer of J_d, we have

$$\begin{aligned}\int_0^\infty (-t) e^{-\hat{w}_k t} \left(\sum_{i=1}^m \hat{a}_i e^{-\hat{w}_i t} - \rho(t) \right) dt &= \int_0^\infty (-t) e^{-\hat{v}_s t} \left(\sum_{j=1}^d \hat{b}_j e^{-\hat{v}_j t} - \rho(t) \right) dt \\ &= \frac{1}{2\hat{b}_s} \frac{\partial J_d}{\partial v_s}(\hat{b}, \hat{v}) = 0, \quad k \in \mathcal{I}_s,\ s = 1, 2, \cdots, d,\end{aligned}$$

for any $(\hat{a}, \hat{w}) \in \mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$. This gives

$$\frac{\partial^2 J_m}{\partial a_k \partial w_k}(\hat{a}, \hat{w}) = \frac{-\hat{a}_k}{2\hat{w}_k^2}. \tag{228}$$

Now we show that, for any $i, j \in \mathcal{I}_s$, $i \neq j$, $s = 1, 2, \cdots, d$, the i-th row and j-th row of $\nabla^2 J_m(\hat{a}, \hat{w})$ are the same. In fact, for any $k = 1, \cdots, m$, let $k \in \mathcal{I}_{s'}$, then by (225),

$$\frac{\partial^2 J_m}{\partial a_i \partial a_k}(\hat{a}, \hat{w}) = \frac{2}{\hat{w}_i + \hat{w}_k} = \frac{2}{\hat{v}_s + \hat{v}_{s'}}, \qquad \frac{\partial^2 J_m}{\partial a_j \partial a_k}(\hat{a}, \hat{w}) = \frac{2}{\hat{w}_j + \hat{w}_k} = \frac{2}{\hat{v}_s + \hat{v}_{s'}}.$$

26. That is, the affine space $\mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$ induced by non-degenerate global minimizers of J_d. See details in the proof of Theorem 57.

For $k \neq i$ and $k \neq j$, (226) gives

$$\frac{\partial^2 J_m}{\partial a_i \partial w_k}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_k}{(\hat{w}_i + \hat{w}_k)^2} = \frac{-2\hat{a}_k}{(\hat{v}_s + \hat{v}_{s'})^2}, \qquad \frac{\partial^2 J_m}{\partial a_j \partial w_k}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_k}{(\hat{w}_j + \hat{w}_k)^2} = \frac{-2\hat{a}_k}{(\hat{v}_s + \hat{v}_{s'})^2}.$$

Together with (228), for $k = i \neq j$,

$$\frac{\partial^2 J_m}{\partial a_i \partial w_k}(\hat{a}, \hat{w}) = \frac{-\hat{a}_i}{2\hat{w}_i^2} = \frac{-\hat{a}_i}{2\hat{v}_s^2}, \qquad \frac{\partial^2 J_m}{\partial a_j \partial w_k}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_i}{(\hat{w}_j + \hat{w}_i)^2} = \frac{-\hat{a}_i}{2\hat{v}_s^2},$$

and similarly for $k = j \neq i$,

$$\frac{\partial^2 J_m}{\partial a_i \partial w_k}(\hat{a}, \hat{w}) = \frac{-2\hat{a}_j}{(\hat{w}_i + \hat{w}_j)^2} = \frac{-\hat{a}_j}{2\hat{v}_s^2}, \qquad \frac{\partial^2 J_m}{\partial a_j \partial w_k}(\hat{a}, \hat{w}) = \frac{-\hat{a}_j}{2\hat{w}_j^2} = \frac{-\hat{a}_j}{2\hat{v}_s^2}.$$

That is to say, there are at most $m+d$ different rows in the symmetric matrix $\nabla^2 J_m(\hat{a}, \hat{w}) \in \mathbb{R}^{2m \times 2m}$, hence $\operatorname{rank}\big(\nabla^2 J_m(\hat{a}, \hat{w})\big) \leq m + d$. Therefore, the number of zero eigenvalues of $\nabla^2 J_m(\hat{a}, \hat{w}) \geq \dim\big\{x \in \mathbb{R}^{2m} : \nabla^2 J_m(\hat{a}, \hat{w}) \cdot x = 0\big\} = 2m - \operatorname{rank}(\nabla^2 J_m(\hat{a}, \hat{w})) \geq m - d$. The proof is completed. ■

Remark 62 *The bound in Theorem 61 is not sharp. The estimate on* $\operatorname{rank}\big(\nabla^2 J_m\big)$ *here is loose since only rows with the same elements are considered. In practice (numerical tests), it is often the case that there are more zero eigenvalues of* $\operatorname{rank}\big(\nabla^2 J_m\big)$ *on the coincided critical affine space* $\mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$.

Remark 63 *Theorem 61 shows that, there are local plateaus around the d-coincided critical affine spaces* $\mathcal{M}_{\mathcal{P},(\hat{b},\hat{v}),(m,d)}$ *for* $d \leq m - 1$. *In addition, the 0-eigenspace of* J_m *is higher-dimensional for smaller d, which may suggest that one can stuck on plateaus more easily.*

A.2 Sufficient Conditions

There is still a gap when connecting Theorem 57 and Theorem 58. That is, it is necessary to guarantee sup $\mathcal{N}$ relatively large, i.e. $J_1, J_2, \cdots, J_d$ all have non-degenerate global minimizers for d as large as possible. Motivated by Kammler (1979), we can give some sufficient conditions by limiting the ground truth ρ within a smaller function space, the so-called completely monotonic functions.

Definition 64 $F \in C[0, \infty] \cap C^\infty(0, \infty)$ *is called completely monotonic, if and only if*

$$(-1)^n F^{(n)}(t) \geq 0, \quad 0 < t < \infty, \ n = 0, 1, \cdots,$$

and $F(\infty) = 0$.

Remark 65 *Several examples of completely monotonic functions:*

- $\rho(t) = 1/(1 + t)^\alpha$ *for any* $\alpha > 0$;

- *The non-degenerate exponential sum with positive coefficients*

$$\rho(t) = \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t}, \quad 0 \le w_1^* < \cdots < w_{m^*}^*,\ a_j^* > 0,\ j = 1, 2, \cdots, m^*.$$

Since the space of exponential sums is not close, we turn to consider the problem of finding a best approximation to a given $\rho \in L^2[0,\infty)$ from the set

$$V_d(\mathbb{R}_+) := \left\{ \hat{\rho} \in C^d[0,\infty) : [(D+w_1)\cdots(D+w_d)]\hat{\rho} = 0 \text{ for some } w_1, \cdots w_d \in \mathbb{R}_+ \right\} \quad (229)$$

with respect to the common L^2-norm, i.e. $\inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho} - \rho\|_{L^2[0,\infty)}$, where D denotes the common differential operator. Obviously $V_d(\mathbb{R}_+) \subset L^2[0,\infty)$ and $V_d(\mathbb{R}_+) \subsetneq V_{d+1}(\mathbb{R}_+)$ for any $d \in \mathbb{N}_+$.

The work Kammler (1979) proves the following theorem.

Theorem 66 *Assume $\rho \in L^2[0,\infty)$ to be completely monotonic. Then there exists a best approximation $\hat{\rho}^0$ to ρ in $V_d(\mathbb{R}_+)$, i.e.*

$$\|\hat{\rho}^0 - \rho\|_{L^2[0,\infty)} = \inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho} - \rho\|_{L^2[0,\infty)}. \quad (230)$$

When $\rho \notin V_d(\mathbb{R}_+)$, any such best approximation admits a non-degenerate form

$$\hat{\rho}^0(t) = \sum_{j=1}^{d} \hat{b}_j e^{-\hat{v}_j t}, \quad 0 < \hat{v}_1 < \cdots < \hat{v}_d,\ \hat{b}_j > 0,\ j = 1, 2, \cdots, d, \quad (231)$$

and satisfies the generalized Aigrain-Williams equations

$$\mathcal{L}[\hat{\rho}^0](\hat{v}_j) = \mathcal{L}[\rho](\hat{v}_j), \quad j = 1, 2, \cdots, d, \quad (232)$$

$$\frac{d}{ds}\mathcal{L}[\hat{\rho}^0](s)\Big|_{s=\hat{v}_j} = \frac{d}{ds}\mathcal{L}[\rho](s)\Big|_{s=\hat{v}_j}, \quad j = 1, 2, \cdots, d. \quad (233)$$

Note that (230) and (231) are pretty similar to Definition 54, except for a different choice of hypothesis function space. Now we show a connection between these two problems.

Theorem 67 *Assume $\rho \in L^2[0,\infty)$ to be completely monotonic, and $\rho \notin V_d(\mathbb{R}_+)$ for some $d \in \mathbb{N}_+$. Then J_d has non-degenerate global minimizers $(\hat{b}, \hat{v}) \in \mathbb{R}^d \times \mathbb{R}_+^d$.*

Proof According to Theorem 66, there exists a non-degenerate best approximation $\hat{\rho}^0$ to ρ from $V_d(\mathbb{R}_+)$, i.e.

$$\|\hat{\rho}^0 - \rho\|_{L^2[0,\infty)} = \inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho} - \rho\|_{L^2[0,\infty)}, \quad (234)$$

$$\hat{\rho}^0(t) = \sum_{j=1}^{d} \hat{b}_j e^{-\hat{v}_j t}, \quad 0 < \hat{v}_1 < \cdots < \hat{v}_d,\ \hat{b}_j > 0,\ j = 1, 2, \cdots, d. \quad (235)$$

We aim to prove $J_d(\hat{b},\hat{v}) = \inf\limits_{(b,v)\in\mathbb{R}^d\times\mathbb{R}_+^d} J_d(b,v)$. Define the following subsets of exponential sums

$$\mathcal{V}_d(\mathbb{R}_+) := \left\{\hat{\rho} : \hat{\rho}(t) = \sum_{i=1}^{d} a_i e^{-w_i t}, a_i \in \mathbb{R}, w_i > 0\right\},$$
$$\mathcal{V}_{d,k}(\mathbb{R}_+) := \left\{\hat{\rho} \in \mathcal{V}_d(\mathbb{R}_+) : w = (w_i) \text{ has } k \text{ different components}\right\}, \quad k = 1, 2, \cdots, d,$$

then we have $\inf\limits_{(b,v)\in\mathbb{R}^d\times\mathbb{R}_+^d} J_d(b,v) = \inf\limits_{\hat{\rho}\in\mathcal{V}_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|^2_{L^2[0,\infty)}$. It is straightforward to verify that $\mathcal{V}_d(\mathbb{R}_+) = \bigcup_{k=1}^d \mathcal{V}_{d,k}(\mathbb{R}_+)$, and $\mathcal{V}_{d,k}(\mathbb{R}_+) = \mathcal{V}_{k,k}(\mathbb{R}_+) \subsetneq V_k(\mathbb{R}_+)$ for $k = 1, \cdots, d$. By (234), we get

$$\|\hat{\rho}^0-\rho\|^2_{L^2[0,\infty)} = \inf_{\hat{\rho}\in V_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|^2_{L^2[0,\infty)} \le \inf_{\hat{\rho}\in\mathcal{V}_{d,d}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|^2_{L^2[0,\infty)}.$$

Since $\hat{\rho}^0 \in \mathcal{V}_{d,d}(\mathbb{R}_+)$, we have

$$J_d(\hat{b},\hat{v}) = \|\hat{\rho}^0-\rho\|^2_{L^2[0,\infty)} = \inf_{\hat{\rho}\in\mathcal{V}_{d,d}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|^2_{L^2[0,\infty)}.$$

The last task is to show $\inf\limits_{\hat{\rho}\in\mathcal{V}_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)} = \inf\limits_{\hat{\rho}\in\mathcal{V}_{d,d}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)}$. In fact, for any $\hat{\rho} \in \mathcal{V}_{k,k}$, $\hat{\rho}(t) = \sum_{i=1}^k a_i e^{-w_i t}$, let $\tilde{a} := (a_1, \cdots, a_k, 0)$, $\quad \tilde{w} := (w_1, \cdots, w_k, 1 + \max\limits_{1\le i\le k} w_i)$, we get $\hat{\rho}(t) := \sum_{i=1}^{k+1} \tilde{a}_i e^{-\tilde{w}_i t} \in \mathcal{V}_{k+1,k+1}$, which implies $\mathcal{V}_{k,k} \subset \mathcal{V}_{k+1,k+1}$. Therefore,

$$\begin{aligned}\inf_{\hat{\rho}\in\mathcal{V}_d(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)} &= \inf_{\hat{\rho}\in\bigcup_{k=1}^d \mathcal{V}_{d,k}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)} = \min_{1\le k\le d}\left\{\inf_{\hat{\rho}\in\mathcal{V}_{d,k}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)}\right\} \\ &= \min_{1\le k\le d}\left\{\inf_{\hat{\rho}\in\mathcal{V}_{k,k}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)}\right\} \ge \inf_{\hat{\rho}\in\mathcal{V}_{d,d}(\mathbb{R}_+)} \|\hat{\rho}-\rho\|_{L^2[0,\infty)},\end{aligned}$$

which completes the proof. ■

Combining Theorem 57 and Theorem 67 immediately gives the following result.

Theorem 68 *Assume $\rho \in L^2[0,\infty)$ to be completely monotonic, and $\rho \notin V_1(\mathbb{R}_+)$. Let $\mathcal{D} := \{d \in \mathbb{N}_+ : \rho \notin V_d(\mathbb{R}_+)\}$, $D_0 := \sup \mathcal{D}$ and write $m' := \min\{m, D_0\}$. Then the total number of coincided critical affine spaces of J_m is at least $\sum_{d=1}^{m'} d! \left\{ m \atop d \right\}$.*

Proof We have $1 \in \mathcal{D}$ and thus $\mathcal{D} \neq \varnothing$, $D_0 \ge 1$. Since $V_d(\mathbb{R}_+) \subsetneq V_{d+1}(\mathbb{R}_+)$ for any $d \in \mathbb{N}_+$, we have $\mathcal{D} = \{1, 2, \cdots, D_0\}$ if $D_0 < \infty$, and $\mathcal{D} = \mathbb{N}_+$ if $D_0 = \infty$.[27] Both of them gives $\{1, 2, \cdots, m'\} \subset \mathcal{D}$, i.e. $\rho \notin V_k(\mathbb{R}_+)$ for any $k \le m'$. By Theorem 67, $J_{(k)}$ has

27. In fact, $V_d(\mathbb{R}_+) \subsetneq V_{d+1}(\mathbb{R}_+)$ for any $d \in \mathbb{N}_+$ implies if $\rho \notin V_d(\mathbb{R}_+)$, $\rho \notin V_k(\mathbb{R}_+)$ for any $k \le d$, i.e. $d \in \mathcal{D} \Rightarrow k \in \mathcal{D}$ for any $k \le d$; otherwise, if $\rho \in V_d(\mathbb{R}_+)$, $\rho \in V_l(\mathbb{R}_+)$ for any $l \ge d$, i.e. $d \notin \mathcal{D} \Rightarrow l \notin \mathcal{D}$ for any $l \ge d$.

non-degenerate global minimizers for any $k \leq m'$, i.e. $m' \in \mathcal{N}$. According to Theorem 57, for any $d \in \mathbb{N}_+$, $1 \leq d \leq m' = \min\{m, m'\} \leq \min\{m, M\}$, there exists at least $d! \left\{ {m \atop d} \right\}$ d-coincided critical affine spaces of J_m. Sum over d gives the total number $\sum_{d=1}^{m'} d! \left\{ {m \atop d} \right\}$. ■

Remark 69 *Examples:*

- *Suppose the target is $\rho(t) = 1/(1+t)^\alpha$, $\alpha > 0$, then $\mathcal{D} = \mathbb{N}_+$ and $D_0 = \infty$. The total number of coincided critical affine spaces of the corresponding J_m is at least $\sum_{d=1}^{m} d! \left\{ {m \atop d} \right\}$.*
- *Suppose the target is an non-degenerate exponential sum with positive coefficients: $\rho(t) = \sum_{j=1}^{m} a_j^* e^{-w_j^* t}$, where $a_j^* > 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j = 1, \cdots, m$. Then $\mathcal{D} = \{1, 2, \cdots, m-1\}$ and $D_0 = m - 1$. The total number of coincided critical affine spaces of the corresponding J_m is at least $\sum_{d=1}^{m-1} d! \left\{ {m \atop d} \right\}$, which is exactly (224).*

An complement for Theorem 58 is as follows.

Theorem 70 *Fix any $m \in \mathbb{N}_+$ relatively large. Consider the loss J_m with the ground truth as a non-degenerate exponential sum with positive coefficients, i.e. $\rho(t) = \sum_{j=1}^{m} a_j^* e^{-w_j^* t}$, where $a_j^* > 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j = 1, \cdots, m$. Then in the landscape of J_m, the number of coincided critical affine spaces is at least $Poly(m)$ times larger than the number of global minimizers.*

Proof By Theorem 58, the only fact we need to show is $m \in \mathcal{N}$. Since $\rho \in L^2[0, \infty)$ is completely monotonic, and $\rho \notin V_k(\mathbb{R}_+)$ for any $k \leq m-1$, then by Theorem 67, $J_{(k)}$ has non-degenerate global minimizers for any $k \leq m-1$, i.e. $m-1 \in \mathcal{N}$. The proof is completed by noticing that J_m obviously has non-degenerate global minimizers, e.g. (a^*, w^*). ■

A.3 A Low-Dimensional Example

To further understand the structure of coincided critical affine spaces, we focus on a specific low-dimensional example in this section. That is

$$\min_{(a,w) \in \mathbb{R}^2 \times \mathbb{R}_+^2} J_2(a, w) = \left\| \sum_{i=1}^{2} a_i e^{-w_i t} - \rho(t) \right\|_{L^2[0,\infty)}^2,$$

with the ground truth to be a non-degenerate exponential sum $\rho(t) = \sum_{j=1}^{m^*} a_j^* e^{-w_j^* t}$, where $a_j^* \neq 0$ and $w_i^* \neq w_j^*$ for any $i \neq j$, $i, j = 1, \cdots, m^*$. As we will show later, the coincided critical affine spaces of J_2 *contain both saddles and degenerate stable points which are not global optimal.*

By Lemma 55, Remark 56 and Theorem 57, we know the 1-coincided critical affine space of J_2 exists, and it can be constructed by taking the non-degenerate global minimizer of J_1, say $(\hat{a}, \hat{w})$ with $\hat{a} \neq 0$ and $\hat{w} > 0$. Then $\mathcal{M}_{(\hat{a},\hat{w}),(2,1)} := \{(a_1, \hat{a} - a_1, \hat{w}, \hat{w}) : a_1 \in \mathbb{R}\} \in \mathbb{R}^4$ is a line,[28] and $\nabla J_2(a_1, \hat{a} - a_1, \hat{w}, \hat{w}) = 0$ for any $a_1 \in \mathbb{R}$. Denote the Hessian of J_2 on the line $\mathcal{M}_{(\hat{a},\hat{w}),(2,1)}$ by $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$, i.e. $\mathcal{A}_{(\hat{a},\hat{w})}(a_1) := \nabla^2 J_2(a_1, \hat{a} - a_1, \hat{w}, \hat{w})$. We investigate the landscape of J_2 on the line $\mathcal{M}_{(\hat{a},\hat{w}),(2,1)}$ by analyzing the eigenvalue distribution of $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$.

Proposition 71 *Suppose $m = m^* = 2$, and $0 < w_1^* < w_2^*$. Let $I_1 := [0, \hat{a}]$ and $I_2 := (-\infty, 0) \cup (\hat{a}, +\infty)$.[29] Then*

1. *If $a_1^* a_2^* < 0$, the minimal eigenvalue of $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$ is 0 for any $a_1 \in I_1$, and negative for any $a_1 \in I_2$;*
2. *If $a_1^* a_2^* > 0$ and $w_2^*/w_1^* < 2 + \sqrt{3}$, the minimal eigenvalue of $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$ is negative for any $a_1 \in I_1$, and 0 for any $a_1 \in I_2$.*

Proof Write $c(w) := \sum_{j=1}^{m^*} a_j^* \left[\frac{1}{2w(w+w_j^*)^2} - \frac{1}{(w+w_j^*)^3} \right]$, and $a_2 := \hat{a} - a_1$. A straightforward computation shows that

$$\mathcal{A}_{(\hat{a},\hat{w})}(a_1) = \begin{bmatrix} \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{-a_1}{2\hat{w}^2} & \frac{-a_1}{2\hat{w}^2} & \frac{a_1^2}{2\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{2\hat{w}^3} \\ \frac{-a_2}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} & \frac{a_1 a_2}{2\hat{w}^3} & \frac{a_2^2}{2\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix}.$$

Considering the congruent transformation of $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$, which does not affect the index of inertia:

$$\begin{aligned} \mathcal{A}_{(\hat{a},\hat{w})}(a_1) &= \begin{bmatrix} \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{1}{\hat{w}} & \frac{1}{\hat{w}} & \frac{-a_1}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} \\ \frac{-a_1}{2\hat{w}^2} & \frac{-a_1}{2\hat{w}^2} & \frac{a_1^2}{2\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{2\hat{w}^3} \\ \frac{-a_2}{2\hat{w}^2} & \frac{-a_2}{2\hat{w}^2} & \frac{a_1 a_2}{2\hat{w}^3} & \frac{a_2^2}{2\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix} \\ &\to \begin{bmatrix} \frac{1}{\hat{w}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{a_1^2}{4\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{4\hat{w}^3} \\ 0 & 0 & \frac{a_1 a_2}{4\hat{w}^3} & \frac{a_2^2}{4\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix}, \end{aligned}$$

we get that $\mathcal{A}_{(\hat{a},\hat{w})}(a_1)$ has one positive eigenvalue $1/\hat{w}$ and one eigenvalue 0. What remains are the eigenvalues of $\mathcal{A}'_{(\hat{a},\hat{w})}(a_1) := \begin{bmatrix} \frac{a_1^2}{4\hat{w}^3} + 4c(\hat{w})a_1 & \frac{a_1 a_2}{4\hat{w}^3} \\ \frac{a_1 a_2}{4\hat{w}^3} & \frac{a_2^2}{4\hat{w}^3} + 4c(\hat{w})a_2 \end{bmatrix}$. To determine their signs, we compute

$$\begin{aligned} \det(\mathcal{A}'_{(\hat{a},\hat{w})}(a_1)) &= a_1(\hat{a} - a_1) \cdot 4c(\hat{w}) \left(\frac{\hat{a}}{4\hat{w}^3} + 4c(\hat{w}) \right) \\ &= \frac{1}{\hat{a}^2 \hat{w}^3} a_1(\hat{a} - a_1) \cdot \hat{a} c(\hat{w}) \cdot \left(\hat{a}^2 + 16\hat{w}^3 \hat{a} c(\hat{w}) \right). \end{aligned} \tag{236}$$

28. Here we omit the corresponding partition $\mathcal{P}$ since it is unique.
29. Suppose $\hat{a} > 0$ here without loss of generality. If $\hat{a} < 0$, we let $I_1 := [\hat{a}, 0]$ and $I_2 := (-\infty, \hat{a}) \cup (0, +\infty)$ and the same results hold.

So we need to analyze the sign of $\hat{a}c(\hat{w})$ and $\hat{a}^2 + 16\hat{w}^3\hat{a}c(\hat{w})$ under different assumptions on (a^*, w^*).

(i) $a_1^* a_2^* < 0$. By the optimality condition of $(\hat{a}, \hat{w})$ for J_1, we have

$$\hat{a} = 2\hat{w}\sum_{j=1}^{m^*} \frac{a_j^*}{\hat{w} + w_j^*} = 4\hat{w}^2 \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w} + w_j^*)^2}, \tag{237}$$

and therefore

$$c(\hat{w}) = \frac{\hat{a}}{8\hat{w}^3} - \sum_{j=1}^{m^*} \frac{a_j^*}{(\hat{w} + w_j^*)^3}.$$

Write $v_j := w_j^*/\hat{w}$, $j = 1, 2$, we get $0 < v_1 < v_2$, and

$$\hat{a} = 2\sum_{j=1}^{m^*} \frac{a_j^*}{1 + v_j} = 4\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^2}, \quad \hat{w}^3 c(\hat{w}) = \frac{\hat{a}}{8} - \sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3}.$$

Therefore

$$\begin{aligned} 8\hat{w}^3\hat{a}c(\hat{w}) &= \hat{a}^2 - 8\hat{a}\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3} \\ &= 16\left[\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^2}\right]^2 - 16\sum_{j=1}^{m^*} \frac{a_j^*}{1 + v_j} \cdot \sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3} \\ &= \frac{-16a_1^* a_2^* (v_1 - v_2)^2}{(1 + v_1)^3 (1 + v_2)^3} \\ &> 0, \end{aligned} \tag{238}$$

which gives $\hat{a}c(\hat{w}) > 0$ and $\hat{a}^2 + 16\hat{w}^3\hat{a}c(\hat{w}) > 8\hat{w}^3\hat{a}c(\hat{w}) > 0$.

(ii) $a_1^* a_2^* > 0, w_2^*/w_1^* < 2 + \sqrt{3}$. By (238), $\hat{a}c(\hat{w}) < 0$. Let $c := a_2^*/a_1^*$, $u_j := 1 + v_j > 1$ and $s := u_2/u_1 > 1$, we have

$$\begin{aligned} \hat{a}^2 + 16\hat{w}^3\hat{a}c(\hat{w}) &= 3\hat{a}^2 - 16\hat{a}\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3} \\ &= 16\left\{3\left[\sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^2}\right]^2 - 2\sum_{j=1}^{m^*} \frac{a_j^*}{1 + v_j} \cdot \sum_{j=1}^{m^*} \frac{a_j^*}{(1 + v_j)^3}\right\} \\ &= \frac{a_1^{*2}}{u_1^4 u_2^4}\left[u_2^4 + c^2 u_1^4 + 6cu_1^2 u_2^2 - 2cu_1^3 u_2 - 2cu_1 u_2^3\right] \\ &= \frac{a_1^{*2}}{u_2^4}\left(s^4 + c^2 + 6cs^2 - 2cs - 2cs^3\right) \\ &= \frac{a_1^{*2}}{u_2^4}\left[c^2 - 2s(s^2 - 3s + 1)c + s^4\right]. \end{aligned}$$

Since $4s^2(s^2-3s+1)^2-4s^4=4s^2(s-1)^2[(s-2)^2-3]$, and $1<s=u_2/u_1=(\hat{w}+w_2^*)/(\hat{w}+w_1^*)<w_2^*/w_1^*<2+\sqrt{3}$, we get $\Delta_c<0$. This implies $c^2-2s(s^2-3s+1)c+s^4>0$ and $\hat{a}^2+16\hat{w}^3\hat{a}c(\hat{w})>0$.

In both (i) and (ii), $\hat{a}^2+16\hat{w}^3\hat{a}c(\hat{w})>0$, which implies that there is at least one positive diagonal element of $\mathcal{A}'_{(\hat{a},\hat{w})}(a_1)$ in a sufficiently small neighborhood of $a_1=0$ and $a_1=\hat{a}$. By the Rayleigh-Ritz Theorem and Weyl's Theorem, $\mathcal{A}'_{(\hat{a},\hat{w})}(a_1)$ has at least one positive eigenvalue in this neighborhood. However, by (236), $\det(\mathcal{A}'_{(\hat{a},\hat{w})}(a_1))$ only changes the sign at $a_1=0$ and $a_1=\hat{a}$. This implies another eigenvalue of $\mathcal{A}'_{(\hat{a},\hat{w})}(a_1)$ changes the sign at $a_1=0$ and $a_1=\hat{a}$ accordingly. By different signs of $\hat{a}c(\hat{w})$ derived in (i) and (ii), and (236), the proof is completed. ■

Remark 72 *From Proposition 71, we deduce that there are both saddles and degenerate stable points of J_2 on the critical affine spaces (line) $\mathcal{M}_{(\hat{a},\hat{w}),(2,1)}$, and each of them in fact forms affine spaces (lines) respectively, but they are certainly not global minimizers. Therefore, the gradient-based algorithms can get stuck around this affine space, except that it meets saddles with negative eigenvalues with large magnitude.*

Appendix B. Momentum Helps Training: a Quadratic Example

In practice, it is often the case that training is trapped in some very flat regions (plateaus), where the loss function has rather small gradients and negative eigenvalues of Hessian. Now we illustrate the escape dynamics (escape from a plateau) via a simple quadratic example.

Consider the loss function $f(x)=(x_1^2-\epsilon x_2^2)/2$ with $0<\epsilon\ll 1$. We check the escaping performance for continuous-time analogues of two optimization algorithms: gradient decent (GD) and momentum (heavy ball) method.

B.1 Gradient Decent

Consider the gradient flow of $f(x)$ with an initial value $x_0=(\delta,1)^\top$, where $0<\delta\ll 1$ and $\delta=\mathcal{O}(\epsilon)$. Thus $\|\nabla f(x_0)\|=\mathcal{O}(\epsilon)$, and

$$\begin{cases} x_1'(\tau)=-x_1(\tau), & x_1(0)=\delta \\ x_2'(\tau)=\epsilon x_2(\tau), & x_2(0)=1 \end{cases} \Rightarrow \begin{cases} x_1(\tau)=\delta e^{-\tau} \\ x_2(\tau)=e^{\epsilon\tau} \end{cases}$$
$$\Rightarrow f(x(\tau))=(\delta^2e^{-2\tau}-\epsilon e^{2\epsilon\tau})/2\triangleq\ell_1(\tau).$$

It is easy to show that there are different timescales of $\ell_1(\tau)$. In fact, when $\tau=\mathcal{O}(1/\epsilon)$, $\ell_1(\tau)=\mathcal{O}(\epsilon^2)e^{-|\mathcal{O}(1/\epsilon)|}-\epsilon e^{|\mathcal{O}(1)|}=\mathcal{O}(\epsilon)$. However, when τ continues to increase, say

$$\tau\geq\frac{1}{2\epsilon}\ln\frac{\delta_0}{\epsilon}\triangleq\tau_1^\epsilon, \tag{239}$$

where $\delta_0>0$ denotes the gap satisfying $\epsilon=o(\delta_0)$, we get $\ell_1(\tau)\leq\ell_1(\tau_1^\epsilon)=\mathcal{O}(\epsilon^2)-\epsilon e^{2\epsilon\cdot\frac{1}{2\epsilon}\ln\frac{\delta_0}{\epsilon}}/2=\mathcal{O}(\epsilon^2)-\delta_0/2<-\delta_0/4$ for any $\tau\geq\tau_1^\epsilon$.

B.2 Momentum

The momentum algorithm has the update rule

$$x_{k+1} = x_k - \eta \nabla f(x_k) + \rho(x_k - x_{k-1}), \tag{240}$$

where $\rho \in \mathbb{R}$, $\eta > 0$ is the learning rate, and f is the objective. The continuous-time analogue can be derived as (see e.g. Su et al. (2014) for more details)

$$\begin{aligned} 0 &= \rho \frac{x_{k+1} - 2x_k + x_{k-1}}{\eta} + \frac{(1-\rho)}{\sqrt{\eta}} \frac{x_{k+1} - x_k}{\sqrt{\eta}} + \nabla f(x_k) \\ &\approx \rho x''(t) + \frac{(1-\rho)}{\sqrt{\eta}} x'(t) + \nabla f(x(t)), \end{aligned}$$

with $x_k := x(k\sqrt{\eta})$ and the step size $\sqrt{\eta}$ of the simple finite differences.[30] Let $x_1 = x_0 - \eta \nabla f(x_0)$, we also get $x'(0) = -\sqrt{\eta} \nabla f(x(0))$.

To facilitate a comparison to GD, we take $\eta = 1$,[31] and $\rho = 1$.[32] Plugging the expression of f, we can solve the ODE

$$\begin{aligned} x''(t) + \nabla f(x(t)) = 0 \Leftrightarrow & \begin{cases} x_1''(\tau) + x_1(\tau) = 0, & x_1(0) = \delta, \quad x_1'(0) = -\delta \\ x_2''(\tau) - \epsilon x_2(\tau) = 0, & x_2(0) = 1, \quad x_2'(0) = \epsilon \end{cases} \\ \Rightarrow & \begin{cases} x_1(\tau) = \delta(\cos\tau - \sin\tau) \\ x_2(\tau) = \frac{1+\sqrt{\epsilon}}{2} e^{\sqrt{\epsilon}\tau} + \frac{1-\sqrt{\epsilon}}{2} e^{-\sqrt{\epsilon}\tau} \end{cases} \\ \Rightarrow & f(x(\tau)) = \frac{1}{2} \left[\delta^2 (\cos\tau - \sin\tau)^2 - \epsilon \left(\frac{1+\sqrt{\epsilon}}{2} e^{\sqrt{\epsilon}\tau} + \frac{1-\sqrt{\epsilon}}{2} e^{-\sqrt{\epsilon}\tau} \right)^2 \right] \\ & \triangleq \ell_2(\tau). \end{aligned}$$

It is not hard to show that there are still different timescales of $\ell_2(\tau)$. In fact, when $\tau = \mathcal{O}(1/\sqrt{\epsilon})$, $\ell_2(\tau) = \mathcal{O}(\epsilon^2)|\mathcal{O}(1)| - \epsilon|\mathcal{O}(1)|(e^{|\mathcal{O}(1)|} + e^{-|\mathcal{O}(1)|})^2 = \mathcal{O}(\epsilon)$. However, when τ continues to increase, say

$$\tau \geq \frac{1}{2\sqrt{\epsilon}} \ln \frac{4\delta_0}{\epsilon} \triangleq \tau_2^\epsilon, \tag{241}$$

we get $\ell_2(\tau_2^\epsilon) = \mathcal{O}(\epsilon^2) - \epsilon(\frac{1+\sqrt{\epsilon}}{2})^2 e^{2\sqrt{\epsilon}\tau}/2 + \mathcal{O}(\epsilon) = \mathcal{O}(\epsilon) - \delta_0 < -\delta_0/2$, hence $\ell_2(\tau) \leq \mathcal{O}(\epsilon) - \delta_0 < -\delta_0/2$ for any $\tau \geq \tau_2^\epsilon$.

Combining Appendix B.1 and Appendix B.2 gives the following statements:

- For both dynamics, there are different timescales in the loss function. That is to say, relatively long time is needed to escape the plateaus;
- Compare (239) and (241), we get different timescales for escaping: $\mathcal{O}\left(1/\epsilon \cdot \ln(1/\epsilon)\right)$ for GD and $\mathcal{O}\left(1/\sqrt{\epsilon} \cdot \ln(1/\epsilon)\right)$ for momentum. Just like the convex case, where momentum improves the convergence rate by weakening the dependence on condition number, we show that momentum can also help to escape rather flat saddles.

30. It is easy to check the error is of order $\mathcal{O}(\sqrt{\eta})$.
31. In the continuous-time analogue of GD (gradient flow), the step size is taken as 1.
32. As is shown later, $\rho = 1$ not only simplifies the analysis, but also helps to obtain the best acceleration.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. On the convergence rate of training recurrent neural networks. In *Advances in Neural Information Processing Systems 32*, pages 1–13, 2019.

Pierre Baldi, Søren Brunak, Paolo Frasconi, Giovanni Soda, and Gianluca Pollastri. Exploiting the past and the future in protein secondary structure prediction. *Bioinformatics*, 15(11):937–946, 1999.

Bernhard Beckermann and Alex Townsend. On the singular values of matrices with displacement structure. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1227–1248, 2017.

Richard Ernest Bellman. *Dynamic Programming*. Princeton University Press, 1957.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Jan Beran. Statistical methods for data with long-range dependence. *Statistical Science*, 7: 404–416, 1992.

Sergei Natanovich Bernstein. On the best approximation of continuous functions by polynomials of given degree. *Soobshcheniya of Kharkov Society of Mathematics*, 2:13, 1920.

Vladimir I. Bogachev. *Measure Theory*, volume 1. Springer Science & Business Media, 2007.

Dietrich Braess. Approximation by exponential sums. In *Nonlinear Approximation Theory*, pages 168–180. Springer, 1986.

Ayaz Hussain Bukhari, Muhammad Asif Zahoor Raja, Muhammad Sulaiman, Saeed Islam, Muhammad Shoaib, and Poom Kumam. Fractional neuro-sequential ARFIMA-LSTM for financial market forecasting. *IEEE Access*, 8:71326–71338, 2020.

Víctor Campos, Brendan Jou, Xavier Giró-i Nieto, Jordi Torres, and Shih-Fu Chang. Skip RNN: Learning to skip state updates in recurrent neural networks. In *International Conference on Learning Representations*, pages 1–17, 2018.

Andrea Ceni, Peter Ashwin, and Lorenzo Livi. Interpreting recurrent neural networks behaviour via excitable network attractors. *Cognitive Computation*, 12:330–356, 2020.

Bo Chang, Minmin Chen, Eldad Haber, and Ed H. Chi. AntisymmetricRNN: A dynamical system view on recurrent neural networks. In *International Conference on Learning Representations*, pages 1–15, 2019.

Tianping Chen and Hong Chen. Approximations of continuous functionals by neural networks with application to dynamical systems. *IEEE Transactions on Neural Networks*, 4 (6):910–918, 1993.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*, 2014.

Tommy W. S. Chow and Xiao-Dong Li. Modeling of continuous time dynamical systems with input by recurrent neural networks. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 47(4):575–578, 2000.

Eugen Diaconescu. The use of NARX neural networks to predict chaotic time series. *WSEAS Transactions on Computers Archive*, 3:182–191, 2008.

Adji B. Dieng, Chong Wang, Jianfeng Gao, and John Paisley. TopicRNN: A recurrent neural network with long-range semantic dependency. In *International Conference on Learning Representations*, pages 1–13, 2017.

Paul Doukhan, George Oppenheim, and Murad Taqqu. *Theory and Applications of Long-Range Dependence*. Springer Science & Business Media, 2002.

Kenji Doya. Universality of fully-connected recurrent neural networks. *IEEE Transactions on Neural Networks*, 1993.

Alex Dytso, Ronit Bustin, H. Vincent Poor, and Shlomo Shamai. Analytical properties of generalized Gaussian distributions. *Journal of Statistical Distributions and Applications*, 5(6):1–40, 2018.

Weinan E. A proposal on machine learning via dynamical systems. *Communications in Mathematics and Statistics*, 5(1):1–11, 2017.

Weinan E, Chao Ma, and Lei Wu. A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics*, 63(7):1235–1258, 2020.

Ken-ichi Funahashi and Yuichi Nakamura. Approximation of dynamical systems by continuous time recurrent neural networks. *Neural Networks*, 6(6):801–806, 1993.

Lukas Gonon, Lyudmila Grigoryeva, and Juan-Pablo Ortega. Approximation bounds for random neural networks and reservoir systems. *arXiv preprint arXiv:2002.05933*, 2020.

Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

Alex Graves and Navdeep Jaitly. Towards end-to-end speech recognition with recurrent neural networks. In *International Conference on Machine Learning*, pages 1764–1772, 2014.

Alex Graves and Jürgen Schmidhuber. Offline handwriting recognition with multidimensional recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 545–552, 2009.

Alex Graves, Abdel-rahman Mohamed, and Geoffrey Hinton. Speech recognition with deep recurrent neural networks. In *International Conference on Acoustics, Speech and Signal Processing*, pages 6645–6649, 2013.

Karol Gregor, Ivo Danihelka, Alex Graves, Danilo Rezende, and Daan Wierstra. DRAW: A recurrent neural network for image generation. In *International Conference on Machine Learning*, pages 1462–1471, 2015.

Eldad Haber and Lars Ruthotto. Stable architectures for deep neural networks. *Inverse Problems*, 34(1):014004, 2017.

Boris Hanin. Which neural net architectures give rise to exploding and vanishing gradients? In *Advances in Neural Information Processing Systems 31*, pages 582–591. Curran Associates, Inc., 2018.

Boris Hanin and David Rolnick. How to start training: The effect of initialization and architecture. In *Advances in Neural Information Processing Systems 31*, pages 571–581. Curran Associates, Inc., 2018.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

Qi-Ming He and Hanqin Zhang. On matrix exponential distributions. *Advances in Applied Probability*, 39:271–292, 2007.

Calypso Herrera, Florian Krach, and Josef Teichmann. Theoretical guarantees for learning conditional expectation using controlled ODE-RNN. *arXiv preprint arXiv:2006.04727*, 2020.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997.

Sepp Hochreiter, Yoshua Bengio, Paolo Frasconi, and Jürgen Schmidhuber. Gradient flow in recurrent nets: The difficulty of learning long-term dependencies. In *A Field Guide to Dynamical Recurrent Neural Networks*. IEEE Press, 2001.

Yuhuang Hu, Adrian Huber, Jithendar Anumula, and Shih-Chii Liu. Overcoming the vanishing gradient problem in plain recurrent networks. *arXiv preprint arXiv:1801.06105*, 2018.

Harold Edwin Hurst. Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, 116:770–799, 1951.

Dunham Jackson. *The Theory of Approximation*, volume 11. American Mathematical Society, 1930.

David W. Kammler. Approximation with sums of exponentials in $L_p[0, \infty)$. *Journal of Approximation Theory*, 16(4):384–408, 1976.

David W. Kammler. Least squares approximation of completely monotonic functions by sums of exponentials. *SIAM Journal on Numerical Analysis*, 16(5):801–818, 1979.

Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, pages 1–15, 2015.

Qianxiao Li and Shuji Hao. An optimal control approach to deep learning and applications to discrete-weight neural networks. In *International Conference on Machine Learning*, pages 2985–2994, 2018.

Qianxiao Li, Long Chen, Cheng Tai, and Weinan E. Maximum principle based algorithms for deep learning. *Journal of Machine Learning Research*, 18(1):5998–6026, 2017.

Shuai Li, Wanqing Li, Chris Cook, Ce Zhu, and Yanbo Gao. Independently recurrent neural network (IndRNN): Building a longer and deeper RNN. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5457–5466, 2018.

Xiao-Dong Li, John K. L. Ho, and Tommy W. S. Chow. Approximation of dynamical time-variant systems by continuous-time recurrent neural networks. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 52:656–660, 2005.

Zhong Li, Jiequn Han, Weinan E, and Qianxiao Li. On the curse of memory in recurrent neural networks: Approximation and optimization analysis. In *International Conference on Learning Representations*, pages 1–43, 2021.

Soon Hoe Lim. Understanding recurrent neural networks using nonequilibrium response theory. *Journal of Machine Learning Research*, 22:1–48, 2021.

Edward N. Lorenz. Predictability: A problem partly solved. In *Proc. Seminar on Predictability*, volume 1, pages 40–58, 1996.

Georgios Loukas and Gülay Öke. Likelihood ratios and recurrent random neural networks in detection of denial of service attacks. In *Society for Modeling and Simulation International*, pages 1–8, 2007.

Lu Lu, Pengzhan Jin, and George Em Karniadakis. DeepONet: Learning nonlinear operators for identifying differential equations based on the universal approximation theorem of operators. *arXiv preprint arXiv:1910.03193*, 2019.

Chao Ma, Jianchun Wang, and Weinan E. Model reduction with memory and the machine learning of dynamical systems. *Communications in Computational Physics*, 25(4):947–962, 2018.

Wolfgang Maass, Prashant Joshi, and Eduardo D. Sontag. Computational aspects of feedback in neural circuits. *PLOS Computational Biology*, 3(1):e165, 2007.

Benoit B. Mandelbrot and John W. Van Ness. Fractional Brownian motions, fractional noises and applications. *SIAM Review*, 10:422–437, 1968.

Michael B. Matthews. Approximating nonlinear fading-memory operators using neural network models. *Circuits, Systems and Signal Processing*, 12:279–307, 1993.

Arvind T. Mohan and Datta V. Gaitonde. A deep learning based approach to reduced order modeling for turbulent flow control using lstm neural networks. *arXiv preprint arXiv:1804.09269*, 2018.

Ch H. Müntz. Über den approximationssatz von Weierstrass. In *Mathematische Abhandlungen Hermann Amandus Schwarz*, pages 303–312. Springer, 1914.

Yuichi Nakamura and Masahiro Nakagawa. Approximation capability of continuous time recurrent neural networks for non-autonomous dynamical systems. In *International Conference on Artificial Neural Networks*, pages 593–602, 2009.

Murphy Yuezhen Niu, Lior Horesh, and Isaac Chuang. Recurrent neural networks in the eye of differential equations. *arXiv preprint arXiv:1904.12933*, 2019.

Colm Art O'Cinneide. Characterization of phase-type distributions. *Stochastic Models*, 6: 1–57, 1990.

Aäron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pages 1310–1318, 2013.

Yulia Rubanova, Ricky T. Q. Chen, and David Duvenaud. Latent ordinary differential equations for irregularly-sampled time series. In *Advances in Neural Information Processing Systems 32*, pages 5320–5330, 2019.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

Gennady Samorodnitsky. Long range dependence. *Foundations and Trends in Stochastic Systems*, 1:163–257, 2006.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. In *International Conference on Artificial Neural Networks*, pages 632–640, 2006.

Anton Maximilian Schäfer and Hans Georg Zimmermann. Recurrent neural networks are universal approximators. *International Journal of Neural Systems*, 17(04):253–263, 2007.

Mike Schuster and Kuldip K. Paliwal. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681, 1997.

Alex Sherstinsky. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *arXiv preprint arXiv:1808.03314*, 2018.

Weijie Su, Stephen Boyd, and Emmanuel Candes. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. In *Advances in Neural Information Processing Systems*, pages 2510–2518, 2014.

Otto Szász. Über die approximation stetiger funktionen durch lineare aggregate von potenzen. *Mathematische Annalen*, 77(4):482–496, 1916.

Sachin S. Talathi and Aniket Vartak. Improving performance of recurrent neural network with ReLU nonlinearity. *arXiv preprint arXiv:1511.03771*, 2015.

Murad S. Taqqu, Vadim Teverovsky, and Walter Willinger. Estimators for long-range dependence: An empirical study. *Fractals*, 3:785–798, 1995.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

Trieu H. Trinh, Andrew M. Dai, Minh-Thang Luong, and Quoc V. Le. Learning longer-term dependencies in RNNs with auxiliary losses. In *International Conference on Machine Learning*, pages 4965–4974, 2018.

Tzu-Hsuan Tseng, Tzu-Hsuan Yang, and Chia-Ping Chen. Verifying the long-range dependency of RNN language models. In *International Conference on Asian Language Processing*, pages 75–78, 2016.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

Robert Zwanzig. *Nonequilibrium Statistical Mechanics*. Oxford University Press, 2001.